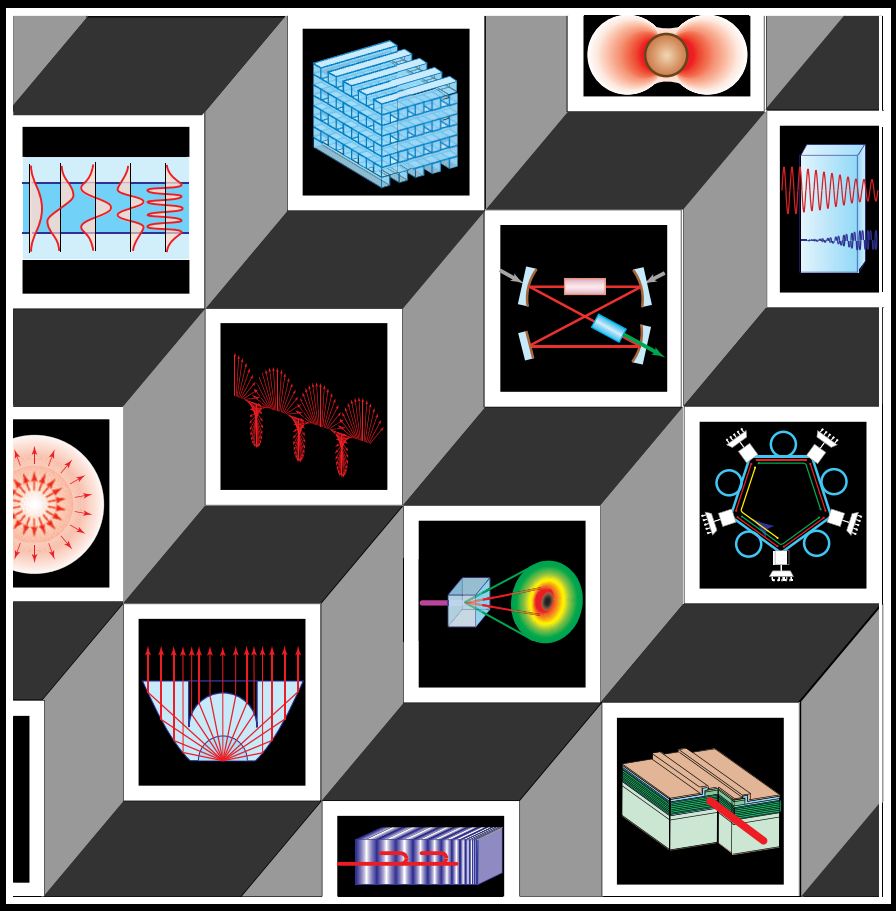
Wiley Series in Pure and Applied Optics

G. Boreman, Editor

FUNDAMENTALS OF PHOTONICS



Third Edition

B. E. A. Saleh

M. C. Teich

WILEY

FUNDAMENTALS OF PHOTONICS

FUNDAMENTALS OF PHOTONICS

THIRD EDITION

BAHAA E. A. SALEH
University of Central Florida
Boston University

MALVIN CARL TEICH
Boston University
Columbia University

WILEY

This edition first published 2019
© 2019 by John Wiley & Sons, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at http://www.wiley.com/go/permissions.

The rights of Bahaa E. A. Saleh and Malvin Carl Teich to be identified as the authors of the editorial material in this work have been asserted in accordance with law.

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data is available.
Volume Set ISBN: 9781119506874
Volume I ISBN: 9781119506867
Volume II ISBN: 9781119506898

Cover design by Wiley
Cover image: Courtesy of B. E. A. Saleh and M. C. Teich

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE TO THE THIRD EDITION xi

PREFACE TO THE SECOND EDITION xx

PREFACE TO THE FIRST EDITION xxiii

PART I: OPTICS 1

1 RAY OPTICS 3

1.1 Postulates of Ray Optics 5
1.2 Simple Optical Components 8
1.3 Graded-Index Optics 20
1.4 Matrix Optics 27
Reading List 37
Problems 38

2 WAVE OPTICS 41

2.1 Postulates of Wave Optics 43
2.2 Monochromatic Waves 44
*2.3 Relation Between Wave Optics and Ray Optics 52
2.4 Simple Optical Components 53
2.5 Interference 61
2.6 Polychromatic and Pulsed Light 71
Reading List 76
Problems 77

3 BEAM OPTICS 79

3.1 The Gaussian Beam 80
3.2 Transmission Through Optical Components 91
3.3 Hermite–Gaussian Beams 99
3.4 Laguerre–Gaussian Beams 102
3.5 Nondiffracting Beams 105
Reading List 108
Problems 108

4 FOURIER OPTICS 110

4.1 Propagation of Light in Free Space 113
4.2 Optical Fourier Transform 124
4.3 Diffraction of Light 129
4.4 Image Formation 137
4.5 Holography 147
Reading List 155
Problems 157

5 ELECTROMAGNETIC OPTICS 160

5.1 Electromagnetic Theory of Light 162
5.2 Electromagnetic Waves in Dielectric Media 166
5.3 Monochromatic Electromagnetic Waves 172
5.4 Elementary Electromagnetic Waves 175
5.5 Absorption and Dispersion 181
5.6 Scattering of Electromagnetic Waves 192
5.7 Pulse Propagation in Dispersive Media 199
Reading List 205
Problems 207

6 POLARIZATION OPTICS 209

6.1 Polarization of Light 211
6.2 Reflection and Refraction 221
6.3 Optics of Anisotropic Media 227
6.4 Optical Activity and Magneto-Optics 240
6.5 Optics of Liquid Crystals 244
6.6 Polarization Devices 247
Reading List 251
Problems 252

7 PHOTONIC-CRYSTAL OPTICS 255

7.1 Optics of Dielectric Layered Media 258
7.2 One-Dimensional Photonic Crystals 277
7.3 Two- and Three-Dimensional Photonic Crystals 291
Reading List 299
Problems 301

8 METAL AND METAMATERIAL OPTICS 303

8.1 Single- and Double-Negative Media 306
8.2 Metal Optics: Plasmonics 320
8.3 Metamaterial Optics 334
*8.4 Transformation Optics 343
Reading List 349
Problems 351

9 GUIDED-WAVE OPTICS 353

9.1 Planar-Mirror Waveguides 355
9.2 Planar Dielectric Waveguides 363
9.3 Two-Dimensional Waveguides 372
9.4 Optical Coupling in Waveguides 376
9.5 Photonic-Crystal Waveguides 385
9.6 Plasmonic Waveguides 386
Reading List 389
Problems 389

10 FIBER OPTICS 391

10.1 Guided Rays 393
10.2 Guided Waves 397
10.3 Attenuation and Dispersion 415
10.4 Holey and Photonic-Crystal Fibers 426
10.5 Fiber Materials 429
Reading List 430
Problems 432

11 RESONATOR OPTICS 433

11.1 Planar-Mirror Resonators 436
11.2 Spherical-Mirror Resonators 447
11.3 Two- and Three-Dimensional Resonators 459
11.4 Microresonators and Nanoresonators 463
Reading List 470
Problems 471

12 STATISTICAL OPTICS 473

12.1 Statistical Properties of Random Light 475
12.2 Interference of Partially Coherent Light 489
*12.3 Transmission of Partially Coherent Light 497
12.4 Partial Polarization 506
Reading List 510
Problems 512

13 PHOTON OPTICS 514

13.1 The Photon 516
13.2 Photon Streams 529
*13.3 Quantum States of Light 541
Reading List 550
Problems 554

PART II: PHOTONICS 559

14 LIGHT AND MATTER 561

14.1 Energy Levels 562
14.2 Occupation of Energy Levels 581
14.3 Interactions of Photons with Atoms 583
14.4 Thermal Light 602
14.5 Luminescence and Scattering 607
Reading List 614
Problems 617

15 LASER AMPLIFIERS 619

15.1 Theory of Laser Amplification 622
15.2 Amplifier Pumping 626
15.3 Representative Laser Amplifiers 636
15.4 Amplifier Nonlinearity 645
*15.5 Amplifier Noise 651
Reading List 653
Problems 655

16 LASERS 657

16.1 Theory of Laser Oscillation 659
16.2 Characteristics of the Laser Output 666
16.3 Types of Lasers 680
16.4 Pulsed Lasers 707
Reading List 723
Problems 728

17 SEMICONDUCTOR OPTICS 731

17.1 Semiconductors 733
17.2 Interactions of Photons with Charge Carriers 766
Reading List 782
Problems 784

18 LEDS AND LASER DIODES 787

18.1 Light-Emitting Diodes 789
18.2 Semiconductor Optical Amplifiers 817
18.3 Laser Diodes 831
18.4 Quantum-Confined Lasers 844
18.5 Microcavity Lasers 854
18.6 Nanocavity Lasers 862
Reading List 864
Problems 868

19 PHOTODETECTORS 871

19.1 Photodetectors 873
19.2 Photoconductors 883
19.3 Photodiodes 887
19.4 Avalanche Photodiodes 895
19.5 Array Detectors 907
19.6 Noise in Photodetectors 909
Reading List 935
Problems 938

20 ACOUSTO-OPTICS 943

20.1 Interaction of Light and Sound 945
20.2 Acousto-Optic Devices 958
*20.3 Acousto-Optics of Anisotropic Media 967
Reading List 972
Problems 972

21 ELECTRO-OPTICS 975

21.1 Principles of Electro-Optics 977
*21.2 Electro-Optics of Anisotropic Media 989
21.3 Electro-Optics of Liquid Crystals 996
*21.4 Photorefractivity 1005
21.5 Electroabsorption 1010
Reading List 1012
Problems 1013

22 NONLINEAR OPTICS 1015

22.1 Nonlinear Optical Media 1017
22.2 Second-Order Nonlinear Optics 1021
22.3 Third-Order Nonlinear Optics 1036
*22.4 Second-Order Nonlinear Optics: Coupled Waves 1047
*22.5 Third-Order Nonlinear Optics: Coupled Waves 1059
*22.6 Anisotropic Nonlinear Media 1066
*22.7 Dispersive Nonlinear Media 1069
Reading List 1074
Problems 1075

23 ULTRAFAST OPTICS 1078

23.1 Pulse Characteristics 1079
23.2 Pulse Shaping and Compression 1088
23.3 Pulse Propagation in Optical Fibers 1102
23.4 Ultrafast Linear Optics 1115
23.5 Ultrafast Nonlinear Optics 1126
23.6 Pulse Detection 1146
Reading List 1159
Problems 1161

24 OPTICAL INTERCONNECTS AND SWITCHES 1163

24.1 Optical Interconnects 1166
24.2 Passive Optical Routers 1178
24.3 Photonic Switches 1187
24.4 Photonic Logic Gates 1211
Reading List 1220
Problems 1222

25 OPTICAL FIBER COMMUNICATIONS 1224

25.1 Fiber-Optic Components 1226
25.2 Optical Fiber Communication Systems 1238
25.3 Modulation and Multiplexing 1257
25.4 Coherent Optical Communications 1266
25.5 Fiber-Optic Networks 1274
Reading List 1281
Problems 1284

A FOURIER TRANSFORM 1287

A.1 One-Dimensional Fourier Transform 1287
A.2 Time Duration and Spectral Width 1290
A.3 Two-Dimensional Fourier Transform 1293
Reading List 1295

B LINEAR SYSTEMS 1296

B.1 One-Dimensional Linear Systems 1296
B.2 Two-Dimensional Linear Systems 1299
Reading List 1300

C MODES OF LINEAR SYSTEMS 1301

Reading List 1305

SYMBOLS AND UNITS 1306

AUTHORS 1331

INDEX 1333

PREFACE TO THE THIRD EDITION

Since the publication of the *Second Edition* in 2007, *Fundamentals of Photonics* has maintained its worldwide prominence as a self-contained, up-to-date, introductory-level textbook that features a blend of theory and applications. It has been reprinted dozens of times and been translated into German and Chinese, as well as Czech and Japanese. The Third Edition incorporates many of the scientific and technological developments in photonics that have taken place in the past decade and strives to be cutting-edge.

Optics and Photonics

Before usage of the term photonics became commonplace at the time of the *First Edition* in the early 1990s, the field was characterized by a collection of appellations that were not always clearly delineated. Terms such as quantum electronics, optoelectronics, electro-optics, and lightwave technology were widely used. Though there was a lack of agreement about the precise meanings of these terms, there was nevertheless a vague consensus regarding their usage. Most of these terms have since receded from general use, although some have retained their presence in the titles of technical journals and academic courses.

Now, more than 25 years later, the term *Optics* along with the term *Photonics*, as well as their combination *Optics & Photonics*, have prevailed. The distinction between optics and photonics remains somewhat fuzzy, however, and there is a degree of overlap between the two arenas. Hence, there is some arbitrariness in the manner in which the chapters of this book are allocated to its two volumes, *Part I: Optics* and *Part II: Photonics*. From a broad perspective, the term *Optics* is taken to signify free-space and guided-wave propagation, and to include topics such as interference, diffraction, imaging, statistical optics, and photon optics. The term *Photonics*, in contrast, is understood to include topics that rely on the interaction of light and matter, and is dedicated to the study of devices and systems. As the miniaturization of components and systems continues to progress and foster emerging technologies such as nanophotonics and biophotonics, the importance of photonics continues to advance.

Printed and Electronic Versions

The *Third Edition* appears in four versions:

1. A printed version.
2. An eBook in the form of an ePDF file that mimics the printed version.
3. An eBook in the form of a standard ePUB.
4. An eBook in the form of an enhanced ePUB with animations for selected figures.

In its *printed* form, the text consists of two volumes, each of which contains the Table of Contents and Index for both volumes along with the Appendices and List of Symbols:

- *Part I: Optics*, contains the first thirteen chapters.
- *Part II: Photonics*, contains the remaining twelve chapters.

The material in the eBook versions is identical to that in the printed version except that all 25 chapters reside in a single electronic file. The various *eBooks* enjoy the following features:

- Hyperlinked table of contents at the beginning of the text.

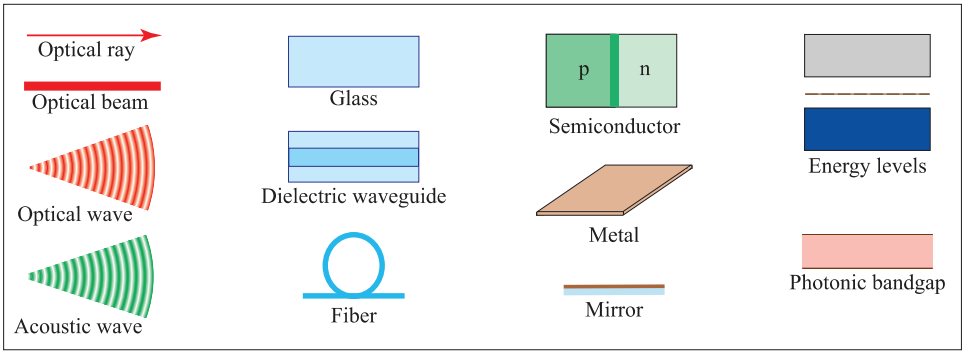
- Hyperlinked table of contents as an optional sidebar.
- Hyperlinked index.
- Hyperlinked section titles, equations, and figures throughout.
- Animations for selected figures in the enhanced ePUB.

Presentation

Exercises, examples, reading lists, and appendices. Each chapter of the *Third Edition* contains exercises, problem sets, and an extensive reading list. Examples are included throughout to emphasize the concepts governing applications of current interest. Appendices summarize the properties of one- and two-dimensional Fourier transforms, linear systems, and modes of linear systems. Important equations are highlighted by boxes and labels to facilitate retrieval.

Symbols, notation, units, and conventions. We make use of the symbols, notation, units, and conventions commonly used in the photonics literature. Because of the broad spectrum of topics covered, different fonts are often used to delineate the multiple meanings of various symbols; a list of symbols, units, abbreviations, and acronyms follows the appendices. We adhere to the International System of Units (SI units). This modern form of the metric system is based on the meter, kilogram, second, ampere, kelvin, candela, and mole, and is coupled with a collection of prefixes (specified on the inside back cover of the text) that indicate multiplication or division by various powers of ten. However, the reader is cautioned that photonics in the service of different areas of science can make use of different conventions and symbols. In Chapter 2, for example, we write the complex wavefunction for a monochromatic plane wave in a form commonly used in electrical engineering, which differs from that used in physics. Another example arises in Chapter 6, where the definitions we use for right (left) circularly polarized light are in accord with general usage in optics, but are opposite those generally used in engineering. These distinctions are often highlighted by *in situ* footnotes. Though the choice of a particular convention is manifested in the form assumed by various equations, it does not of course affect the results.

Color coding of illustrations. The color code used in illustrations is summarized in the chart presented below. Light beams and optical-field distributions are displayed in red (except when light beams of multiple wavelengths are involved, as is often the case in nonlinear optics). When optical fields are represented, white indicates negative values but when intensity is portrayed, white indicates zero. Acoustic beams and fields are similarly represented, but by with green rather than red. Glass and glass fibers are depicted in light blue; darker shades represent larger refractive indices. Semiconductors are cast in green, with various shades representing different doping levels. Metal and mirrors are indicated as copper. Semiconductor energy-band diagrams are portrayed in blue and gray whereas photonic bandgaps are illustrated in pink.



Color chart

Intended Audience

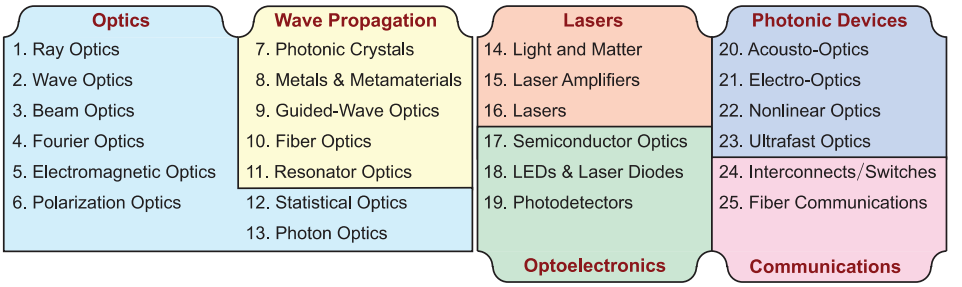
As with the previous editions, the *Third Edition* is meant to serve as:

- An introductory textbook for students of electrical engineering, applied physics, physics, or optics at the senior or first-year graduate level.
- A self-contained work for self-study.
- A textbook suitable for use in programs of continuing professional development offered by industry, universities, and professional societies.

The reader is assumed to have a background in engineering, physics, or optics, including courses in modern physics, electricity and magnetism, and wave motion. Some knowledge of linear systems and elementary quantum mechanics is helpful but not essential. The intent is to provide an introduction to optics and photonics that emphasizes the concepts that govern applications of current interest. The book should therefore not be considered as a compendium encompassing all photonic devices and systems. Indeed, some areas of photonics are not included at all, and many of the individual chapters could easily have been expanded into free-standing monographs.

Organization

The *Third Edition* comprises 25 chapters compartmentalized into six divisions, as depicted in the diagram below.



In recognition of the different levels of mathematical sophistication of the intended audience, we have endeavored to present difficult concepts in two steps: at an introductory level that provides physical insight and motivation, followed by a more advanced analysis. This approach is exemplified by the treatment in Chapter 21 (*Electro-Optics*), in which the subject is first presented using scalar notation and then treated again using tensor notation. Sections dealing with material of a more advanced nature are indicated by asterisks and may be omitted if desired. Summaries are provided at points where recapitulation is deemed useful because of the involved nature of the material.

The form of the book is modular so that it can be used by readers with different needs; this also provides instructors an opportunity to select topics for different courses. Essential material from one chapter is often briefly summarized in another to make each chapter as self-contained as possible. At the beginning of Chapter 25 (*Optical Fiber Communications*), for example, relevant material from earlier chapters describing optical fibers, light sources, optical amplifiers, photodetectors, and photonic integrated circuits is briefly reviewed. This places important information about the components of such systems at the disposal of the reader in advance of presenting system-design and performance considerations.

Contents

A principal feature of the *Third Edition* is a new chapter entitled *Metal and Metamaterial Optics*, an area that has had a substantial and increasing impact on photonics.

The new chapter comprises theory and applications for single- and double-negative media, metal optics, plasmonics, metamaterial optics, and transformation optics.

All chapters have been thoroughly vetted and updated. A chapter-by-chapter compilation of new material in the *Third Edition* is provided below.

- ***Chapter 1 (Ray Optics)***. Ray-optics descriptions for optical components such as biprisms, axicons, LED light collimators, and Fresnel lenses have been added. The connection between characterizing an arbitrary paraxial optical system by its ray-transfer matrix and its cardinal points has been established. A matrix-optics analysis for imaging with an arbitrary paraxial optical system has been included.
- ***Chapter 2 (Wave Optics)***. A wave-optics analysis of transmission through biprisms and axicons has been added. A treatment of the Fresnel zone plate from the perspective of interference has been introduced. An analysis of the Michelson–Fabry–Perot (LIGO) interferometer used for the detection of gravitational waves in the distant universe has been incorporated.
- ***Chapter 3 (Beam Optics)***. An enhanced description of Laguerre–Gaussian beams has been provided. The basic features of several additional optical beams have been introduced: optical vortex, Ince–Gaussian, nondiffracting Bessel, Bessel–Gaussian, and Airy.
- ***Chapter 4 (Fourier Optics)***. An analysis of Fresnel diffraction from a periodic aperture (Talbot effect) has been included. Nondiffracting waves and Bessel beams have been introduced from a Fourier-optics perspective. A discussion of computer-generated holography has been added.
- ***Chapter 5 (Electromagnetic Optics)***. A new section on the dipole wave, the basis of near-field optics, has been incorporated. A new section on scattering that includes Rayleigh and Mie scattering, along with attenuation in a medium with scatterers, has been added.
- ***Chapter 6 (Polarization Optics)***. The material dealing with the dispersion relation in anisotropic media has been reworked to simplify the presentation.
- ***Chapter 7 (Photonic-Crystal Optics)***. The behavior of the dielectric-slab beam-splitter has been elucidated. A discussion relating to fabrication methods for 3D photonic crystals has been incorporated.
- ***Chapter 8 (Metal and Metamaterial Optics)***. This new chapter, entitled Metal and Metamaterial Optics, provides a venue for the examination of single- and double-negative media, metal optics, plasmonics, metamaterial optics, and transformation optics. Topics considered include evanescent waves, surface plasmon polaritons, localized surface plasmons, nanoantennas, metasurfaces, subwavelength imaging, and optical cloaking.
- ***Chapter 9 (Guided-Wave Optics)***. A new section on waveguide arrays that details the mutual coupling of multiple waveguides and introduces the notion of supermodes has been inserted. A new section on plasmonic waveguides that includes metal–insulator–metal and metal-slab waveguides, along with periodic metal–dielectric arrays, has been incorporated.
- ***Chapter 10 (Fiber Optics)***. A discussion of multicore fibers, fiber couplers, and photonic lanterns has been added. A brief discussion of the applications of photonic-crystal fibers has been provided. A new section on multimaterial fibers, including conventional and hybrid mid-infrared fibers, specialty fibers, multimaterial fibers, and multifunctional fibers, has been introduced.
- ***Chapter 11 (Resonator Optics)***. A section on plasmonic resonators has been added.
- ***Chapter 12 (Statistical Optics)***. The sections on optical coherence tomography and unpolarized light have been reorganized.
- ***Chapter 13 (Photon Optics)***. A brief description of single-photon imaging has been added. The discussion of quadrature-squeezed and photon-number-squeezed

light has been enhanced and examples of the generation and applications of these forms of light have been provided. A section that describes two-photon light, entangled photons, two-photon optics, and the generation and applications thereof, has been incorporated. Examples of two-photon polarization, two-photon spatial optics, and two-beam optics have been appended.

- ***Chapter 14 (Light and Matter)***. The title of this chapter was changed from *Photons and Atoms* to *Light and Matter*. Brief descriptions of the Zeeman effect, Stark effect, and ionization energies have been added. The discussion of lanthanide-ion manifolds has been enhanced. Descriptions of Doppler cooling, optical molasses, optical tweezers, optical lattices, atom interferometry, and atom amplifiers have been incorporated into the section on laser cooling, laser trapping, and atom optics.
- ***Chapter 15 (Laser Amplifiers)***. Descriptions of quasi-three-level and in-band pumping have been added. The sections on representative laser amplifiers, including ruby, neodymium-doped glass, erbium-doped silica fiber, and Raman fiber devices, have been enhanced.
- ***Chapter 16 (Lasers)***. Descriptions of tandem pumping, transition-ion-doped zinc-chalcogenide lasers, silicon Raman lasers, and master-oscillator power-amplifiers (MOPAs) have been added. Descriptions of inner-shell photopumping and X-ray free-electron lasers have been incorporated. A new section on optical frequency combs has been provided.
- ***Chapter 17 (Semiconductor Optics)***. The section on organic semiconductors has been enhanced. A discussion of group-IV photonics, including graphene and 2D materials such as transition-metal dichalcogenides, has been added. A brief discussion of quantum-dot single-photon emitters has been incorporated.
- ***Chapter 18 (LEDs and Laser Diodes)***. The title of this chapter was changed from *Semiconductor Photon Sources* to *LEDs and Laser Diodes*. A new section on the essentials of LED lighting has been incorporated. Brief discussions of the following topics are now included: resonant-cavity LEDs, silicon-photonics light sources, quantum-dot semiconductor amplifiers, external-cavity wavelength-tunable laser diodes, broad-area laser diodes, and laser-diode bars and stacks. A discussion of the semiconductor-laser linewidth-enhancement factor has been added. A new section on nanolasers has been introduced.
- ***Chapter 19 (Photodetectors)***. The title of this chapter was changed from *Semiconductor Photon Detectors* to *Photodetectors*. Brief discussions of the following topics have been added: organic, plasmonic, group-IV-based, and graphene-enhanced photodetectors; edge vs. normal illumination; photon-trapping microstructures; SACM and superlattice APDs; multiplied dark current; and $1/f$ detector noise. New examples include multi-junction photovoltaic solar cells; Ge-on-Si photodiodes; graphene-Si Schottky-barrier photodiodes; and SAM, SACM, and staircase APDs. A new section on single-photon and photon-number-resolving detectors details the operation of SPADs, SiPMs, and TESs.
- ***Chapter 20 (Acousto-Optics)***. The identical forms of the photoelastic matrix in acousto-optics and the Kerr-effect matrix in electro-optics has been highlighted for cubic isotropic media.
- ***Chapter 21 (Electro-Optics)***. New sections on passive- and active-matrix liquid-crystal displays have been introduced and their operation has been elucidated. The performance of active-matrix liquid-crystal displays (AMLCDs) has been compared with that of active-matrix organic light-emitting displays (AMOLEDs).
- ***Chapter 22 (Nonlinear Optics)***. New material relating to guided-wave nonlinear optics has been introduced. Quasi-phase matching in periodically poled integrated optical waveguides, and the associated improvement in wave-mixing efficiency, is now considered. The section pertaining to Raman gain has been enhanced.
- ***Chapter 23 (Ultrafast Optics)***. New examples have been incorporated that con-

sider chirped pulse amplification in a petawatt laser and the generation of high-energy solitons in a photonic-crystal rod. A new section on high-harmonic generation and attosecond optics has been added. The section on pulse detection has been reorganized.

- ***Chapter 24 (Optical Interconnects and Switches)***. The role of optical interconnects at the inter-board, inter-chip, and intrachip scale of computer systems is delineated. All-optical switching now incorporates nonparametric and parametric photonic switches that operate on the basis of manifold nonlinear-optical effects. Photonic-crystal and plasmonic photonic switches are discussed. The treatment of photonic logic gates now includes an analysis of embedded bistable systems and examples of bistability in fiber-based-interferometric and microring laser systems.
- ***Chapter 25 (Optical Fiber Communications)***. The material on fiber-optic components has been updated and rewritten, and the role of photonic integrated circuits is delineated. A new section on space-division multiplexing in multicore and multimode fibers has been added. The section on coherent detection has been expanded and now emphasizes digital coherent receivers with spectrally efficient coding.

Representative Courses

The different chapters of the book may be combined in various ways for use in courses of semester or quarter duration. Representative examples of such courses are presented below. Some of these courses may be offered as part of a sequence. Other selections may be made to suit the particular objectives of instructors and students.

Optics

1. Ray Optics	8. Metals & Metamaterials	14. Light and Matter	20. Acousto-Optics
2. Wave Optics	9. Guided-Wave Optics	15. Laser Amplifiers	21. Electro-Optics
3. Beam Optics	10. Fiber Optics	16. Lasers	22. Nonlinear Optics
4. Fourier Optics	11. Resonator Optics	17. Semiconductor Optics	23. Ultrafast Optics
5. Electromagnetic Optics	12. Statistical Optics	18. LEDs & Laser Diodes	24. Interconnects/Switches
6. Polarization Optics	13. Photon Optics	19. Photodetectors	25. Fiber Communications
7. Photonic Crystals			

The first six chapters of the book are suitable for an introductory course on *Optics*. These may be supplemented by Chapter 12 (*Statistical Optics*) to introduce incoherent and partially coherent light, and by Chapter 13 (*Photon Optics*) to introduce the photon. The introductory sections of Chapters 9 and 10 (*Guided-Wave Optics* and *Fiber Optics*, respectively) may be added to cover some applications.

Guided-Wave Optics

1. Ray Optics	8. Metals & Metamaterials	14. Light and Matter	20. Acousto-Optics
2. Wave Optics	9. Guided-Wave Optics	15. Laser Amplifiers	21. Electro-Optics
3. Beam Optics	10. Fiber Optics	16. Lasers	22. Nonlinear Optics
4. Fourier Optics	11. Resonator Optics	17. Semiconductor Optics	23. Ultrafast Optics
5. Electromagnetic Optics	12. Statistical Optics	18. LEDs & Laser Diodes	24. Interconnects/Switches
6. Polarization Optics	13. Photon Optics	19. Photodetectors	25. Fiber Communications
7. Photonic Crystals			

A course on *Guided-Wave Optics* might begin with an introduction to wave propagation in layered and periodic media in Chapter 7 (*Photonic-Crystal Optics*), and could include Chapter 8 (*Metal and Metamaterial Optics*). This would be followed by Chapters 9, 10, and 11 (*Guided-Wave Optics*, *Fiber Optics*, and *Resonator Optics*, respectively). The introductory sections of Chapters 21 and 24 (*Electro-Optics* and *Optical Interconnects and Switches*) would provide additional material.

Lasers

1. Ray Optics	8. Metals & Metamaterials	14. Light and Matter	20. Acousto-Optics
2. Wave Optics	9. Guided-Wave Optics	15. Laser Amplifiers	21. Electro-Optics
3. Beam Optics	10. Fiber Optics	16. Lasers	22. Nonlinear Optics
4. Fourier Optics	11. Resonator Optics	17. Semiconductor Optics	23. Ultrafast Optics
5. Electromagnetic Optics	12. Statistical Optics	18. LEDs & Laser Diodes	24. Interconnects/Switches
6. Polarization Optics	13. Photon Optics	19. Photodetectors	25. Fiber Communications
7. Photonic Crystals			

A course on *Lasers* could begin with *Beam Optics* and *Resonator Optics* (Chapters 3 and 11, respectively), followed by *Light and Matter* (Chapter 14). The initial portion of *Photon Optics* (Chapter 13) could be assigned. The heart of the course would be the material contained in *Laser Amplifiers* and *Lasers* (Chapters 15 and 16, respectively). The course might also include material drawn from *Semiconductor Optics* and *LEDs and Laser Diodes* (Chapters 17 and 18, respectively). An introduction to femtosecond lasers could be extracted from some sections of *Ultrafast Optics* (Chapter 23).

Optoelectronics

1. Ray Optics	8. Metals & Metamaterials	14. Light and Matter	20. Acousto-Optics
2. Wave Optics	9. Guided-Wave Optics	15. Laser Amplifiers	21. Electro-Optics
3. Beam Optics	10. Fiber Optics	16. Lasers	22. Nonlinear Optics
4. Fourier Optics	11. Resonator Optics	17. Semiconductor Optics	23. Ultrafast Optics
5. Electromagnetic Optics	12. Statistical Optics	18. LEDs & Laser Diodes	24. Interconnects/Switches
6. Polarization Optics	13. Photon Optics	19. Photodetectors	25. Fiber Communications
7. Photonic Crystals			

The chapters on *Semiconductor Optics*, *LEDs and Laser Diodes*, and *Photodetectors* (Chapters 17, 18, and 19, respectively) form a suitable basis for a course on *Optoelectronics*. This material would be supplemented with optics background from earlier chapters and could include topics such as liquid-crystal devices (Secs. 6.5 and 21.3), electroabsorption modulators (Sec. 21.5), and an introduction to photonic devices used for switching and/or communications (Chapters 24 and 25, respectively).

Photonic Devices

1. Ray Optics	8. Metals & Metamaterials	14. Light and Matter	20. Acousto-Optics
2. Wave Optics	9. Guided-Wave Optics	15. Laser Amplifiers	21. Electro-Optics
3. Beam Optics	10. Fiber Optics	16. Lasers	22. Nonlinear Optics
4. Fourier Optics	11. Resonator Optics	17. Semiconductor Optics	23. Ultrafast Optics
5. Electromagnetic Optics	12. Statistical Optics	18. LEDs & Laser Diodes	24. Interconnects/Switches
6. Polarization Optics	13. Photon Optics	19. Photodetectors	25. Fiber Communications
7. Photonic Crystals			

Photonic Devices is a course that would consider the devices used in *Acousto-Optics*, *Electro-Optics*, and *Nonlinear Optics* (Chapters 20, 21, and 22, respectively). It might also include devices used in optical routing and switching, as discussed in *Optical Interconnects and Switches* (Chapter 24).

Nonlinear & Ultrafast Optics

1. Ray Optics	8. Metals & Metamaterials	14. Light and Matter	20. Acousto-Optics
2. Wave Optics	9. Guided-Wave Optics	15. Laser Amplifiers	21. Electro-Optics
3. Beam Optics	10. Fiber Optics	16. Lasers	22. Nonlinear Optics
4. Fourier Optics	11. Resonator Optics	17. Semiconductor Optics	23. Ultrafast Optics
5. Electromagnetic Optics	12. Statistical Optics	18. LEDs & Laser Diodes	24. Interconnects/Switches
6. Polarization Optics	13. Photon Optics	19. Photodetectors	25. Fiber Communications
7. Photonic Crystals			

The material contained in Chapters 21–23 (*Electro-Optics*, *Nonlinear Optics*, and *Ultrafast Optics*, respectively) is suitable for an in-depth course on *Nonlinear and Ultrafast Optics*. These chapters

could be supplemented by the material pertaining to electro-optic and all-optical switching in Chapter 24 (*Optical Interconnects and Switches*).

Fiber-Optic Communications

1. Ray Optics	8. Metals & Metamaterials	14. Light and Matter	20. Acousto-Optics
2. Wave Optics	9. Guided-Wave Optics	15. Laser Amplifiers	21. Electro-Optics
3. Beam Optics	10. Fiber Optics	16. Lasers	22. Nonlinear Optics
4. Fourier Optics	11. Resonator Optics	17. Semiconductor Optics	23. Ultrafast Optics
5. Electromagnetic Optics	12. Statistical Optics	18. LEDs & Laser Diodes	24. Interconnects/Switches
6. Polarization Optics	13. Photon Optics	19. Photodetectors	25. Fiber Communications
7. Photonic Crystals			

The heart of a course on *Fiber-Optic Communications* would be the material contained in Chapter 25 (*Optical Fiber Communications*). Background for this course would comprise material drawn from Chapters 9, 10, 18, and 19 (*Guided-Wave Optics*, *Fiber Optics*, *LEDs and Laser Diodes*, and *Photodetectors*, respectively), along with material contained in Secs. 15.3C and 15.3D (doped-fiber and Raman fiber amplifiers, respectively). If fiber-optic networks were to be emphasized, Sec. 24.3 (photonic switches) would be a valuable adjunct.

Optical Information Processing

1. Ray Optics	8. Metals & Metamaterials	14. Light and Matter	20. Acousto-Optics
2. Wave Optics	9. Guided-Wave Optics	15. Laser Amplifiers	21. Electro-Optics
3. Beam Optics	10. Fiber Optics	16. Lasers	22. Nonlinear Optics
4. Fourier Optics	11. Resonator Optics	17. Semiconductor Optics	23. Ultrafast Optics
5. Electromagnetic Optics	12. Statistical Optics	18. LEDs & Laser Diodes	24. Interconnects/Switches
6. Polarization Optics	13. Photon Optics	19. Photodetectors	25. Fiber Communications
7. Photonic Crystals			

Background material for a course on *Optical Information Processing* would be drawn from *Wave Optics* and *Beam Optics* (Chapters 2 and 3, respectively). The course could cover coherent image formation and processing from *Fourier Optics* (Chapter 4) along with incoherent and partially coherent imaging from *Statistical Optics* (Chapter 12). The focus could then shift to devices used for analog data processing, such as those considered in *Acousto-Optics* (Chapter 20). The course could then finish with material on switches and gates used for digital data processing, such as those considered in *Optical Interconnects and Switches* (Chapter 24).

Acknowledgments

We are indebted to many colleagues for providing us with valuable suggestions regarding improvements for the *Third Edition*: Rodrigo Amezcua-Correa, Luca Argenti, Joe C. Campbell, Zenghu Chang, Demetrios Christodoulides, Peter J. Delfyett, Dirk Englund, Eric R. Fossum, Majeed M. Hayat, Pieter G. Kik, Akhlesh Lakhtakia, Guifang Li, Steven B. Lowen, M. G. "Jim" Moharam, Rüdiger Paschotta, Kosmas L. Tsakmakidis, Shin-Tson Wu, Timothy M. Yarnall, and Boris Y. Zeldovich. We are also grateful to many of our former students and postdoctoral associates who posed excellent questions that helped us hone our presentation in the *Third Edition*, including John David Giese, Barry D. Jacobson, Samik Mukherjee, Adam Palmer, and Jian Yin.

We extend our special thanks to Mark Feuer, Joseph W. Goodman, and Mohammed F. Saleh who graciously provided us with in-depth critiques of various chapters.

Amy Hendrickson provided invaluable assistance with the Latex style files and eBook formatting. We are grateful to our Editors at John Wiley & Sons, Inc., who offered valuable suggestions and support throughout the course of production: Brett Kurzmann, Sarah Keegan, Nick Prindle, and Melissa Yanuzzi.

Finally, we are most appreciative of the generous support provided by CREOL, the College of Optics & Photonics at the University of Central Florida, the Boston University Photonics Center, and the Boston University College of Engineering.

Photo Credits

Many of the images on the chapter opening pages were carried forward from the First and Second Editions. Additional credits for the chapter opening pages of the *Third Edition* include: Wikimedia Commons (Laguerre and Bessel in Chapter 3, Stokes in Chapter 6, Drude in Chapter 8, Tyndall in Chapter 9, and Born in Chapter 12); *La Science Illustrée*, Volume 4 of the French weekly published in the second period of 1889 (Colladon in Chapter 9); Courtesy of Corning Incorporated (Schultz, Keck, & Maurer in Chapter 10); Courtesy of the Faculty History Project at Bentley Historical Library, University of Michigan (Franken in Chapter 22); and Courtesy of Peg Skorpinski, photographer, and the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley (Kaminow in Chapter 25). Self-photographs were provided by Viktor Georgievich Veselago (Veselago in Chapter 8); Sir John Pendry (Pendry in Chapter 8); Paul B. Corkum (Corkum in Chapter 23); Gérard Mourou (Mourou in Chapter 23); and the authors.

BAHAA E. A. SALEH

Orlando, Florida

MALVIN CARL TEICH

Boston, Massachusetts
June 4, 2018

PREFACE TO THE SECOND EDITION

Since the publication of the *First Edition* in 1991, *Fundamentals of Photonics* has been reprinted some 20 times, translated into Czech and Japanese, and used worldwide as a textbook and reference. During this period, major developments in photonics have continued apace, and have enabled technologies such as telecommunications and applications in industry and medicine. The *Second Edition* reports some of these developments, while maintaining the size of this single-volume tome within practical limits.

In its new organization, *Fundamentals of Photonics* continues to serve as a self-contained and up-to-date introductory-level textbook, featuring a logical blend of theory and applications. Many readers of the *First Edition* have been pleased with its abundant and well-illustrated figures. This feature has been enhanced in the *Second Edition* by the introduction of full color throughout the book, offering improved clarity and readability.

While each of the 22 chapters of the *First Edition* has been thoroughly updated, the principal feature of the *Second Edition* is the addition of two new chapters: one on photonic-crystal optics and another on ultrafast optics. These deal with developments that have had a substantial and growing impact on photonics over the past decade.

The new chapter on **photonic-crystal optics** provides a foundation for understanding the optics of layered media, including Bragg gratings, with the help of a matrix approach. Propagation of light in one-dimensional periodic media is examined using Bloch modes with matrix and Fourier methods. The concept of the photonic bandgap is introduced. Light propagation in two- and three-dimensional photonic crystals, and the associated dispersion relations and bandgap structures, are developed. Sections on photonic-crystal waveguides, holey fibers, and photonic-crystal resonators have also been added at appropriate locations in other chapters.

The new chapter on **ultrafast optics** contains sections on picosecond and femtosecond optical pulses and their characterization, shaping, and compression, as well as their propagation in optical fibers, in the domain of linear optics. Sections on ultrafast nonlinear optics include pulsed parametric interactions and optical solitons. Methods for the detection of ultrafast optical pulses using available detectors, which are relatively slow, are reviewed.

In addition to these two new chapters, the chapter on **optical interconnects and switches** has been completely rewritten and supplemented with topics such as wavelength and time routing and switching, FBGs, WGRs, SOAs, TOADs, and packet switches. The chapter on **optical fiber communications** has also been significantly updated and supplemented with material on WDM networks; it now offers concise descriptions of topics such as dispersion compensation and management, optical amplifiers, and soliton optical communications.

Continuing advances in device-fabrication technology have stimulated the emergence of **nanophotonics**, which deals with optical processes that take place over subwavelength (nanometer) spatial scales. Nanophotonic devices and systems include quantum-confined structures, such as quantum dots, nanoparticles, and nanoscale periodic structures used to synthesize *metamaterials* with exotic optical properties such as negative refractive index. They also include configurations in which light (or its interaction with matter) is confined to nanometer-size (rather than micrometer-size) regions near boundaries, as in *surface plasmon* optics. Evanescent fields, such as those produced at a surface where total internal reflection occurs, also exhibit such

confinement. Evanescent fields are present in the immediate vicinity of subwavelength-size apertures, such as the open tip of a tapered optical fiber. Their use allows imaging with resolution beyond the diffraction limit and forms the basis of *near-field optics*. Many of these emerging areas are described at suitable locations in the *Second Edition*.

New sections have been added in the process of updating the various chapters. New topics introduced in the early chapters include: Laguerre–Gaussian beams; near-field imaging; the Sellmeier equation; fast and slow light; optics of conductive media and plasmonics; doubly negative metamaterials; the Poincaré sphere and Stokes parameters; polarization mode dispersion; whispering-gallery modes; microresonators; optical coherence tomography; and photon orbital angular momentum.

In the chapters on laser optics, new topics include: rare-earth and Raman fiber amplifiers and lasers; EUV, X-ray, and free-electron lasers; and chemical and random lasers. In the area of optoelectronics, new topics include: gallium nitride-based structures and devices; superluminescent diodes; organic and white-light LEDs; quantum-confined lasers; quantum cascade lasers; microcavity lasers; photonic-crystal lasers; array detectors; low-noise APDs; SPADs; and QWIPs.

The chapter on nonlinear optics has been supplemented with material on parametric-interaction tuning curves; quasi-phase-matching devices; two-wave mixing and cross-phase modulation; THz generation; and other nonlinear optical phenomena associated with narrow optical pulses, including chirp pulse amplification and supercontinuum light generation. The chapter on electro-optics now includes a discussion of electroabsorption modulators.

Appendix C on **modes of linear systems** has been expanded and now offers an overview of the concept of modes as they appear in numerous locations within the book. Finally, additional exercises and problems have been provided, and these are now numbered disjointly to avoid confusion.

Acknowledgments

We are grateful to many colleagues for providing us with valuable comments about draft chapters for the *Second Edition* and for drawing our attention to errors in the *First Edition*: Mete Atatüre, Michael Bär, Robert Balahura, Silvia Carrasco, Stephen Chinn, Thomas Daly, Gianni Di Giuseppe, Adel El-Nadi, John Fourkas, Majeed Hayat, Tony Heinz, Erich Ippen, Martin Jaspan, Gerd Keiser, Jonathan Kane, Paul Kelley, Ted Moustakas, Allen Mullins, Magued Nasr, Roy Olivier, Roberto Paiella, Peter W. E. Smith, Stephen P. Smith, Kenneth Suslick, Anna Swan, Tristan Tayag, Tommaso Toffoli, and Brad Tousley.

We extend our special thanks to those colleagues who graciously provided us with in-depth critiques of various chapters: Ayman Abouraddy, Luca Dal Negro, and Paul Prucnal.

We are indebted to the legions of students and postdoctoral associates who have posed so many excellent questions that helped us hone our presentation. In particular, many improvements were initiated by suggestions from Mark Booth, Jasper Cabalu, Michael Cunha, Darryl Goode, Chris LaFratta, Rui Li, Eric Lynch, Nan Ma, Nishant Mohan, Julie Praino, Yunjie Tong, and Ranjith Zachariah. We are especially grateful to Mohammed Saleh, who diligently read much of the manuscript and provided us with excellent suggestions for improvement throughout.

Wai Yan (Eliza) Wong provided logistical support and a great deal of assistance in crafting diagrams and figures. Many at Wiley, including George Telecki, our Editor, and Rachel Witmer have been most helpful, patient, and encouraging. We appreciate the attentiveness and thoroughness that Melissa Yanuzzi brought to the production process. Don DeLand of the Integre Technical Publishing Company provided invaluable assistance in setting up the Latex style files.

We are most appreciative of the financial support provided by the National Science Foundation (NSF), in particular the Center for Subsurface Sensing and Imaging Systems (CenSSIS), an NSF-supported Engineering Research Center; the Defense Advanced Research Projects Agency (DARPA); the National Reconnaissance Office (NRO); the U.S. Army Research Office (ARO); the David & Lucile Packard Foundation; the Boston University College of Engineering; and the Boston University Photonics Center.

Photo Credits. Most of the portraits were carried forward from the First Edition with the benefit of permissions provided for all editions. Additional credits are: Godfrey Kneller 1689 portrait (Newton); Siegfried Bendixen 1828 lithograph (Gauss); Engraving in the Small Portraits Collection, History of Science Collections, University of Oklahoma Libraries (Fraunhofer); Stanford University, Courtesy AIP Emilio Segrè Visual Archives (Bloch); Eli Yablonovitch (Yablonovitch); Sajeev John (John); Charles Kuen Kao (Kao); Philip St John Russell (Russell); Ecole Polytechnique (Fabry); Observatoire des Sciences de l'Univers (Perot); AIP Emilio Segrè Visual Archives (Born); Lagrelius & Westphal 1920 portrait (Bohr); AIP Emilio Segrè Visual Archives, Weber Collection (W. L. Bragg); Linn F. Mollenauer (Mollenauer); Roger H. Stolen (Stolen); and James P. Gordon (Gordon). In Chapter 24, the Bell Symbol was reproduced with the permission of BellSouth Intellectual Property Marketing Corporation, the AT&T logo is displayed with the permission of AT&T, and Lucent Technologies permitted us use of their logo. Stephen G. Eick kindly provided the image used at the beginning of Chapter 25. The photographs of Saleh and Teich were provided courtesy of Boston University.

BAHAA E. A. SALEH

MALVIN CARL TEICH

Boston, Massachusetts
December 19, 2006

PREFACE TO THE FIRST EDITION

Optics is an old and venerable subject involving the generation, propagation, and detection of light. Three major developments, which have been achieved in the last thirty years, are responsible for the rejuvenation of optics and for its increasing importance in modern technology: the invention of the laser, the fabrication of low-loss optical fibers, and the introduction of semiconductor optical devices. As a result of these developments, new disciplines have emerged and new terms describing these disciplines have come into use: **electro-optics**, **optoelectronics**, **quantum electronics**, **quantum optics**, and **lightwave technology**. Although there is a lack of complete agreement about the precise usages of these terms, there is a general consensus regarding their meanings.

Photonics

Electro-optics is generally reserved for optical devices in which electrical effects play a role (lasers, and electro-optic modulators and switches, for example). *Optoelectronics*, on the other hand, typically refers to devices and systems that are essentially electronic in nature but involve light (examples are light-emitting diodes, liquid-crystal display devices, and array photodetectors). The term *quantum electronics* is used in connection with devices and systems that rely principally on the interaction of light with matter (lasers and nonlinear optical devices used for optical amplification and wave mixing serve as examples). Studies of the quantum and coherence properties of light lie within the realm of *quantum optics*. The term *lightwave technology* has been used to describe devices and systems that are used in optical communications and optical signal processing.

In recent years, the term **photonics** has come into use. This term, which was coined in analogy with electronics, reflects the growing tie between optics and electronics forged by the increasing role that semiconductor materials and devices play in optical systems. *Electronics* involves the control of electric-charge flow (in vacuum or in matter); *photonics* involves the control of photons (in free space or in matter). The two disciplines clearly overlap since electrons often control the flow of photons and, conversely, photons control the flow of electrons. The term *photonics* also reflects the importance of the photon nature of light in describing the operation of many optical devices.

Scope

This book provides an introduction to the fundamentals of photonics. The term *photonics* is used broadly to encompass all of the aforementioned areas, including the following:

- The *generation* of coherent light by lasers, and incoherent light by luminescence sources such as light-emitting diodes.
- The *transmission* of light in free space, through conventional optical components such as lenses, apertures, and imaging systems, and through waveguides such as optical fibers.
- The *modulation*, switching, and scanning of light by the use of electrically, acoustically, or optically controlled devices.
- The *amplification* and *frequency conversion* of light by the use of wave interactions in nonlinear materials.
- The *detection* of light.

These areas have found ever-increasing applications in optical communications, signal processing, computing, sensing, display, printing, and energy transport.

Approach and Presentation

The underpinnings of photonics are provided in a number of chapters that offer concise introductions to:

- The four theories of light (each successively more advanced than the preceding): ray optics, wave optics, electromagnetic optics, and photon optics.
- The theory of interaction of light with matter.
- The theory of semiconductor materials and their optical properties.

These chapters serve as basic building blocks that are used in other chapters to describe the *generation* of light (by lasers and light-emitting diodes); the *transmission* of light (by optical beams, diffraction, imaging, optical waveguides, and optical fibers); the *modulation* and switching of light (by the use of electro-optic, acousto-optic, and nonlinear-optic devices); and the *detection* of light (by means of photodetectors). Many applications and examples of real systems are provided so that the book is a blend theory and practice. The final chapter is devoted to the study of fiber-optic communications, which provides an especially rich example in which the generation, transmission, modulation, and detection of light are all part of a single photonic system used for the transmission of information.

The theories of light are presented at progressively increasing levels of difficulty. Thus light is described first as rays, then scalar waves, then electromagnetic waves, and finally, photons. Each of these descriptions has its domain of applicability. Our approach is to draw from the simplest theory that adequately describes the phenomenon or intended application. Ray optics is therefore used to describe imaging systems and the confinement of light in waveguides and optical resonators. Scalar wave theory provides a description of optical beams, which are essential for the understanding of lasers, and of Fourier optics, which is useful for describing coherent optical systems and holography. Electromagnetic theory provides the basis for the polarization and dispersion of light, and the optics of guided waves, fibers, and resonators. Photon optics serves to describe the interactions of light with matter, explaining such processes as light generation and detection, and light mixing in nonlinear media.

Intended Audience

Fundamentals of Photonics is meant to serve as:

- An introductory textbook for students in electrical engineering or applied physics at the senior or first-year graduate level.
- A self-contained work for self-study.
- A text for programs of continuing professional development offered by industry, universities, and professional societies.

The reader is assumed to have a background in engineering or applied physics, including courses in modern physics, electricity and magnetism, and wave motion. Some knowledge of linear systems and elementary quantum mechanics is helpful but not essential. Our intent has been to provide an introduction to photonics that emphasizes the concepts governing applications of current interest. The book should, therefore, not be considered as a compendium that encompasses all photonic devices and systems. Indeed, some areas of photonics are not included at all, and many of the individual chapters could easily have been expanded into separate monographs.

Problems, Reading Lists, and Appendices

A set of problems is provided at the end of each chapter. Problems are numbered in accordance with the chapter sections to which they apply. Quite often, problems deal with ideas or applications not mentioned in the text, analytical derivations, and numerical computations designed to illustrate the magnitudes of important quantities. Problems marked with asterisks are of a more advanced nature. A number of exercises also appear within the text of each chapter to help the reader develop a better understanding of (or to introduce an extension of) the material.

Appendices summarize the properties of one- and two-dimensional Fourier transforms, linear-systems theory, and modes of linear systems (which are important in polarization devices, optical waveguides, and resonators); these are called upon at appropriate points throughout the book. Each chapter ends with a reading list that includes a selection of important books, review articles, and a few classic papers of special significance.

Acknowledgments

We are grateful to many colleagues for reading portions of the text and providing helpful comments: Govind P. Agrawal, David H. Auston, Rasheed Azzam, Nikolai G. Basov, Franco Cerrina, Emmanuel Desurvire, Paul Diament, Eric Fossum, Robert J. Keyes, Robert H. Kingston, Rodney Loudon, Leonard Mandel, Leon McCaughan, Richard M. Osgood, Jan Peřina, Robert H. Rediker, Arthur L. Schawlow, S. R. Seshadri, Henry Stark, Ferrel G. Stremler, John A. Tataronis, Charles H. Townes, Patrick R. Trischitta, Wen I. Wang, and Edward S. Yang.

We are especially indebted to John Whinnery and Emil Wolf for providing us with many suggestions that greatly improved the presentation.

Several colleagues used portions of the notes in their classes and provided us with invaluable feedback. These include Etan Bourkoff at Johns Hopkins University (now at the University of South Carolina), Mark O. Freeman at the University of Colorado, George C. Papen at the University of Illinois, and Paul R. Prucnal at Princeton University.

Many of our students and former students contributed to this material in various ways over the years and we owe them a great debt of thanks: Gaetano L. Aiello, Mohamad Asi, Richard Campos, Buddy Christyono, Andrew H. Cordes, Andrew David, Ernesto Fontenla, Evan Goldstein, Matthew E. Hansen, Dean U. Hekel, Conor Heneghan, Adam Heyman, Bradley M. Jost, David A. Landgraf, Kanghua Lu, Ben Nathanson, Winslow L. Sargeant, Michael T. Schmidt, Raul E. Sequeira, David Small, Kraisin Songwatana, Nikola S. Subotic, Jeffrey A. Tobin, and Emily M. True. Our thanks also go to the legions of unnamed students who, through a combination of vigilance and the desire to understand the material, found countless errors.

We particularly appreciate the many contributions and help of those students who were intimately involved with the preparation of this book at its various stages of completion: Niraj Agrawal, Suzanne Keilson, Todd Larchuk, Guifang Li, and Philip Tham.

We are grateful for the assistance given to us by a number of colleagues in the course of collecting the photographs used at the beginnings of the chapters: E. Scott Barr, Nicolaas Bloembergen, Martin Carey, Marjorie Graham, Margaret Harrison, Ann Kottner, G. Thomas Holmes, John Howard, Theodore H. Maiman, Edward Palik, Martin Parker, Aleksandr M. Prokhorov, Jarus Quinn, Lesley M. Richmond, Claudia Schüler, Patrick R. Trischitta, J. Michael Vaughan, and Emil Wolf. Specific photo credits are as follows: AIP Meggers Gallery of Nobel Laureates (Gabor, Townes, Basov, Prokhorov, W. L. Bragg); AIP Niels Bohr Library (Rayleigh, Fraunhofer, Maxwell, Planck, Bohr, Einstein in Chapter 14, W. H. Bragg); Archives de l'Académie des Sciences de Paris (Fabry); The Astrophysical Journal (Perot);

AT&T Bell Laboratories (Shockley, Brattain, Bardeen); Bettmann Archives (Young, Gauss, Tyndall); Bibliothèque Nationale de Paris (Fermat, Fourier, Poisson); Burndy Library (Newton, Huygens); Deutsches Museum (Hertz); ETH Bibliothek (Einstein in Chapter 13); Bruce Fritz (Saleh); Harvard University (Bloembergen); Heidelberg University (Pockels); Kelvin Museum of the University of Glasgow (Kerr); Theodore H. Maiman (Maiman); Princeton University (von Neumann); Smithsonian Institution (Fresnel); Stanford University (Schawlow); Emil Wolf (Born, Wolf). Corning Incorporated kindly provided the photograph used at the beginning of Chapter 10. We are grateful to GE for the use of their logotype, which is a registered trademark of the General Electric Company, at the beginning of Chapter 18. The IBM logo at the beginning of Chapter 18 is being used with special permission from IBM. The right-most logotype at the beginning of Chapter 18 was supplied courtesy of Lincoln Laboratory, Massachusetts Institute of Technology. AT&T Bell Laboratories kindly permitted us use of the diagram at the beginning of Chapter 25.

We greatly appreciate the continued support provided to us by the National Science Foundation, the Center for Telecommunications Research, and the Joint Services Electronics Program through the Columbia Radiation Laboratory.

Finally, we extend our sincere thanks to our editors, George Telecki and Bea Shube, for their guidance and suggestions throughout the course of preparation of this book.

BAHAA E. A. SALEH

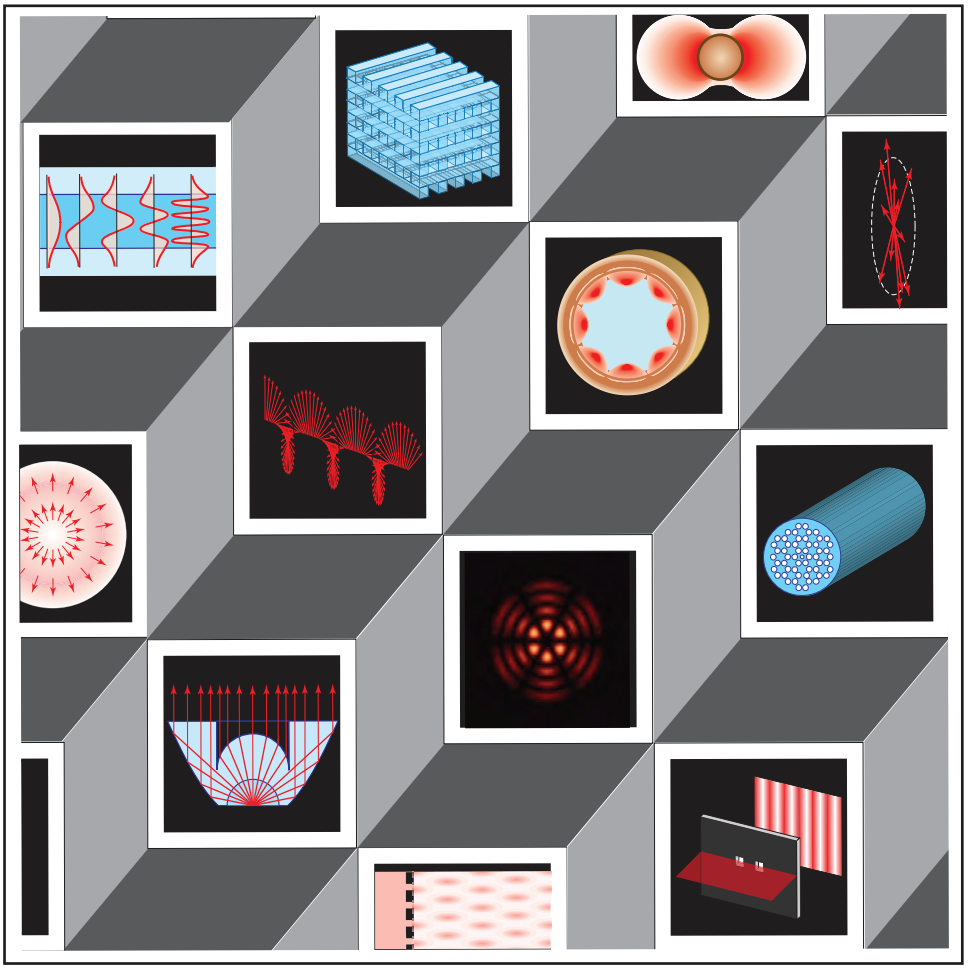
Madison, Wisconsin

MALVIN CARL TEICH

New York, New York
April 3, 1991

FUNDAMENTALS OF PHOTONICS

Part I: Optics
(Chapters 1–13)



CHAPTER

1

RAY OPTICS

1.1 POSTULATES OF RAY OPTICS 5

1.2 SIMPLE OPTICAL COMPONENTS 8
- A. Mirrors
- B. Planar Boundaries
- C. Spherical Boundaries and Lenses
- D. Light Guides

1.3 GRADED-INDEX OPTICS 20
- A. The Ray Equation
- B. Graded-Index Optical Components
- *C. The Eikonal Equation

1.4 MATRIX OPTICS 27
- A. The Ray-Transfer Matrix
- B. Matrices of Simple Optical Components
- C. Matrices of Cascaded Optical Components
- D. Periodic Optical Systems

Sir Isaac Newton (1642–1727) set forth a theory of optics in which light emissions consist of collections of corpuscles that propagate rectilinearly.

Pierre de Fermat (1601–1665) enunciated a rule, known as Fermat's Principle, in which light rays travels along the path of least time relative to neighboring paths.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Light can be described as an electromagnetic wave phenomenon governed by the same theoretical principles that govern all other forms of electromagnetic radiation, such as radio waves and X-rays. This conception of light is called **electromagnetic optics**. Electromagnetic radiation propagates in the form of two mutually coupled *vector* waves, an electric-field wave and a magnetic-field wave. Nevertheless, it is possible to describe many optical phenomena using a simplified *scalar* wave theory in which light is described by a single scalar wavefunction. This approximate way of treating light is called scalar wave optics, or simply **wave optics**.

When light waves propagate through and around objects whose dimensions are much greater than the wavelength of the light, the wave nature is not readily discerned and the behavior of light can be adequately described by rays obeying a set of geometrical rules. This model of light is called **ray optics**. From a mathematical perspective, ray optics is the limit of wave optics when the wavelength is infinitesimally small.

Thus, electromagnetic optics encompasses wave optics, which in turn encompasses ray optics, as illustrated in Fig. 1.0-1. Ray optics and wave optics are approximate theories that derive validity from their successes in producing results that approximate those based on the more rigorous electromagnetic theory.

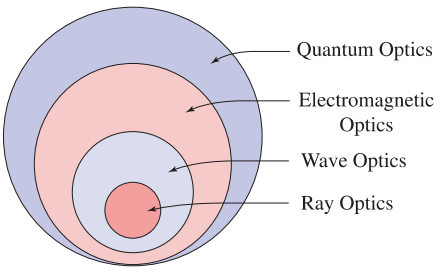


Figure 1.0-1 The theory of quantum optics provides an explanation for virtually all optical phenomena. The electromagnetic theory of light (electromagnetic optics) provides the most complete treatment of light within the confines of classical optics. Wave optics is a scalar approximation of electromagnetic optics. Ray optics is the limit of wave optics when the wavelength is very short.

Although electromagnetic optics provides the most complete treatment of light within the confines of **classical optics**, certain optical phenomena are characteristically quantum mechanical in nature and cannot be explained classically. These nonclassical phenomena are described by a quantum version of electromagnetic theory known as **quantum electrodynamics**. For optical phenomena, this theory is also referred to as **quantum optics**.

Historically, the theories of optics developed roughly in the following order: (1) ray optics $\rightarrow$ (2) wave optics $\rightarrow$ (3) electromagnetic optics $\rightarrow$ (4) quantum optics. These models are progressively more complex and sophisticated, and were developed successively to provide explanations for the outcomes of increasingly subtle and precise optical experiments. The optimal choice of a model is the simplest one that satisfactorily describes a particular phenomenon, but it is sometimes difficult to know *a priori* which model achieves this. Experience is often the best guide.

For pedagogical reasons, the initial chapters of this book follow the historical order indicated above. Each model of light begins with a set of postulates (provided without proof), from which a large body of results are generated. The postulates of each model are shown to arise as special cases of the next-higher-level model. In this chapter we begin with ray optics.

This Chapter

Ray optics is the simplest theory of light. Light is described by rays that travel in different optical media in accordance with a set of geometrical rules. Ray optics is therefore also called **geometrical optics**. Ray optics is an approximate theory. Although it adequately describes most of our daily experiences with light, there are many phenomena that ray optics cannot adequately construe (as amply attested to by the remaining chapters of this book).

Ray optics is concerned with the *locations* and *directions* of light rays. It is therefore useful in studying *image formation* — the collection of rays from each point of an object and their redirection by an optical component onto a corresponding point of an image. Ray optics permits us to determine the conditions under which light is guided within a given medium, such as a glass fiber. In isotropic media, optical rays point in the direction of the flow of *optical energy*. Ray bundles can be constructed in which the density of rays is proportional to the density of light energy. When light is generated isotropically from a point source, for example, the energy associated with the rays in a given cone is proportional to the solid angle of the cone. Rays may be traced through an optical system to determine the optical energy crossing a given area.

This chapter begins with a set of postulates from which we derive the simple rules that govern the propagation of light rays through optical media. In Sec. 1.2 these rules are applied to simple optical components, such as mirrors and planar or spherical boundaries between different optical media. Ray propagation in inhomogeneous (graded-index) optical media is examined in Sec. 1.3. Graded-index optics is the basis of a technology that has become an important part of modern optics.

Optical components are often centered about an optical axis, with respect to which the rays travel at small inclinations. Such rays are called **paraxial rays** and the assumption that the rays have this property is the basis of **paraxial optics**. The change in the position and inclination of a paraxial ray as it travels through an optical system can be efficiently described by the use of a 2×2-matrix algebra. Section 1.4 is devoted to this algebraic tool, which is known as **matrix optics**.

1.1 POSTULATES OF RAY OPTICS

Postulates of Ray Optics

- Light travels in the form of rays. The rays are emitted by light sources and can be observed when they reach an optical detector.
- An optical medium is characterized by a quantity $n \geq 1$, called the **refractive index**. The refractive index $n = c_o/c$ where c_o is the speed of light in free space and c is the speed of light in the medium. Therefore, the time taken by light to travel a distance d is $d/c = nd/c_o$. It is proportional to the product nd, which is known as the **optical pathlength**.
- In an inhomogeneous medium the refractive index $n(\mathbf{r})$ is a function of the position $\mathbf{r} = (x, y, z)$. The optical pathlength along a given path between two points A and B is therefore

$$\text{Optical pathlength} = \int_A^B n(\mathbf{r})\, ds, \tag{1.1-1}$$

where ds is the differential element of length along the path. The time taken by light to travel from A to B is proportional to the optical pathlength.

- **Fermat's Principle**. Optical rays traveling between two points, A and B, follow a path such that the time of travel (or the optical pathlength) between the two points is an extremum relative to neighboring paths. This is expressed mathematically as

$$\delta \int_A^B n(\mathbf{r})\, ds = 0, \tag{1.1-2}$$

where the symbol δ, which is read "the variation of," signifies that the optical pathlength is either minimized or maximized, or is a point of inflection. It is, however, usually a minimum, in which case:

Light rays travel along the path of least time.

Sometimes the minimum time is shared by more than one path, which are then all followed simultaneously by the rays. An example in which the pathlength is maximized is provided in Prob. 1.1-2.

In this chapter we use the postulates of ray optics to determine the rules governing the propagation of light rays, their reflection and refraction at the boundaries between different media, and their transmission through various optical components. A wealth of results applicable to numerous optical systems are obtained without the need for any other assumptions or rules regarding the nature of light.

Propagation in a Homogeneous Medium

In a homogeneous medium the refractive index is the same everywhere, and so is the speed of light. The path of minimum time, required by Fermat's principle, is therefore also the path of minimum distance. The principle of the *path of minimum distance* is known as **Hero's principle**. The path of minimum distance between two points is a straight line so that *in a homogeneous medium, light rays travel in straight lines* (Fig. 1.1-1).

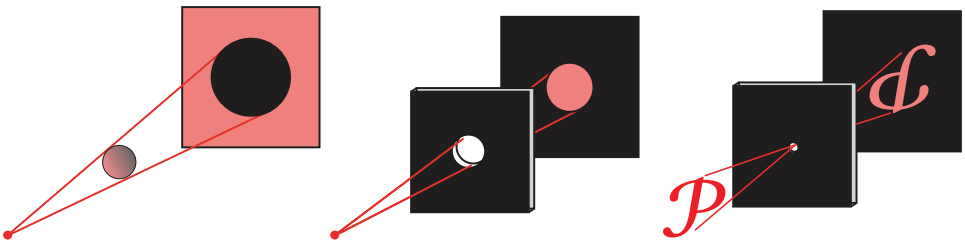


Figure 1.1-1 Light rays travel in straight lines. Shadows are perfect projections of stops.

Reflection from a Mirror

Mirrors are made of certain highly polished metallic surfaces, or metallic or dielectric films deposited on a substrate such as glass. Light reflects from mirrors in accordance with the law of reflection:

> *The reflected ray lies in the plane of incidence; the angle of reflection equals the angle of incidence.*

The plane of incidence is the plane formed by the incident ray and the normal to the mirror at the point of incidence. The angles of incidence and reflection, θ and θ', are defined in Fig. 1.1-2(a). To prove the law of reflection we simply use Hero's principle. Examine a ray that travels from point A to point C after reflection from the planar mirror in Fig. 1.1-2(b). According to Hero's principle, for a mirror of infinitesimal thickness, the distance $\overline{AB}+\overline{BC}$ must be minimum. If C' is a mirror image of C, then $\overline{BC} = \overline{BC'}$, so that $\overline{AB} + \overline{BC'}$ must be a minimum. This occurs when $\overline{ABC'}$ is a straight line, i.e., when B coincides with B' so that $\theta = \theta'$.

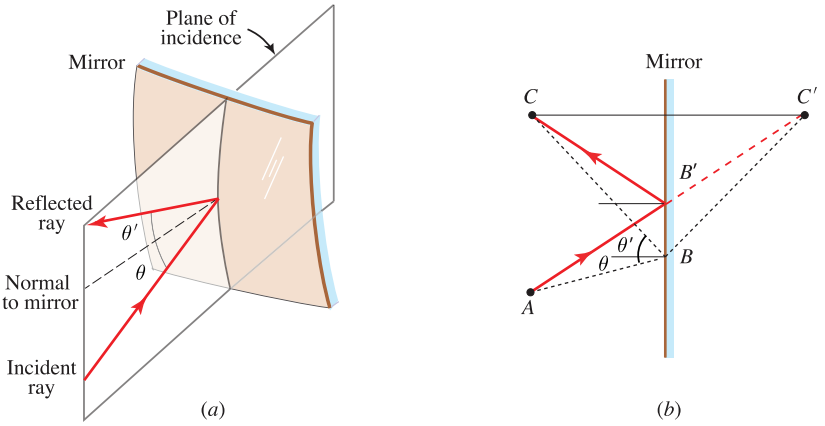


Figure 1.1-2 (a) Reflection from the surface of a curved mirror. (b) Geometrical construction to prove the law of reflection.

Reflection and Refraction at the Boundary Between Two Media

At the boundary between two media of refractive indices n_1 and n_2 an incident ray is split into two — a reflected ray and a refracted (or transmitted) ray (Fig. 1.1-3). The reflected ray obeys the law of reflection. The refracted ray obeys the law of refraction:

> *The refracted ray lies in the plane of incidence; the angle of refraction θ_2 is related to the angle of incidence θ_1 by Snell's law,*

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \tag{1.1-3}$$

Snell's Law

The proportion in which the light is reflected and refracted is not described by ray optics.

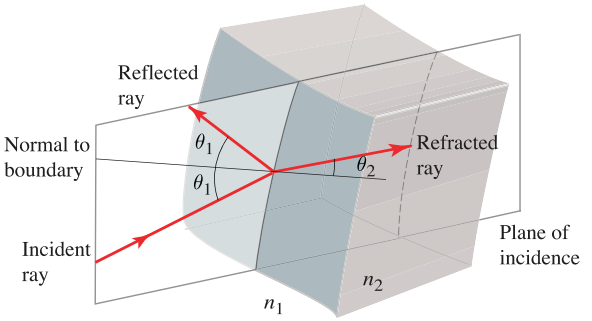


Figure 1.1-3 Reflection and refraction at the boundary between two media.

EXERCISE 1.1-1

Proof of Snell's Law. The proof of Snell's law is an exercise in the application of Fermat's principle. Referring to Fig. 1.1-4, we seek to minimize the optical pathlength $n_1\overline{AB}+n_2\overline{BC}$ between points A and C. We therefore have the following optimization problem: Minimize $n_1 d_1 \sec\theta_1 + n_2 d_2 \sec\theta_2$ with respect to the angles θ_1 and θ_2, subject to the condition $d_1 \tan\theta_1 + d_2 \tan\theta_2 = d$. Show that the solution of this constrained minimization problem yields Snell's law.

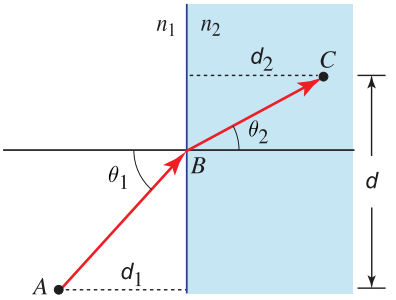


Figure 1.1-4 Construction to prove Snell's law.

The three simple rules — propagation in straight lines and the laws of reflection and refraction — are applied in Sec. 1.2 to several geometrical configurations of mirrors and transparent optical components, without further recourse to Fermat's principle.

1.2 SIMPLE OPTICAL COMPONENTS

A. Mirrors

Planar Mirrors

A planar mirror reflects the rays originating from a point P_1 such that the reflected rays appear to originate from a point P_2 behind the mirror, called the image (Fig. 1.2-1).

Paraboloidal Mirrors

The surface of a paraboloidal mirror is a reflective paraboloid of revolution. It has the useful property of focusing all incident rays parallel to its axis to a single point, called the **focus** or **focal point**. The distance $\overline{PF} \equiv f$ defined in Fig. 1.2-2 is known as

the **focal length**. Paraboloidal mirrors are often used as light-collecting elements in telescopes. They are also used to render parallel the rays from a point source of light, such as a flashlight bulb or a light-emitting diode, located at the focus. When used in this manner, the device is known as a **collimator**.

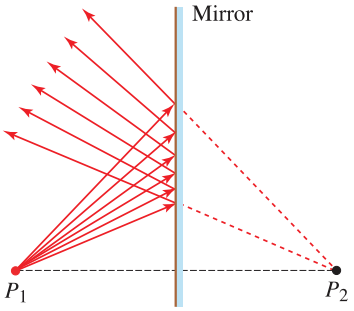


Figure 1.2-1 Reflection of light from a planar mirror.

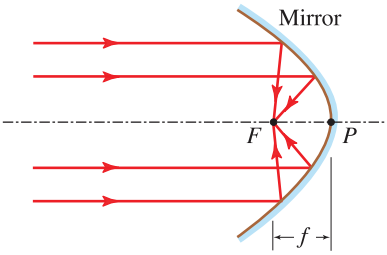


Figure 1.2-2 Focusing of light by a paraboloidal mirror.

Elliptical Mirrors

An elliptical mirror reflects all the rays emitted from one of its two foci, e.g., P_1, and images them onto the other focus, P_2 (Fig. 1.2-3). In accordance with Hero's principle, the distances traveled by the light from P_1 to P_2 along any of the paths are equal.

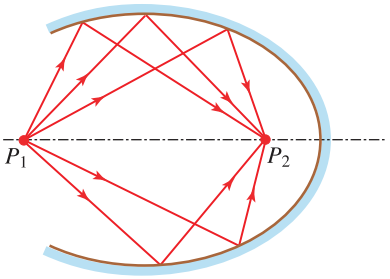


Figure 1.2-3 Reflection from an elliptical mirror.

Spherical Mirrors

A spherical mirror is easier to fabricate than a paraboloidal mirror or an elliptical mirror. However, it has neither the focusing property of the paraboloidal mirror nor the imaging property of the elliptical mirror. As illustrated in Fig. 1.2-4, parallel rays meet the axis at different points; their envelope (the dashed curve) is called the caustic curve. Nevertheless, parallel rays close to the axis are approximately focused onto a single point F at distance $(-R)/2$ from the mirror center C. By convention, the **radius of curvature** R is negative for concave mirrors and positive for convex mirrors.

Paraxial Rays Reflected from Spherical Mirrors

Rays that make small angles (such that $\sin\theta \approx \theta$) with the mirror's axis are called **paraxial rays**. In the **paraxial approximation**, where only paraxial rays are considered, a spherical mirror has a focusing property like that of the paraboloidal mirror *and* an imaging property like that of the elliptical mirror. The body of rules that results from this approximation forms **paraxial optics**, also called **first-order optics** or **Gaussian optics**.

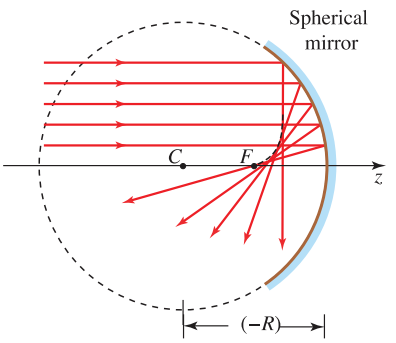


Figure 1.2-4 Reflection of parallel rays from a concave spherical mirror.

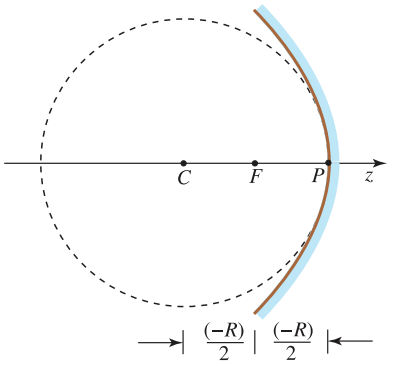


Figure 1.2-5 A spherical mirror approximates a paraboloidal mirror for paraxial rays.

A spherical mirror of radius R therefore acts like a paraboloidal mirror of focal length $f = R/2$. This is, in fact, plausible since at points near the axis, a parabola can be approximated by a circle with radius equal to the parabola's radius of curvature (Fig. 1.2-5).

All paraxial rays originating from each point on the axis of a spherical mirror are reflected and focused onto a single corresponding point on the axis. This can be seen (Fig. 1.2-6) by examining a ray emitted at an angle θ_1 from a point P_1 at a distance z_1 away from a concave mirror of radius R, and reflecting at angle $(-\theta_2)$ to meet the axis at a point P_2 that is a distance z_2 away from the mirror. The angle θ_2 is negative since the ray is traveling downward. Since the three angles of a triangle add to 180°, we have $\theta_1 = \theta_0 - \theta$ and $(-\theta_2) = \theta_0 + \theta$, so that $(-\theta_2) + \theta_1 = 2\theta_0$. If θ_0 is sufficiently small, the approximation $\tan\theta_0 \approx \theta_0$ may be used, so that $\theta_0 \approx y/(-R)$, from which

$$(-\theta_2) + \theta_1 \approx \frac{2y}{(-R)}\,, \tag{1.2-1}$$

where y is the height of the point at which the reflection occurs. Recall that R is negative since the mirror is concave. Similarly, if θ_1 and θ_2 are small, $\theta_1 \approx y/z_1$ and $(-\theta_2) = y/z_2$, so that (1.2-1) yields $y/z_1 + y/z_2 \approx 2y/(-R)$, whereupon

$$\frac{1}{z_1} + \frac{1}{z_2} \approx \frac{2}{(-R)}\,. \tag{1.2-2}$$

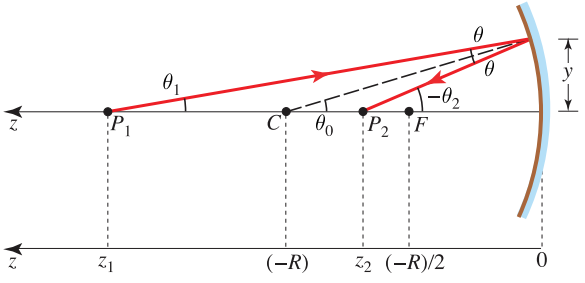


Figure 1.2-6 Reflection of paraxial rays from a concave spherical mirror of radius $R < 0$.

This relation holds regardless of y (i.e., regardless of θ_1) as long as the approximation is valid. This means that all paraxial rays originating from point P_1 arrive at P_2. The distances z_1 and z_2 are measured in a coordinate system in which the z axis points to the left. Points of negative z therefore lie to the right of the mirror.

According to (1.2-2), rays that are emitted from a point very far out on the z axis ($z_1 = \infty$) are focused to a point F at a distance $z_2 = (-R)/2$. This means that within the paraxial approximation, all rays coming from infinity (parallel to the axis of the mirror) are focused to a point at a distance f from the mirror, which is known as its focal length:

$$f = \frac{(-R)}{2}, \tag{1.2-3}$$

Focal Length
Spherical Mirror

Equation (1.2-2) is usually written in the form

$$\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}, \tag{1.2-4}$$

Imaging Equation
(Paraxial Rays)

which is known as the imaging equation. Both the incident and the reflected rays must be paraxial for this equation to hold.

EXERCISE 1.2-1

Image Formation by a Spherical Mirror. Show that, within the paraxial approximation, rays originating from a point $P_1 = (y_1, z_1)$ are reflected to a point $P_2 = (y_2, z_2)$, where z_1 and z_2 satisfy (1.2-4) and $y_2 = -y_1 z_2/z_1$ (Fig. 1.2-7). This means that rays from each point in the plane $z = z_1$ meet at a single corresponding point in the plane $z = z_2$, so that the mirror acts as an image-formation system with magnification $-z_2/z_1$. Negative magnification means that the image is inverted.

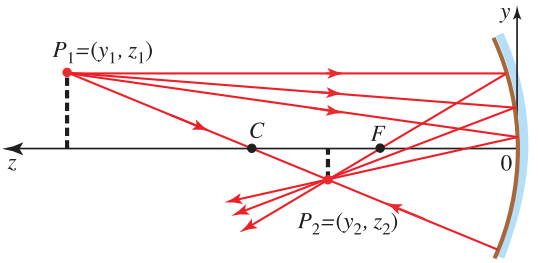


Figure 1.2-7 Image formation by a spherical mirror. Four particular rays are illustrated.

B. Planar Boundaries

The relation between the angles of refraction and incidence, θ_2 and θ_1, at a planar boundary between two media of refractive indices n_1 and n_2 is governed by Snell's law (1.1-3). This relation is plotted in Fig. 1.2-8 for two cases:

- *External Refraction* ($n_1 < n_2$). When the ray is incident from the medium of smaller refractive index, $\theta_2 < \theta_1$ and the refracted ray bends away from the boundary.

- *Internal Refraction* $(n_1 > n_2)$. If the incident ray is in a medium of higher refractive index, $\theta_2 > \theta_1$ and the refracted ray bends toward the boundary.

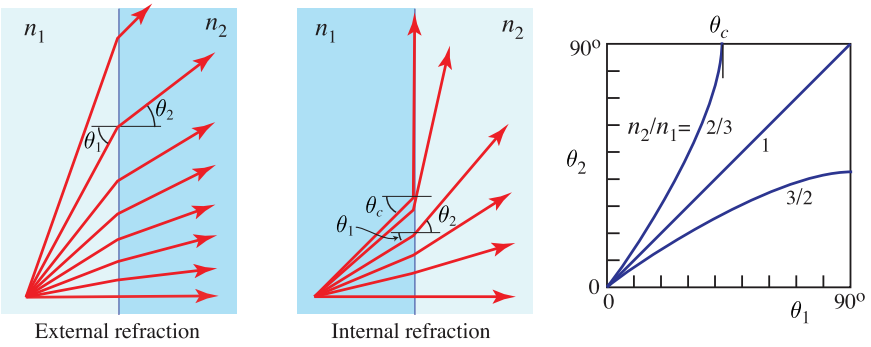


Figure 1.2-8 Relation between the angles of refraction and incidence.

The refracted rays bend in such a way as to minimize the optical pathlength, i.e., to increase the pathlength in the lower-index medium at the expense of pathlength in the higher-index medium. In both cases, when the angles are small (i.e., the rays are paraxial), the relation between θ_2 and θ_1 is approximately linear, $n_1\theta_1 \approx n_2\theta_2$, or $\theta_2 \approx (n_1/n_2)\theta_1$.

Total Internal Reflection

For internal refraction $(n_1 > n_2)$, the angle of refraction is greater than the angle of incidence, $\theta_2 > \theta_1$, so that as θ_1 increases, θ_2 reaches 90° when $\theta_1 = \theta_c$, the **critical angle** (see Fig. 1.2-8). This occurs when $n_1 \sin\theta_c = n_2 \sin(\pi/2) = n_2$, so that

$$\theta_c = \sin^{-1}\frac{n_2}{n_1}. \qquad (1.2\text{-}5)$$

Critical Angle

When $\theta_1 > \theta_c$, Snell's law (1.1-3) cannot be satisfied and refraction does not occur. The incident ray is then totally reflected as if the surface were a perfect mirror [Fig. 1.2-9(a)]. This phenomenon, called **total internal reflection (TIR)**, is the basis of many optical devices and systems, such as reflecting prisms [Fig. 1.2-9(b)], light-emitting diode collimators (Fig. 1.2-14), and optical fibers (Sec. 1.2D). Electromagnetic optics (Fresnel's equations in Chapter 6) reveals that all of the energy is carried by the reflected light so that the process of total internal reflection is highly efficient.

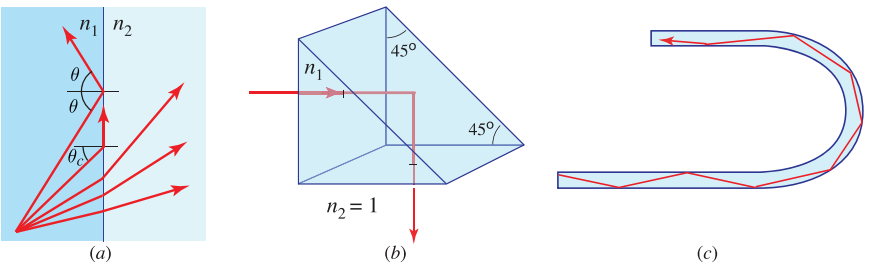


Figure 1.2-9 (a) Total internal reflection at a planar boundary. (b) The reflecting prism. If $n_1 > \sqrt{2}$ and $n_2 = 1$ (air), then $\theta_c < 45°$; since $n_2 \approx 1.5 > \sqrt{2}$ for glass, and $\theta_1 = 45°$, the ray is totally reflected. (c) Rays are guided by total internal reflection from the internal surface of an optical fiber.

Prisms

A prism of apex angle α and refractive index n (Fig. 1.2-10) deflects a ray incident at an angle θ by an angle

$$\theta_d = \theta - \alpha + \sin^{-1}\left[\sqrt{n^2 - \sin^2\theta}\ \sin\alpha - \sin\theta\cos\alpha\right]. \tag{1.2-6}$$

This equation is arrived at by using Snell's law twice, at the two refracting surfaces of the prism. When α is very small (thin prism) and θ is also very small (paraxial approximation), (1.2-6) may be approximated by

$$\theta_d \approx (n-1)\alpha. \tag{1.2-7}$$

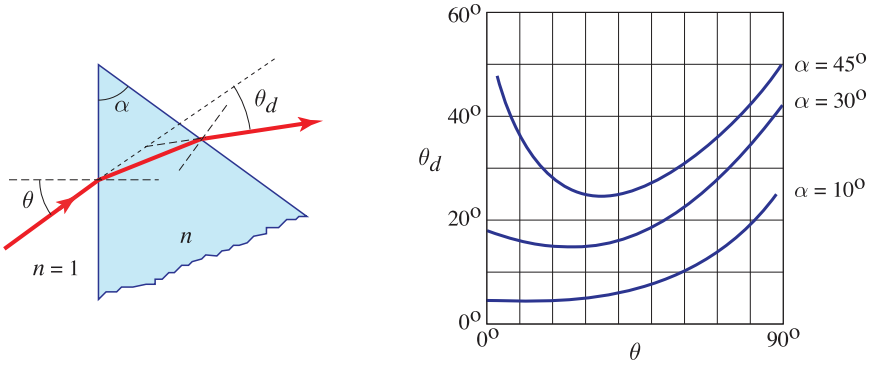
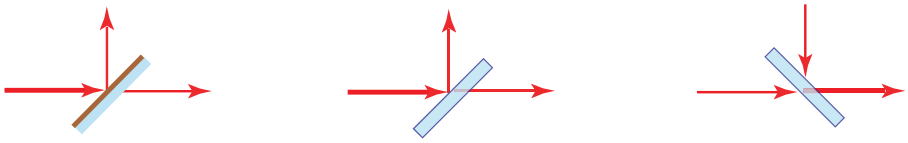


Figure 1.2-10 (a) Ray deflection by a prism. (b) Graph of (1.2-6) for the deflection angle θ_d as a function of the angle of incidence θ, for different apex angles α and $n = 1.5$. When both α and θ are small the angle of deflection $\theta_d \approx (n-1)\alpha$, which is approximately independent of θ, as is evident for the $\alpha = 10°$ curve. When $\theta = 0°$ and $\alpha = 45°$, total internal reflection occurs, as illustrated in Fig. 1.2-9(b).

Beamsplitters

The beamsplitter is an optical component that splits an incident ray into a reflected ray and a transmitted ray, as illustrated in Fig. 1.2-11. The relative proportions of light transmitted and reflected are established by Fresnel's equations in electromagnetic optics (Chapter 6). Beamsplitters are also frequently used to combine two light rays into one [Fig. 1.2-11(c)]. Beamsplitters are usually constructed by depositing a thin semitransparent metallic or dielectric film on a glass substrate. A thin bare glass plate, such as a microscope slide, can also serve as a beamsplitter although the fraction of light reflected is small. Transparent plastic materials are often used in place of glass.



(a) Partially reflective mirror (b) Thin glass plate (c) Beam combiner

Figure 1.2-11 Beamsplitters and beam combiners.

Beam Directors

Simple optical components can be used to direct rays in particular directions. The devices illustrated in Fig. 1.2-12 redirect incident rays into rays tilted at fixed angles with respect to each other. The **biprism** depicted in Fig. 1.2-12(a) is the juxtaposition of a prism and an identical inverted prism. The **Fresnel biprism** portrayed in Fig. 1.2-12(b) is formed from rows of adjacently placed tiny prisms. This device is equivalent to a biprism but is thinner and lighter. The cone-shaped optic depicted in Fig. 1.2-12(c), known as an **axicon**, converts incident rays into a collection of circularly symmetric rays directed toward its central axis in the form of a cone. It has the same cross section as the biprism, namely an isosceles triangle.

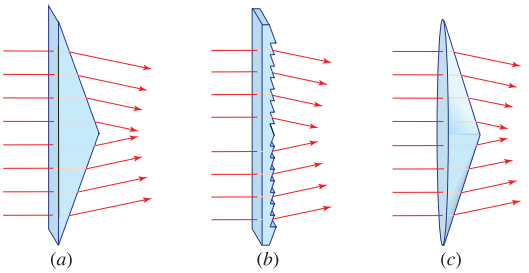


Figure 1.2-12 (a) Biprism. (b) Fresnel biprism. (c) Plano-convex axicon.

C. Spherical Boundaries and Lenses

We now examine the refraction of rays from a spherical boundary of radius R between two media of refractive indices n_1 and n_2. By convention, R is positive for a convex boundary and negative for a concave boundary. The results are obtained by applying Snell's law, which relates the angles of incidence and refraction relative to the normal to the surface, defined by the radius vector from the center C. These angles are to be distinguished from the angles θ_1 and θ_2, which are defined relative to the z axis. Considering only paraxial rays making small angles with the axis of the system so that $\sin\theta \approx \theta$ and $\tan\theta \approx \theta$, the following properties may be shown to hold:

- A ray making an angle θ_1 with the z axis and meeting the boundary at a point of height y where it makes an angle θ_0 with the radius vector [see Fig. 1.2-13(a)] changes direction at the boundary so that the refracted ray makes an angle θ_2 with the z axis and an angle θ_3 with the radius vector. With the help of Exercise 1.2-2, we obtain

$$\theta_2 \approx \frac{n_1}{n_2}\theta_1 - \frac{n_2 - n_1}{n_2}\frac{y}{R}. \tag{1.2-8}$$

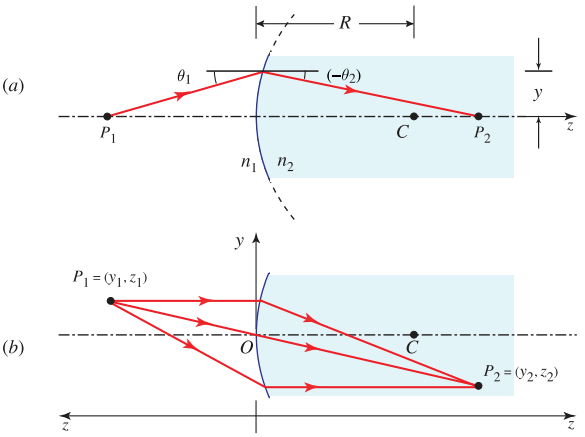


Figure 1.2-13 Refraction at a convex spherical boundary ($R > 0$).

- All paraxial rays originating from a point $P_1 = (y_1, z_1)$ in the $z = z_1$ plane meet at a point $P_2 = (y_2, z_2)$ in the $z = z_2$ plane (see Exercise 1.2-2), where

$$\frac{n_1}{z_1} + \frac{n_2}{z_2} \approx \frac{n_2 - n_1}{R} \tag{1.2-9}$$

and

$$y_2 = -\frac{n_1}{n_2}\frac{z_2}{z_1} y_1. \tag{1.2-10}$$

The $z = z_1$ and $z = z_2$ planes are said to be conjugate planes. Every point in the first plane has a corresponding point (image) in the second with magnification $-(n_1/n_2)(z_2/z_1)$. Again, negative magnification means that the image is inverted. By convention P_1 is measured in a coordinate system pointing to the left and P_2 in a coordinate system pointing to the right (e.g., if P_2 lies to the left of the boundary, then z_2 would be negative).

The similarities between these properties and those of the spherical mirror are evident. It is important to remember that the image formation properties described above are approximate. They hold only for paraxial rays. Rays of large angles do not obey these paraxial laws; the deviation results in image distortion called **aberration**.

EXERCISE 1.2-2

Image Formation. Derive (1.2-8). Prove that paraxial rays originating from P_1 pass through P_2 when (1.2-9) and (1.2-10) are satisfied.

EXERCISE 1.2-3

Aberration-Free Imaging Surface. Determine the equation of a convex aspherical (nonspherical) surface between media of refractive indices n_1 and n_2 such that all rays (not necessarily paraxial) from an axial point P_1 at a distance z_1 to the left of the surface are imaged onto an axial point P_2 at a distance z_2 to the right of the surface [Fig. 1.2-13(a)]. *Hint:* In accordance with Fermat's principle the optical pathlengths between the two points must be equal for all paths.

EXAMPLE 1.2-1. ***Collimator for LED Light.*** Light emitted by an LED (Sec. 18.1) is often collimated by making use of an optic whose surface takes the form of a paraboloid of revolution (Fig. 1.2-14). The LED is placed at the focus of the paraboloid by inserting its hemispherical dome (darker blue) into a recess formed in the narrow end of the optic. Rays emanating from the sides of the LED dome impinge on the paraboloidal boundary at angles of incidence greater than the critical angle and are thus reflected out of the device via total internal reflection. Rays emanating from the central portion of the LED dome are refracted out of the device at the spherical boundary. Optical systems that combine reflection and refraction are known as **catadioptric systems**.

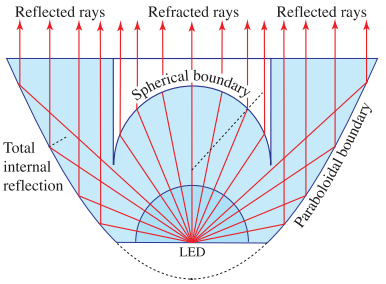


Figure 1.2-14 Cross section of a collimator for LED light. LED collimators come in many configurations but most make use of both total internal reflection and refraction to provide rays of light that are approximately parallel at the exit. Such devices are often fabricated from molded acrylic or polycarbonate plastic, which have refractive indices similar to that of glass ($n \approx 1.5$). The diameter of the narrow end of the optic illustrated is ≈ 1 cm.

Spherical Lenses

A **spherical lens** is bounded by two spherical surfaces. It is, therefore, defined completely by the radii R_1 and R_2 of its two surfaces, its thickness Δ, and the refractive index n of the material (Fig. 1.2-15). A glass lens in air can be regarded as a combination of two spherical boundaries, air-to-glass and glass-to-air.

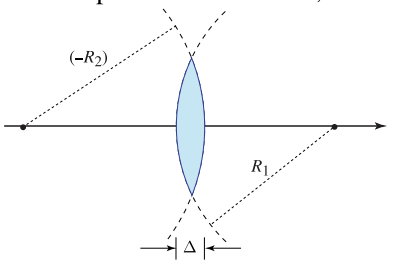


Figure 1.2-15 A biconvex spherical lens.

A ray crossing the first surface at height y and angle θ_1 with the z axis [Fig. 1.2-16(a)] is traced by applying (1.2-8) at the first surface to obtain the inclination angle θ of the refracted ray, which we extend until it meets the second surface. We then use (1.2-8) once more with θ replacing θ_1 to obtain the inclination angle θ_2 of the ray after refraction from the second surface. The results are in general complex. When the lens is thin, however, it can be assumed that the incident ray emerges from the lens at about the same height y at which it enters. Under this assumption, the following relations obtain:

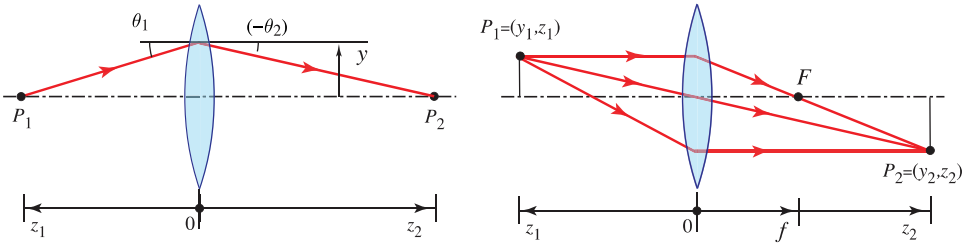
- The angles of the refracted and incident rays are related by (see Exercise 1.2-4)

$$\theta_2 = \theta_1 - \frac{y}{f}, \tag{1.2-11}$$

where f, called the **focal length**, is given by

$$\boxed{\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right).} \tag{1.2-12}$$

Focal Length
Thin Spherical Lens



(a) (b)

Figure 1.2-16 (a) Ray bending by a thin lens. (b) Image formation by a thin lens.

- All rays originating from a point $P_1 = (y_1, z_1)$ meet at a point $P_2 = (y_2, z_2)$ [Fig. 1.2-16(b)] (see Exercise 1.2-4), where

$$\boxed{\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}} \tag{1.2-13}$$

Imaging Equation

and

$$y_2 = -\frac{z_2}{z_1} y_1. \tag{1.2-14}$$

Magnification

These results are identical to those for the spherical mirror [see (1.2-4) and Exercise 1.2-1].

These equations indicate that each point in the $z = z_1$ plane is imaged onto a corresponding point in the $z = z_2$ plane with the magnification factor $-z_2/z_1$. The magnification is unity when $z_1 = z_2 = 2f$. The focal length f of a lens therefore completely determines its effect on paraxial rays. As indicated earlier, P_1 and P_2 are measured in coordinate systems pointing to the left and right, respectively, and the radii of curvatures R_1 and R_2 are positive for convex surfaces and negative for concave surfaces. For the biconvex lens shown in Fig. 1.2-15, R_1 is positive and R_2 is negative, so that the two terms of (1.2-12) add and provide a positive f.

EXERCISE 1.2-4

Proof of the Thin Lens Formulas. Using (1.2-8), along with the definition of the focal length given in (1.2-12), prove (1.2-11) and (1.2-13).

It is emphasized once again that the foregoing relations hold only for paraxial rays. The presence of nonparaxial rays results in aberrations, as illustrated in Fig. 1.2-17.

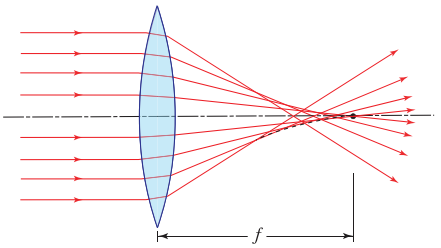


Figure 1.2-17 Nonparaxial rays do not meet at the paraxial focus. The dashed envelope of the refracted rays is known as the caustic curve.

Convex and Concave Lenses

Lenses are transparent optical devices that bend rays in a manner prescribed by the shapes of their surfaces. Most common lenses, such as the biconvex lens considered above, are spherical lenses. Lenses that consist of a single piece of material (glass and plastic are favored in the visible) are called simple lenses, while lenses that comprise multiple simple lenses, usually along a common axis, are known as compound lenses.

The surface of a lens can be convex or concave, depending on whether it projects out of, or recedes into the body of the lens, respectively, or it can be planar, indicating that it has a flat surface. A **cylindrical lens** is curved in only one direction; it thus has a focal length f for rays in the y–z plane, and no focusing power for rays in the x–z plane. A lens in which one surface is convex and the other concave is called a meniscus lens (these are often used for spectacles). A lens in which one or both surfaces have a shape that is neither spherical nor cylindrical is known as an **aspheric lens**.

Several different types of lenses are illustrated in Fig. 1.2-18. Biconvex and plano-convex lenses result in the convergence of rays and are useful for image formation, as depicted in Fig. 1.2-16. Biconcave and plano-concave lenses lead to the divergence of rays and are used in projection and focal-length expansion. A **Fresnel lens** is constructed by removing the nonrefracting portions of a conventional lens. Hence, the Fresnel-lens equivalent [Fig. 1.2-18(e)] of a plano-convex lens [Fig. 1.2-18(b)] is a flattened set of concentric surfaces with identical curvature at all locations on the surface (except at the stepwise discontinuities). The Fresnel design allows for the construction of thin, light, and inexpensive plastic lenses with sizes that range from meters to micrometers and short focal lengths. Fresnel lenses can be converging, diverging, or cylindrical.

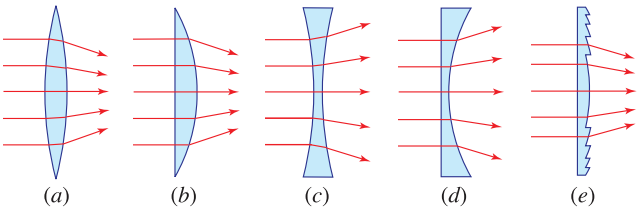


Figure 1.2-18 Lenses: (a) Biconvex; (b) Plano-convex; (c) Concave; (d) Plano-concave. (e) Fresnel-lens counterpart of the plano-convex lens displayed in (b); the curvatures are the same everywhere on the two surfaces.

D. Light Guides

Light may be guided from one location to another by use of a set of lenses or mirrors, as illustrated schematically in Fig. 1.2-19. Since refractive elements (such as lenses) are usually partially reflective and since mirrors are partially absorptive, the cumulative loss of optical power will be significant when the number of guiding elements is large. Components in which these effects are minimized can be fabricated (e.g., antireflection-coated lenses), but the system is generally cumbersome and costly.

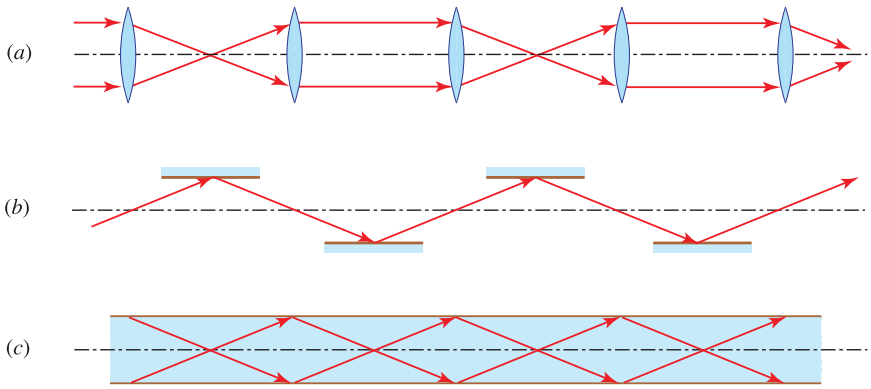


Figure 1.2-19 Guiding light: (a) lenses; (b) mirrors; (c) total internal reflection.

An ideal mechanism for guiding light is that of total internal reflection at the boundary between two media of different refractive indices. Rays are reflected repeatedly without undergoing refraction. Glass fibers of high chemical purity are used to guide light for tens of kilometers with relatively low loss of optical power.

An optical fiber is a light conduit made of two concentric glass (or plastic) cylinders (Fig. 1.2-20). The inner, called the core, has a refractive index n_1, and the outer, called

the cladding, has a slightly smaller refractive index, $n_2 < n_1$. Light rays traveling in the core are totally reflected from the cladding if their angle of incidence is greater than the critical angle, $\overline{\theta} > \theta_c = \sin^{-1}(n_2/n_1)$. The rays making an angle $\theta = 90° - \overline{\theta}$ with the optical axis are therefore confined in the fiber core if $\theta < \overline{\theta_c}$, where $\overline{\theta_c} = 90° - \theta_c = \cos^{-1}(n_2/n_1)$. Optical fibers are used in optical communication systems (see Chapters 10 and 25). Some important properties of optical fibers are derived in Exercise 1.2-5.

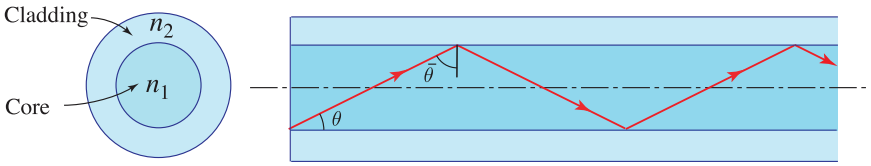


Figure 1.2-20 The optical fiber. Light rays are guided by multiple total internal reflections. Here θ represents the angle measured from the axis of the optical fiber so that its complement $\overline{\theta} = 90° - \theta$ is the angle of incidence at the dielectric interface.

EXERCISE 1.2-5

Numerical Aperture and Angle of Acceptance of an Optical Fiber. An optical fiber is illuminated by light from a source (e.g., a light-emitting diode, LED). The refractive indices of the core and cladding of the fiber are n_1 and n_2, respectively, and the refractive index of air is 1 (see Fig. 1.2-21). Show that the half-angle θ_a of the cone of rays accepted by the fiber (transmitted through the fiber without undergoing refraction at the cladding) is given by

$$\text{NA} = \sin\theta_a = \sqrt{n_1^2 - n_2^2}\,. \tag{1.2-15}$$

Numerical Aperture
Optical Fiber

The angle θ_a is called the **acceptance angle** and the parameter $\text{NA} \equiv \sin\theta_a$ is known as the **numerical aperture** of the fiber. Calculate the numerical aperture and acceptance angle for a silica-glass fiber with $n_1 = 1.475$ and $n_2 = 1.460$. Silica glass, also known as fused silica, is amorphous silicon dioxide (SiO_2). It is widely used because of its excellent optical and mechanical properties. Moreover, its refractive index can be readily modified by doping (e.g., with GeO_2).

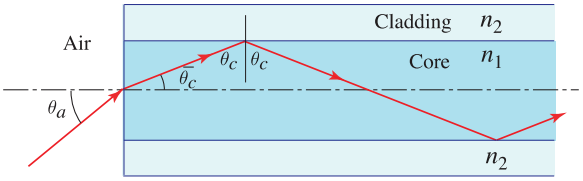


Figure 1.2-21 Acceptance angle of an optical fiber.

Trapping of Light in Media of High Refractive Index

It is often difficult for light originating inside a medium of large refractive index to be extracted into air, especially if the surfaces of the medium are parallel. This occurs since certain rays undergo multiple total internal reflections without ever refracting into air. The principle is illustrated in Exercise 1.2-6.

EXERCISE 1.2-6

Light Trapped in a Light-Emitting Diode.

(a) Assume that light is generated in all directions inside a material of refractive index n cut in the shape of a parallelepiped (Fig. 1.2-22). The material is surrounded by air with unity refractive index. This process occurs in light-emitting diodes (see Sec. 18.1B). What is the angle of the cone of light rays (inside the material) that will emerge from each face? What happens to the other rays? What is the numerical value of this angle for GaAs ($n = 3.6$)?

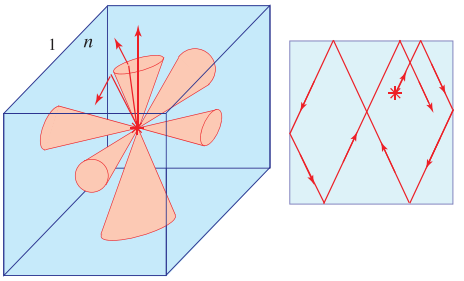


Figure 1.2-22 Trapping of light in a parallelepiped of high refractive index.

(b) Assume that when light is generated isotropically the amount of optical power associated with the rays in a given cone is proportional to the solid angle of the cone. Show that the ratio of the optical power that is extracted from the material to the total generated optical power is $3\left(1-\sqrt{1-1/n^2}\right)$, provided that $n > \sqrt{2}$. What is the numerical value of this ratio for GaAs?

1.3 GRADED-INDEX OPTICS

A graded-index (GRIN) material has a refractive index that varies with position in accordance with a continuous function $n(\mathbf{r})$. These materials are often fabricated by adding impurities (dopants) of controlled concentrations. In a GRIN medium the optical rays follow curved trajectories, instead of straight lines. By appropriate choice of $n(\mathbf{r})$, a GRIN plate can have the same effect on light rays as a conventional optical component, such as a prism or lens.

A. The Ray Equation

To determine the trajectories of light rays in an inhomogeneous medium with refractive index $n(\mathbf{r})$, we use Fermat's principle,

$$\delta \int_A^B n(\mathbf{r})\, ds = 0, \tag{1.3-1}$$

where ds is a differential length along the ray trajectory between A and B. If the trajectory is described by the function $x(s)$, $y(s)$, and $z(s)$, where s is the length of the trajectory (Fig. 1.3-1), then using the calculus of variations it can be shown that[†] $x(s)$,

[†] See, e.g., R. Weinstock, *Calculus of Variations: With Applications to Physics and Engineering*, 1952; Dover, 1974.

$y(s)$, and $z(s)$ must satisfy three partial differential equations,

$$\frac{d}{ds}\left(n\frac{dx}{ds}\right)=\frac{\partial n}{\partial x},\qquad \frac{d}{ds}\left(n\frac{dy}{ds}\right)=\frac{\partial n}{\partial y},\qquad \frac{d}{ds}\left(n\frac{dz}{ds}\right)=\frac{\partial n}{\partial z}. \tag{1.3-2}$$

By defining the vector $\mathbf{r}(s)$, whose components are $x(s)$, $y(s)$, and $z(s)$, (1.3-2) may be written in the compact vector form

$$\boxed{\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right)=\nabla n,} \tag{1.3-3}$$

(1.3-3) Ray Equation

where ∇n, the gradient of n, is a vector with Cartesian components $\partial n/\partial x$, $\partial n/\partial y$, and $\partial n/\partial z$. Equation (1.3-3) is known as the **ray equation**.

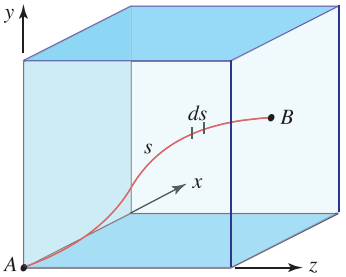


Figure 1.3-1 The ray trajectory is described parametrically by three functions $x(s)$, $y(s)$, and $z(s)$, or by two functions $x(z)$ and $y(z)$.

One approach to solving the ray equation is to describe the trajectory by two functions $x(z)$ and $y(z)$, write $ds=dz\sqrt{1+(dx/dz)^2+(dy/dz)^2}$, and substitute in (1.3-3) to obtain two partial differential equations for $x(z)$ and $y(z)$. The algebra is generally not trivial, but it simplifies considerably when the paraxial approximation is used.

The Paraxial Ray Equation

In the paraxial approximation, the trajectory is almost parallel to the z axis, so that $ds\approx dz$ (Fig. 1.3-2). The ray equations (1.3-2) then simplify to

$$\boxed{\frac{d}{dz}\left(n\frac{dx}{dz}\right)\approx\frac{\partial n}{\partial x},\qquad \frac{d}{dz}\left(n\frac{dy}{dz}\right)\approx\frac{\partial n}{\partial y}.} \tag{1.3-4}$$

(1.3-4) Paraxial Ray Equations

Given $n=n(x,y,z)$, these two partial differential equations may be solved for the trajectory $x(z)$ and $y(z)$.

In the limiting case of a homogeneous medium for which n is independent of x, y, z, (1.3-4) gives $d^2x/dz^2=0$ and $d^2y/dz^2=0$, from which it follows that x and y are linear functions of z, so that the trajectories are straight lines. More interesting cases will be examined subsequently.

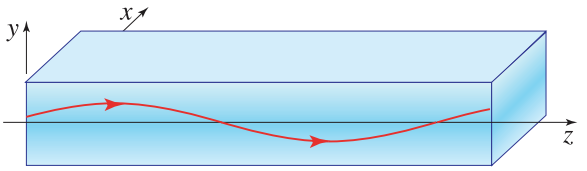


Figure 1.3-2 Trajectory of a paraxial ray in a graded-index medium.

B. Graded-Index Optical Components

Graded-Index Slab

Consider a slab of material whose refractive index $n = n(y)$ is uniform in the x and z directions but varies continuously in the y direction (Fig. 1.3-3). The trajectories of paraxial rays in the y–z plane are described by the paraxial ray equation

$$\frac{d}{dz}\left(n\frac{dy}{dz}\right) = \frac{dn}{dy}, \tag{1.3-5}$$

from which

$$\frac{d^2y}{dz^2} = \frac{1}{n(y)}\frac{dn(y)}{dy}. \tag{1.3-6}$$

Given $n(y)$ and initial conditions (y and dy/dz at $z = 0$), (1.3-6) can be solved for the function $y(z)$, which describes the ray trajectories.

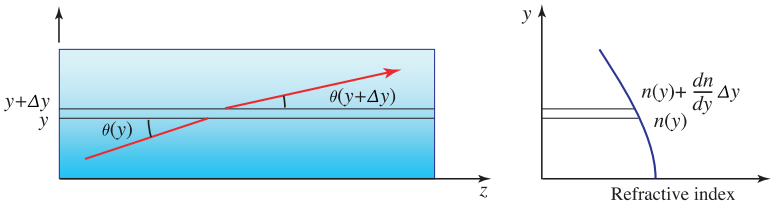


Figure 1.3-3 Refraction in a graded-index slab.

□ **Derivation of the Paraxial Ray Equation in a Graded-Index Slab Using Snell's Law.** Equation (1.3-6) may also be derived by the direct use of Snell's law (Fig. 1.3-3). Let $\theta(y) \approx dy/dz$ be the angle that the ray makes with the z axis at the position (y, z). After traveling through a layer of thickness Δy the ray changes its angle to $\theta(y + \Delta y)$. The two angles are related by Snell's law where θ, as defined in Fig. 1.3-3, is the complement of the angle of incidence (refraction):

$$\begin{aligned} n(y)\cos\theta(y) &= n(y+\Delta y)\cos\theta(y+\Delta y) \\ &= \left[n(y) + \frac{dn}{dy}\Delta y\right]\left[\cos\theta(y) - \frac{d\theta}{dy}\Delta y\,\sin\theta(y)\right], \end{aligned} \tag{1.3-7}$$

where we have applied the expansion $f(y+\Delta y) = f(y) + (df/dy)\,\Delta y$ to the functions $f(y) = n(y)$ and $f(y) = \cos\theta(y)$. In the limit $\Delta y \to 0$, after eliminating the term in $(\Delta y)^2$, we obtain the differential equation

$$\frac{dn}{dy} = n\frac{d\theta}{dy}\tan\theta. \tag{1.3-8}$$

For paraxial rays θ is very small so that $\tan\theta \approx \theta$. Substituting $\theta = dy/dz$ in (1.3-8), we obtain (1.3-6). ■

EXAMPLE 1.3-1. ***Slab with Parabolic Index Profile.*** An important particular distribution for the graded refractive index is

$$n^2(y) = n_0^2\left(1 - \alpha^2 y^2\right). \tag{1.3-9}$$

This is a symmetric function of y that has its maximum value at $y = 0$ (Fig. 1.3-4). A glass slab with this profile is known by the trade name SELFOC. Usually, α is chosen to be sufficiently small so that $\alpha^2 y^2 \ll 1$ for all y of interest. Under this condition, $n(y) = n_0\sqrt{1 - \alpha^2 y^2} \approx n_0(1 - \frac{1}{2}\alpha^2 y^2)$; i.e., $n(y)$ is a parabolic distribution. Also, because $n(y) - n_0 \ll n_0$, the fractional change of the refractive index is very small. Taking the derivative of (1.3-9), the right-hand side of (1.3-6) is $(1/n)dn/dy = -(n_0/n)^2\alpha^2 y \approx -\alpha^2 y$, so that (1.3-6) becomes

$$\frac{d^2y}{dz^2} \approx -\alpha^2 y. \tag{1.3-10}$$

The solutions of this equation are harmonic functions with period $2\pi/\alpha$. Assuming an initial position $y(0) = y_0$ and an initial slope $dy/dz = \theta_0$ at $z = 0$ inside the GRIN medium,

$$y(z) = y_0 \cos \alpha z + \frac{\theta_0}{\alpha} \sin \alpha z, \tag{1.3-11}$$

from which the slope of the trajectory is

$$\theta(z) = \frac{dy}{dz} = -y_0 \alpha \sin \alpha z + \theta_0 \cos \alpha z. \tag{1.3-12}$$

The ray oscillates about the center of the slab with a period (distance) $2\pi/\alpha$ known as the **pitch**, as illustrated in Fig. 1.3-4.

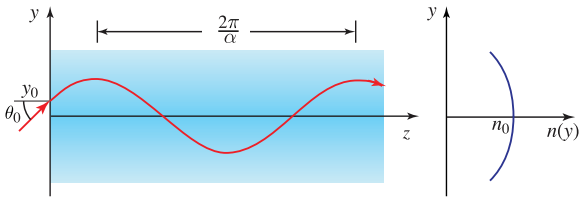


Figure 1.3-4 Trajectory of a ray in a GRIN slab of parabolic index profile (SELFOC).

The maximum excursion of the ray is $y_{\max} = \sqrt{y_0^2 + (\theta_0/\alpha)^2}$ and the maximum angle is $\theta_{\max} = \alpha y_{\max}$. The validity of this approximate analysis is ensured if $\theta_{\max} \ll 1$. If $2y_{\max}$ is smaller than the thickness of the slab, the ray remains confined and the slab serves as a light guide. Figure 1.3-5 shows the trajectories of a number of rays transmitted through a SELFOC slab. Note that all rays have the same pitch. This GRIN slab may be used as a lens, as demonstrated in Exercise 1.3-1.

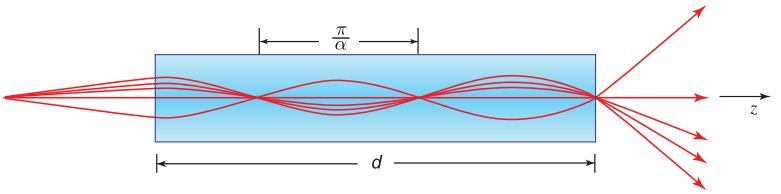


Figure 1.3-5 Trajectories of rays from an external point source in a SELFOC slab.

EXERCISE 1.3-1

The GRIN Slab as a Lens. Show that a SELFOC slab of length $d < \pi/2\alpha$ and refractive index given by (1.3-9) acts as a cylindrical lens (a lens with focusing power in the y–z plane) of focal length

$$f \approx \frac{1}{n_0 \alpha \sin(\alpha d)}. \tag{1.3-13}$$

Show that the principal point (defined in Fig. 1.3-6) lies at a distance from the slab edge $\overline{AH} \approx (1/n_0\alpha)\tan(\alpha d/2)$. Sketch the ray trajectories in the special cases $d = \pi/\alpha$ and $\pi/2\alpha$.

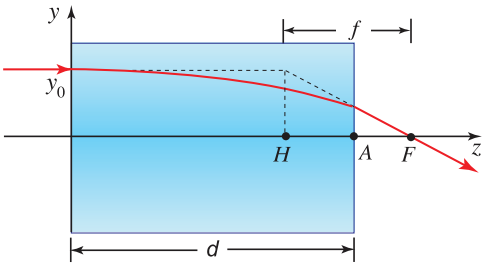


Figure 1.3-6 The SELFOC slab used as a lens; F is the focal point and H is the principal point.

Graded-Index Fibers

A graded-index fiber is a glass cylinder with a refractive index n that varies as a function of the radial distance from its axis. In the paraxial approximation, the ray trajectories are governed by the paraxial ray equations (1.3-4). Consider, for example, the distribution

$$n^2 = n_0^2\left[1 - \alpha^2\left(x^2 + y^2\right)\right]. \tag{1.3-14}$$

Substituting (1.3-14) into (1.3-4) and assuming that $\alpha^2(x^2 + y^2) \ll 1$ for all x and y of interest, we obtain

$$\frac{d^2x}{dz^2} \approx -\alpha^2 x, \qquad \frac{d^2y}{dz^2} \approx -\alpha^2 y. \tag{1.3-15}$$

Both x and y are therefore harmonic functions of z with period $2\pi/\alpha$. The initial positions (x_0, y_0) and angles ($\theta_{x0} = dx/dz$ and $\theta_{y0} = dy/dz$) at $z = 0$ determine the amplitudes and phases of these harmonic functions. Because of the circular symmetry, there is no loss of generality in choosing $x_0 = 0$. The solution of (1.3-15) is then

$$\begin{aligned} x(z) &= \frac{\theta_{x0}}{\alpha}\sin\alpha z \\ y(z) &= \frac{\theta_{y0}}{\alpha}\sin\alpha z + y_0\cos\alpha z. \end{aligned} \tag{1.3-16}$$

If $\theta_{x0} = 0$, i.e., the incident ray lies in a meridional plane (a plane passing through the axis of the cylinder, in this case the y–z plane), the ray continues to lie in that plane following a sinusoidal trajectory similar to that in the GRIN slab [Fig. 1.3-7(a)].

On the other hand, if $\theta_{y0} = 0$, and $\theta_{x0} = \alpha y_0$, then

$$\begin{aligned} x(z) &= y_0\sin\alpha z \\ y(z) &= y_0\cos\alpha z, \end{aligned} \tag{1.3-17}$$

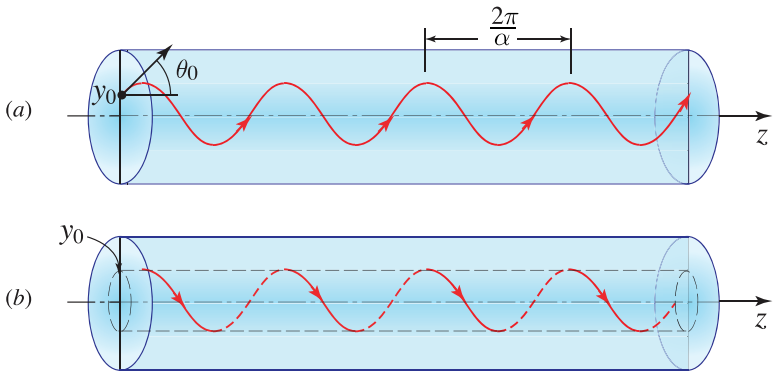


Figure 1.3-7 (a) Meridional and (b) helical rays in a graded-index fiber with parabolic index profile.

so that the ray follows a helical trajectory lying on the surface of a cylinder of radius y_0 [Fig. 1.3-7(b)]. In both cases the ray remains confined within the fiber, so that the fiber serves as a light guide. Other helical patterns are generated with different incident rays.

Graded-index fibers and their use in optical fiber communications are discussed in Chapters 10 and 25.

EXERCISE 1.3-2

Numerical Aperture of the Graded-Index Fiber. Consider a graded-index fiber with the index profile provided in (1.3-14) and radius a. A ray is incident from air into the fiber at its center, which then makes an angle θ_0 with the fiber axis in the medium (see Fig. 1.3-8). Show, in the paraxial approximation, that the numerical aperture is

$$\text{NA} \equiv \sin\theta_a \approx n_0 a\alpha, \tag{1.3-18}$$

Numerical Aperture
Graded-Index Fiber

where θ_a is the maximum acceptance angle for which the ray trajectory is confined within the fiber. Compare this to the numerical aperture of a step-index fiber such as the one discussed in Exercise 1.2-5. To make the comparison fair, take the refractive indices of the core and cladding of the step-index fiber to be $n_1 = n_0$ and $n_2 = n_0\sqrt{1-\alpha^2 a^2} \approx n_0(1 - \frac{1}{2}\alpha^2 a^2)$, respectively.

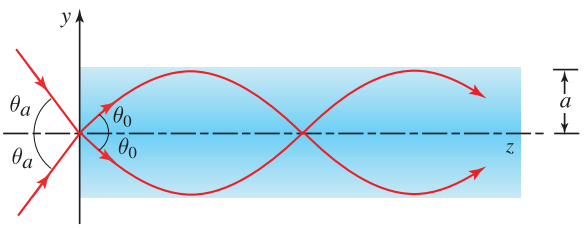


Figure 1.3-8 Acceptance angle of a graded-index optical fiber.

*C. The Eikonal Equation

The ray trajectories are often characterized by the surfaces to which they are normal. Let $S(\mathbf{r})$ be a scalar function such that its equilevel surfaces, $S(\mathbf{r}) = \text{constant}$, are everywhere normal to the rays (Fig. 1.3-9). If $S(\mathbf{r})$ is known, the ray trajectories can readily be constructed since the normal to the equilevel surfaces at a position $\mathbf{r}$ is in the direction of the gradient vector $\nabla S(\mathbf{r})$. The function $S(\mathbf{r})$, called the **eikonal**, is akin to the potential function $V(\mathbf{r})$ in electrostatics; the role of the optical rays is played by the lines of electric field $\mathbf{E} = -\nabla V$.

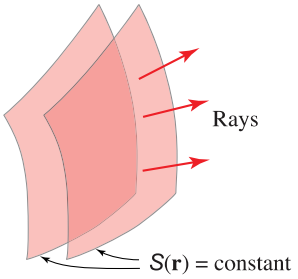


Figure 1.3-9 Ray trajectories are normal to the surfaces of constant $S(\mathbf{r})$.

To satisfy Fermat's principle (which is the main postulate of ray optics) the eikonal $S(\mathbf{r})$ must satisfy a partial differential equation known as the **eikonal equation**,

$$\left(\frac{\partial S}{\partial x}\right)^2 + \left(\frac{\partial S}{\partial y}\right)^2 + \left(\frac{\partial S}{\partial z}\right)^2 = n^2, \tag{1.3-19}$$

which is usually written in the vector form

$$\boxed{|\nabla S|^2 = n^2,} \tag{1.3-20}$$

Eikonal Equation

where $|\nabla S^2| = \nabla S \cdot \nabla S$. The proof of the eikonal equation from Fermat's principle is a mathematical exercise that lies beyond the scope of this book.[†] Conversely, Fermat's principle (and the ray equation) can be shown to follow from the eikonal equation. Therefore, either Fermat's principle or the eikonal equation may be regarded as the principal postulate of ray optics.

Integrating the eikonal equation (1.3-20) along a ray trajectory between points A and B gives

$$S(\mathbf{r}_B) - S(\mathbf{r}_A) = \int_A^B |\nabla S|\, ds = \int_A^B n\, ds = \text{optical pathlength between } A \text{ and } B. \tag{1.3-21}$$

This means that the difference $S(\mathbf{r}_B) - S(\mathbf{r}_A)$ represents the optical pathlength between A and B. In the electrostatics analogy, the optical pathlength plays the role of the potential difference.

To determine the ray trajectories in an inhomogeneous medium of refractive index $n(\mathbf{r})$, we can either solve the ray equation (1.3-3), as we have done earlier, or solve the eikonal equation for $S(\mathbf{r})$, from which we calculate the gradient ∇S.

[†] See, e.g., M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th expanded and corrected ed. 2002.

If the medium is homogeneous, i.e., $n(\mathbf{r})$ is constant, the magnitude of ∇S is constant, so that the wavefront normals (rays) must be straight lines. The surfaces $S(\mathbf{r}) =$ constant may be parallel planes or concentric spheres, as illustrated in Fig. 1.3-10.

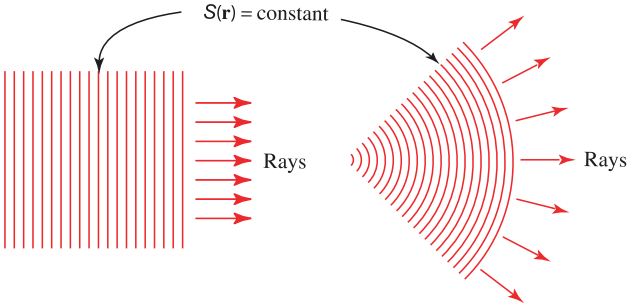


Figure 1.3-10 Rays and surfaces of constant $S(\mathbf{r})$ in a homogeneous medium.

The eikonal equation is revisited from the point-of-view of the relation between ray optics and wave optics in Sec. 2.3.

1.4 MATRIX OPTICS

Matrix optics is a technique for tracing paraxial rays. The rays are assumed to travel only within a single plane, so that the formalism is applicable to systems with planar geometry and to meridional rays in circularly symmetric systems.

A ray is described by its position and its angle with respect to the optical axis. These variables are altered as the ray travels through the system. In the paraxial approximation, the position and angle at the input and output planes of an optical system are related by two *linear* algebraic equations. As a result, the optical system is described by a 2×2 matrix called the ray-transfer matrix.

The convenience of using matrix methods lies in the fact that the ray-transfer matrix of a cascade of optical components (or systems) is a product of the ray-transfer matrices of the individual components (or systems). Matrix optics therefore provides a formal mechanism for describing complex optical systems in the paraxial approximation.

A. The Ray-Transfer Matrix

Consider a circularly symmetric optical system formed by a succession of refracting and reflecting surfaces all centered about the same axis (optical axis). The z axis lies along the optical axis and points in the general direction in which the rays travel. Consider rays in a plane containing the optical axes, say the y–z plane. We proceed to trace a ray as it travels through the system, i.e., as it crosses the transverse planes at different axial distances. A ray crossing the transverse plane at z is completely characterized by the coordinate of y of its crossing point and the angle θ (Fig. 1.4-1).

An optical system is a set of optical components placed between two transverse planes at z_1 and z_2, referred to as the input and output planes, respectively. The system is characterized completely by its effect on an incoming ray of arbitrary position and direction (y_1, θ_1). It steers the ray so that it has new position and direction (y_2, θ_2) at the output plane (Fig. 1.4-2).

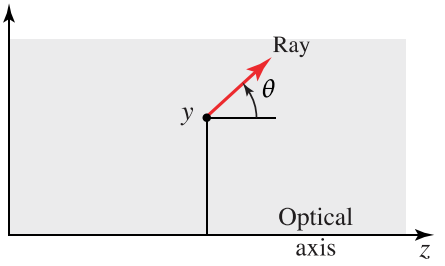


Figure 1.4-1 A ray is characterized by its coordinate y and its angle θ.

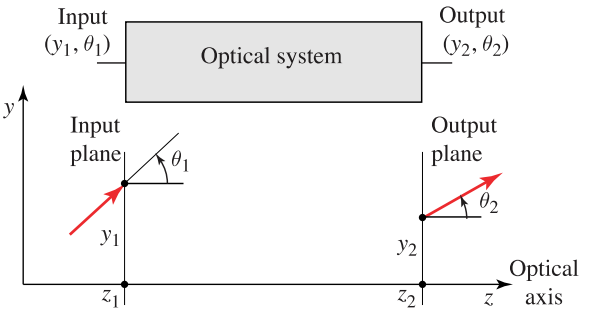


Figure 1.4-2 A ray enters an optical system at location z_1 with position y_1 and angle θ_1 and leaves at position y_2 and angle θ_2.

In the paraxial approximation, when all angles are sufficiently small so that $\sin\theta \approx \theta$, the relation between (y_2, θ_2) and (y_1, θ_1) is linear and can generally be written in the form

$$y_2 = Ay_1 + B\theta_1 \tag{1.4-1}$$

$$\theta_2 = Cy_1 + D\theta_1, \tag{1.4-2}$$

where A, B, C, and D are real numbers. Equations (1.4-1) and (1.4-2) may be conveniently written in matrix form as

$$\begin{bmatrix} y_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ \theta_1 \end{bmatrix}. \tag{1.4-3}$$

The matrix $\mathbf{M}$, whose elements are A, B, C, and D, characterizes the optical system completely since it permits (y_2, θ_2) to be determined for any (y_1, θ_1). It is known as the **ray-transfer matrix**. As will be seen in the examples provided in Sec. 1.4B, angles that turn out to be negative point downward from the z axis in their direction of travel. Radii that turn out to be negative indicate concave surfaces whereas those that are positive indicate convex surfaces.

EXERCISE 1.4-1

Special Forms of the Ray-Transfer Matrix. Consider the following situations in which one of the four elements of the ray-transfer matrix vanishes:

(a) Show that $A = 0$ represents a *focusing system*, in which all rays entering the system at a particular angle, whatever their position, leave at a single position.

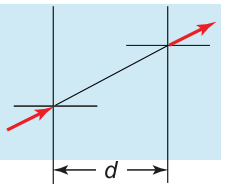
(b) Show that $B = 0$ represents an *imaging system*, in which all rays entering the system at a particular position, whatever their angle, leave at a single position.

(c) What are the special features of a system for which $C = 0$ or $D = 0$?

B. Matrices of Simple Optical Components

Free-Space Propagation

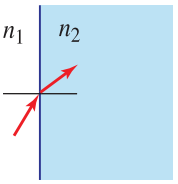
Since rays travel along straight lines in a medium of uniform refractive index such as free space, a ray traversing a distance d is altered in accordance with $y_2 = y_1 + \theta_1 d$ and $\theta_2 = \theta_1$. The ray-transfer matrix is therefore



$$\mathbf{M} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}. \qquad (1.4\text{-}4)$$

Refraction at a Planar Boundary

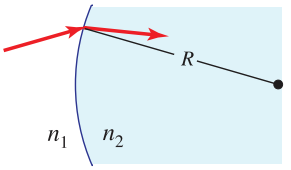
At a planar boundary between two media of refractive indices n_1 and n_2, the ray angle changes in accordance with Snell's law $n_1 \sin\theta_1 = n_2 \sin\theta_2$. In the paraxial approximation, $n_1\theta_1 \approx n_2\theta_2$. The position of the ray is not altered, $y_2 = y_1$. The ray-transfer matrix is



$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{bmatrix}. \qquad (1.4\text{-}5)$$

Refraction at a Spherical Boundary

The relation between θ_1 and θ_2 for paraxial rays refracted at a spherical boundary between two media is provided in (1.2-8). The ray height is not altered, $y_2 \approx y_1$. The ray-transfer matrix is

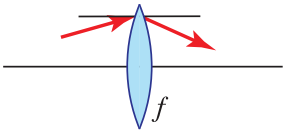


Convex: $R > 0$; concave: $R < 0$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{(n_2 - n_1)}{n_2 R} & \frac{n_1}{n_2} \end{bmatrix}. \qquad (1.4\text{-}6)$$

Transmission Through a Thin Lens

The relation between θ_1 and θ_2 for paraxial rays transmitted through a thin lens of focal length f is given in (1.2-11). Since the height remains unchanged $(y_2 = y_1)$, we have



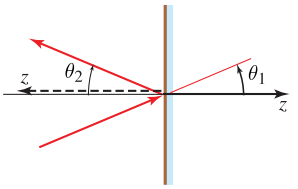
Convex: $f > 0$; concave: $f < 0$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix}. \qquad (1.4\text{-}7)$$

Reflection from a Planar Mirror

Upon reflection from a planar mirror, the ray position is not altered, $y_2 = y_1$. Adopting the convention that the z axis points in the general direction of travel of the rays, i.e., toward the mirror for the incident rays and away from it for the reflected rays, we

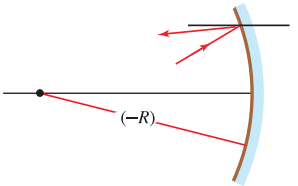
conclude that $\theta_2 = \theta_1$. The ray-transfer matrix is therefore the identity matrix



$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \qquad (1.4\text{-}8)$$

Reflection from a Spherical Mirror

Using (1.2-1), and the convention that the z axis follows the general direction of the rays as they reflect from mirrors, we similarly obtain



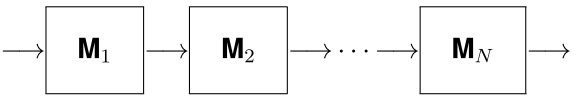
Concave: $R < 0$; convex: $R > 0$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ \frac{2}{R} & 1 \end{bmatrix}. \qquad (1.4\text{-}9)$$

Note the similarity between the ray-transfer matrices of a spherical mirror (1.4-9) and a thin lens (1.4-7). A mirror with radius of curvature R bends rays in a manner that is identical to that of a thin lens with focal length $f = -R/2$.

C. Matrices of Cascaded Optical Components

A cascade of N optical components or systems whose ray-transfer matrices are $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_N$ is equivalent to a single optical system of ray-transfer matrix

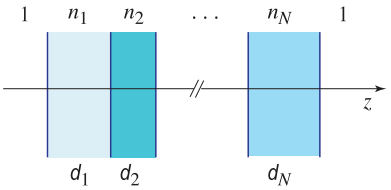


$$\mathbf{M} = \mathbf{M}_N \cdots \mathbf{M}_2 \mathbf{M}_1. \qquad (1.4\text{-}10)$$

Note the order of matrix multiplication: The matrix of the system that is crossed by the rays is first placed to the right, so that it operates on the column matrix of the incident ray first. A sequence of matrix multiplications is not, in general, commutative, although it is associative.

EXERCISE 1.4-2

A Set of Parallel Transparent Plates. Consider a set of N parallel planar transparent plates of refractive indices $n_1, n_2, \ldots, n_N$ and thicknesses $d_1, d_2, \ldots, d_N$, placed in air $(n = 1)$ normal to the z axis. Using induction, show that the ray-transfer matrix is

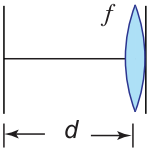


$$\mathbf{M} = \begin{bmatrix} 1 & \sum_{i=1}^{N} \frac{d_i}{n_i} \\ 0 & 1 \end{bmatrix}. \qquad (1.4\text{-}11)$$

Note that the order in which the plates are placed does not affect the overall ray-transfer matrix. What is the ray-transfer matrix of an inhomogeneous transparent plate of thickness d_0 and refractive index $n(z)$?

EXERCISE 1.4-3

A Gap Followed by a Thin Lens. Show that the ray-transfer matrix of a distance d of free space followed by a lens of focal length f is



$$\mathbf{M} = \begin{bmatrix} 1 & d \\ -\frac{1}{f} & 1 - \frac{d}{f} \end{bmatrix}. \qquad (1.4\text{-}12)$$

EXERCISE 1.4-4

Imaging with a Thin Lens. Derive an expression for the ray-transfer matrix of a system comprised of free space/thin lens/free space, as shown in Fig. 1.4-3. Show that if the imaging condition $(1/d_1 + 1/d_2 = 1/f)$ is satisfied, all rays originating from a single point in the input plane reach the output plane at the single point y_2, regardless of their angles. Also show that if $d_2 = f$, all parallel incident rays are focused by the lens onto a single point in the output plane.

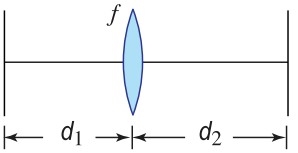


Figure 1.4-3 Single-lens imaging system.

Imaging with an Arbitrary Paraxial Optical System

A paraxial system comprising an arbitrary set of cascaded optical elements is characterized completely by the four elements A, B, C, D of its ray-transfer matrix $\mathbf{M}$. Alternatively, the system may be characterized by the locations of four *cardinal points*: two focal points that determine the transmission of rays between its input and output planes. In accordance with (1.4-3), an incoming ray parallel to the optical axis $(\theta_1 = 0)$ at height y_1 exits the system at height $y_2 = Ay_1$ and angle $\theta_2 = Cy_1$. This ray crosses the axis at a point F called the **back focal point**, which is located a distance $y_2/\theta_2 = A/C$ from the system's back vertex V, as shown in Fig. 1.4-4(a). The intersection of the extensions of the incoming and outgoing rays defines the **back principal point** H, which lies at a distance $f = y_1/\theta_2 = -1/C$ to the left of F, and is known as the **back focal length**. The back principal point H is thus located at a distance $h = -1/C + A/C$ to the left of the back vertex V. Note that the locations of the focal and principal points are independent of y_1 as long as the paraxial approximation is applicable.

Similarly, rays parallel to the axis but entering the system in the opposite direction (from right to left) are focused to the **front focal point** F' and define the **front principal point** H', which lies at a distance h' from the front vertex V'. The front focal point lies at a distance f' to the left of H', where f' is the **front focal length**. These distances may be expressed in terms of the elements of the inverse ray-transfer matrix

$$\mathbf{M}^{-1} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} = \frac{1}{\det[\mathbf{M}]} \begin{bmatrix} D & -B \\ -C & DA \end{bmatrix} \qquad (1.4\text{-}13)$$

via the relations $-f' = -1/C'$ and $-h' = -1/C' + A'/C'$. The determinant of $\mathbf{M}$, denoted $\det[\mathbf{M}]$, is given by $AD - BC$.

In summary, the focal lengths and locations of the principal points may be determined from the $ABCD$ parameters by use of the following relations:

$$f = -1/C, \qquad h = (1 - A)f \tag{1.4-14}$$

$$f' = \det[\mathbf{M}]\, f, \qquad h' = -f' + Df. \tag{1.4-15}$$

Negative signs indicate directions opposite to those denoted by the arrows in Fig. 1.4-4(a). The four distances may alternatively be established by tracing two rays, parallel to the optical axis but pointing in opposite directions, through the system. The $ABCD$ parameters may be determined from f, f', h, and h' by inverting (1.4-14) and (1.4-15).

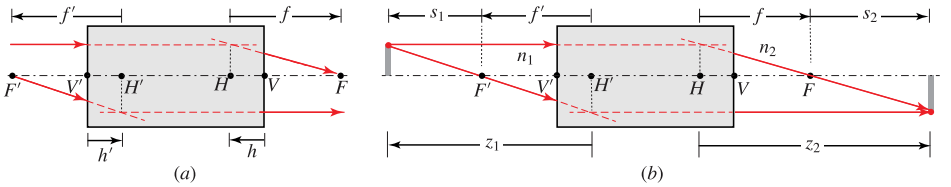


Figure 1.4-4 (a) Paraxial system representing an arbitrary set of cascaded optical elements. The designations F, V, and H represent the focal, vertex, and principal points, respectively, whereas f and h represent the focal length and distance from the principal point to the vertex, respectively. Primed quantities refer to the input plane while unprimed quantities refer to the output plane. (b) Imaging with this system. The refractive indices of the media in which the optical system is embedded are denoted n_1 and n_2, as shown.

The imaging condition is determined by considering the geometry portrayed in Fig. 1.4-4(b). Since $s_2/f = f'/s_1$, the imaging condition is simply $s_1 s_2 = ff'$, or equivalently $(z_1 - f')(z_2 - f) = ff'$, which leads to

$$\frac{f'}{z_1} + \frac{f}{z_2} = 1. \tag{1.4-16}$$

If the refractive indices of the media within which the system is embedded are equal, then $\det[\mathbf{M}] = 1$ and, in accordance with (1.4-15), we have $f' = f$. The imaging condition in (1.4-16) then reduces to the familiar imaging equation $1/z_1 + 1/z_2 = 1/f$ [see (1.2-4)]; note, however, that here the distances z_1 and z_2 are measured from the principal points H' and H, respectively.

EXERCISE 1.4-5

Imaging with a Thick Lens. Consider a glass lens of refractive index n, thickness d, and two spherical surfaces of equal radii R. Determine the ray-transfer matrix of the lens assuming that it is placed in air (unity refractive index). Show that the back and front focal lengths are equal ($f' = f$) and that the principal points are located at equal distances from the vertices ($h' = h$), where

$$\frac{1}{f} = \frac{(n-1)}{R}\left[2 - \frac{n-1}{n}\frac{d}{R}\right] \tag{1.4-17}$$

$$h = \frac{(n-1)fd}{nR}. \tag{1.4-18}$$

Demonstrate that the transfer matrix of the system between two conjugate planes at distances z_1 and z_2 from the principal points of the lens (i.e., at distances $d_1 = z_1 - h'$ and $d_2 = z_2 - h$ from the vertices) that satisfies the imaging equation yields $B = 0$, indicating that it does indeed satisfy the imaging condition [see Exercise 1.4-1(b)].

D. Periodic Optical Systems

A periodic optical system is a cascade of identical unit systems. An example is a sequence of equally spaced identical relay lenses used to guide light, as shown in Fig. 1.2-19(a). Another example is the reflection of light between two mirrors that form an optical resonator (see Sec. 11.2A); in that case, the ray repeatedly traverses the same unit system (a round trip of reflections). Even a homogeneous medium, such as a glass fiber, may be considered as a periodic system if it is divided into contiguous identical segments of equal length. We proceed to formulate a general theory of ray propagation in periodic optical systems using matrix methods.

Difference Equation for the Ray Position

A periodic system is composed of a cascade of identical unit systems (stages), each with a ray-transfer matrix (A, B, C, D), as shown in Fig. 1.4-5. A ray enters the system with initial position y_0 and slope θ_0. To determine the position and slope (y_m, θ_m) of the ray at the exit of the mth stage, we apply the $ABCD$ matrix m times,

$$\begin{bmatrix} y_m \\ \theta_m \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^m \begin{bmatrix} y_0 \\ \theta_0 \end{bmatrix}. \tag{1.4-19}$$

We can also iteratively apply the relations

$$y_{m+1} = Ay_m + B\theta_m \tag{1.4-20}$$

$$\theta_{m+1} = Cy_m + D\theta_m \tag{1.4-21}$$

to determine (y_1, θ_1) from (y_0, θ_0), then (y_2, θ_2) from (y_1, θ_1), and so on, using a software routine.

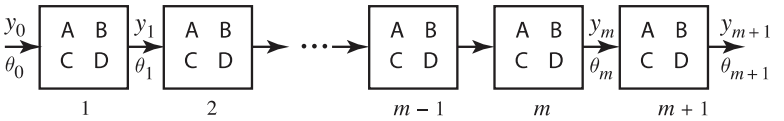


Figure 1.4-5 A cascade of identical optical systems.

It is of interest to derive equations that govern the dynamics of the position y_m, $m = 0, 1, \ldots$, irrespective of the angle θ_m. This is achieved by eliminating θ_m from (1.4-20) and (1.4-21). From (1.4-20)

$$\theta_m = \frac{y_{m+1} - Ay_m}{B}. \tag{1.4-22}$$

Replacing m with $m + 1$ in (1.4-22) yields

$$\theta_{m+1} = \frac{y_{m+2} - Ay_{m+1}}{B}. \tag{1.4-23}$$

Substituting (1.4-22) and (1.4-23) into (1.4-21) gives

$$\boxed{y_{m+2} = 2by_{m+1} - F^2 y_m,} \tag{1.4-24}$$

Recurrence Relation for Ray Position

where

$$b = \frac{A + D}{2} \tag{1.4-25}$$

$$F^2 = AD - BC = \det[\mathbf{M}], \tag{1.4-26}$$

and $\det[\mathbf{M}]$ is the determinant of $\mathbf{M}$.

Equation (1.4-24) is a linear difference equation governing the ray position y_m. It can be solved iteratively by computing y_2 from y_0 and y_1, then y_3 from y_1 and y_2, and so on. The quantity y_1 may be computed from y_0 and θ_0 by use of (1.4-20) with $m = 0$.

It is useful, however, to derive an explicit expression for y_m by solving the difference equation (1.4-24). As with linear differential equations, a solution satisfying a linear difference equation and the initial conditions is a unique solution. It is therefore appropriate to make a judicious guess for the solution of (1.4-24). We use a trial solution of the geometric form

$$y_m = y_0 h^m, \tag{1.4-27}$$

where h is a constant. Substituting (1.4-27) into (1.4-24) immediately shows that the trial solution is suitable provided that h satisfies the quadratic algebraic equation

$$h^2 - 2bh + F^2 = 0, \tag{1.4-28}$$

from which

$$h = b \pm j\sqrt{F^2 - b^2}. \tag{1.4-29}$$

The results can be presented in a more compact form by defining the variable

$$\varphi = \cos^{-1}(b/F), \tag{1.4-30}$$

so that $b = F\cos\varphi$, $\sqrt{F^2 - b^2} = F\sin\varphi$, and therefore $h = F(\cos\varphi \pm j\sin\varphi) = F\exp(\pm j\varphi)$, whereupon (1.4-27) becomes $y_m = y_0 F^m \exp(\pm jm\varphi)$.

A general solution may be constructed from the two solutions with positive and negative signs by forming their linear combination. The sum of the two exponential functions can always be written as a harmonic (circular) function, so that

$$y_m = y_{\max} F^m \sin(m\varphi + \varphi_0), \tag{1.4-31}$$

where $y_{\max}$ and φ_0 are constants to be determined from the initial conditions y_0 and y_1. In particular, setting $m = 0$ we obtain $y_{\max} = y_0/\sin\varphi_0$.

The parameter F is related to the determinant of the ray-transfer matrix of the unit system by $F = \sqrt{\det[\mathbf{M}]}$. It can be shown that regardless of the unit system, $\det[\mathbf{M}] = n_1/n_2$, where n_1 and n_2 are the refractive indices of the initial and final sections of the unit system. This general result is easily verified for the ray-transfer matrices of all the optical components considered in this section. Since the determinant of a product of two matrices is the product of their determinants, it follows that the relation $\det[\mathbf{M}] = n_1/n_2$ is applicable to any cascade of these optical components. For example, if $\det[\mathbf{M}_1] = n_1/n_2$ and $\det[\mathbf{M}_2] = n_2/n_3$, then $\det[\mathbf{M}_2\mathbf{M}_1] = (n_2/n_3)(n_1/n_2) = n_1/n_3$. In most applications the first and last stages are air ($n = 1$) so that $n_1 = n_2$, which leads to $\det[\mathbf{M}] = 1$ and $F = 1$. In that case the solution for the ray position is

$$\boxed{y_m = y_{\max}\sin(m\varphi + \varphi_0).} \tag{1.4-32}$$

Ray Position
Periodic System

We shall henceforth assume that $F = 1$. The corresponding solution for the ray angle is obtained by use of the relation $\theta_m = (y_{m+1} - A y_m)/B$, which is derived from (1.4-20).

Condition for a Harmonic Trajectory

For y_m to be a harmonic (instead of hyperbolic) function, $\varphi = \cos^{-1} b$ must be real. This requires that

$$|b| \le 1 \qquad \text{or} \qquad \tfrac{1}{2}|A + D| \le 1. \tag{1.4-33}$$

Stability Condition

If, instead, $|b| > 1$, φ is then imaginary and the solution is a hyperbolic function (cosh or sinh), which increases without bound, as illustrated in Fig. 1.4-6(a). A harmonic solution ensures that y_m is bounded for all m, with a maximum value of $y_{\max}$. The bound $|b| \le 1$ therefore provides a condition of **stability** (boundedness) of the ray trajectory.

Since y_m and y_{m+1} are both harmonic functions, so too is the ray angle corresponding to (1.4-32), by virtue of (1.4-22) and trigonometric identities. Thus, $\theta_m = \theta_{\max} \sin(m\varphi + \varphi_1)$, where the constants $\theta_{\max}$ and φ_1 are determined by the initial conditions. The maximum angle $\theta_{\max}$ must be sufficiently small so that the paraxial approximation, which underlies this analysis, is applicable.

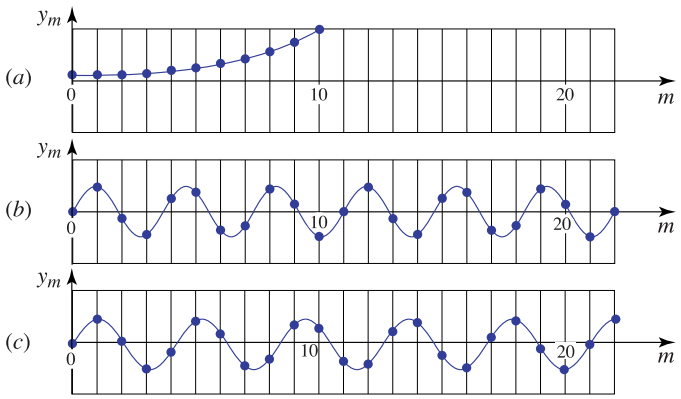


Figure 1.4-6 Examples of trajectories in periodic optical systems: (a) unstable trajectory ($b > 1$); (b) stable and periodic trajectory ($\varphi = 6\pi/11$; period = 11 stages); (c) stable but nonperiodic trajectory ($\varphi = 1.5$).

Condition for a Periodic Trajectory

The harmonic function (1.4-32) is periodic in m if it is possible to find an integer s such that $y_{m+s} = y_m$ for all m. The smallest integer is the period. The ray then retraces its path after s stages. This condition is satisfied if $s\varphi = 2\pi q$, where q is an integer. Thus, the necessary and sufficient condition for a periodic trajectory is that $\varphi/2\pi$ is a rational number q/s. If $\varphi = 6\pi/11$, for example, then $\varphi/2\pi = {}^3/_{11}$ and the trajectory is periodic with period $s = 11$ stages. This case is illustrated in Fig. 1.4-6(b). Periodic optical systems will be revisited in Chapter 7.

Summary

A paraxial ray ($\theta_{\max} \ll 1$) traveling through a cascade of identical unit optical systems, each with a ray-transfer matrix with elements (A, B, C, D) such that $AD - BC = 1$, follows a harmonic (and therefore bounded) trajectory if the condition $|\frac{1}{2}(A+D)| \le 1$, called the stability condition, is satisfied. The position at the mth stage is then $y_m = y_{\max}\sin(m\varphi + \varphi_0)$, $m = 0, 1, 2, \ldots$, where $\varphi = \cos^{-1}[\frac{1}{2}(A+D)]$. The constants $y_{\max}$ and φ_0 are determined from the initial positions y_0 and $y_1 = Ay_0 + B\theta_0$, where θ_0 is the initial ray inclination. The ray angles are related to the positions by $\theta_m = (y_{m+1} - Ay_m)/B$ and follow a harmonic function $\theta_m = \theta_{\max}\sin(m\varphi + \varphi_1)$. The ray trajectory is periodic with period s if $\varphi/2\pi$ is a rational number q/s.

EXAMPLE 1.4-1. ***A Sequence of Equally Spaced Identical Lenses.*** A set of identical lenses of focal length f separated by distance d, as shown in Fig. 1.4-7, may be used to relay light between two locations. The unit system, a distance of d of free space followed by a lens, has a ray-transfer matrix given by (1.4-12); $A = 1$, $B = d$, $C = -1/f$, $D = 1 - d/f$. The parameter $b = \frac{1}{2}(A+D) = 1 - d/2f$ and the determinant is unity. The condition for a stable ray trajectory, $|b| \le 1$ or $-1 \le b \le 1$, is therefore

$$0 \le d \le 4f, \tag{1.4-34}$$

so that the spacing between the lenses must be smaller than four times the focal length. Under this condition the positions of paraxial rays obey the harmonic function

$$y_m = y_{\max}\sin(m\varphi + \varphi_0), \qquad \varphi = \cos^{-1}\left(1 - \frac{d}{2f}\right). \tag{1.4-35}$$

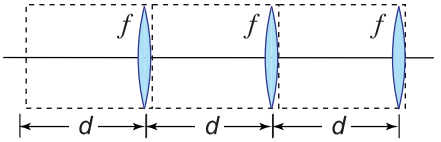


Figure 1.4-7 A periodic sequence of lenses.

When $d = 2f$, $\varphi = \pi/2$, and $\varphi/2\pi = 1/4$, so that the trajectory of an arbitrary ray is periodic with period equal to four stages. When $d = f$, $\varphi = \pi/3$, and $\varphi/2\pi = 1/6$, so that the ray trajectory is periodic and retraces itself each six stages. These cases are illustrated in Fig. 1.4-8.

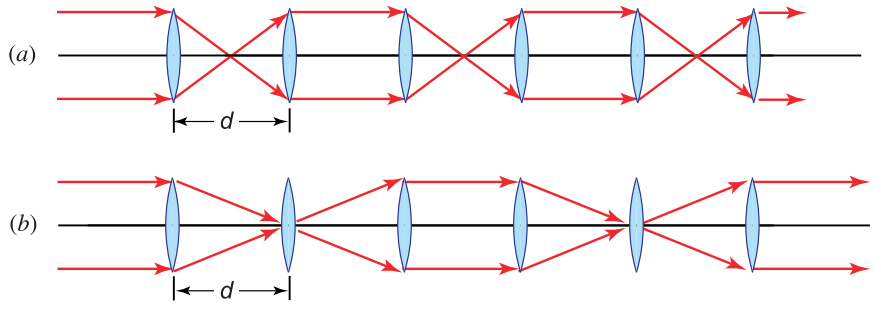


Figure 1.4-8 Examples of stable ray trajectories in a periodic lens system: (a) $d = 2f$; (b) $d = f$.

EXERCISE 1.4-6

A Periodic Set of Pairs of Different Lenses. Examine the trajectories of paraxial rays through a periodic system comprising a sequence of lens pairs with alternating focal lengths f_1 and f_2, as shown in Fig. 1.4-9. Show that the ray trajectory is bounded (stable) if

$$0 \le \left(1 - \frac{d}{2f_1}\right)\left(1 - \frac{d}{2f_2}\right) \le 1. \tag{1.4-36}$$

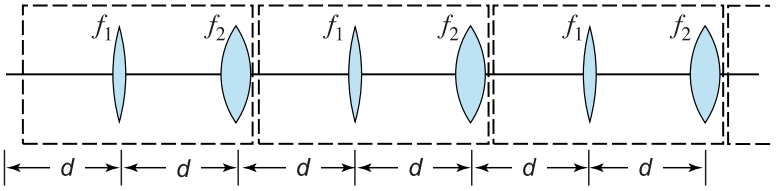


Figure 1.4-9 A periodic sequence of lens pairs.

EXERCISE 1.4-7

An Optical Resonator. Paraxial rays are reflected repeatedly between two spherical mirrors of radii R_1 and R_2 separated by a distance d (Fig. 1.4-10). Regarding this as a periodic system whose unit system is a single round trip between the mirrors, determine the condition of stability for the ray trajectory. Optical resonators will be studied in detail in Chapter 11.

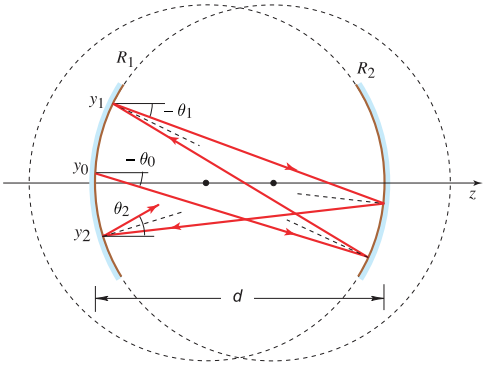


Figure 1.4-10 The optical resonator as a periodic optical system.

READING LIST

General Optics

F. L. Pedrotti, L. M. Pedrotti, and L. S. Pedrotti, *Introduction to Optics*, Cambridge University Press, 3rd ed. 2018.

T.-C. Poon and T. Kim, *Engineering Optics with MATLAB*, World Scientific, 2nd ed. 2018.

E. Hecht, *Optics*, Pearson, 5th ed. 2016.

S. D. Gupta, N. Ghosh, and A. Banerjee, *Wave Optics: Basic Concepts and Contemporary Trends*, CRC Press/Taylor & Francis, 2016.

B. D. Guenther, *Modern Optics*, Oxford University Press, 2nd ed. 2015.

C. A. DiMarzio, *Optics for Engineers*, CRC Press/Taylor & Francis, 2011.

M. Mansuripur, *Classical Optics and Its Applications*, Cambridge University Press, 2nd ed. 2009.

A. Walther, *The Ray and Wave Theory of Lenses*, Cambridge University Press, 1995, paperback ed. 2006.

A. Siciliano, *Optics: Problems and Solutions*, World Scientific, 2006.

M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th expanded and corrected ed. 2002.

D. T. Moore, ed., *Selected Papers on Gradient-Index Optics*, SPIE Optical Engineering Press (Milestone Series Volume 67), 1993.

F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw–Hill, 1937, 4th revised ed. 1991.

R. W. Wood, *Physical Optics*, Macmillan, 3rd ed. 1934; Optical Society of America, 1988.

A. Sommerfeld, *Lectures on Theoretical Physics: Optics*, Academic Press, paperback ed. 1954.

Geometrical Optics

P. D. Lin, *Advanced Geometrical Optics*, Springer-Verlag, 2017.

V. N. Mahajan, *Fundamentals of Geometrical Optics*, SPIE Optical Engineering Press, 2014.

E. Dereniak and T. D. Dereniak, *Geometrical and Trigonometric Optics*, Cambridge University Press, 2008.

Yu. A. Kravtsov, *Geometrical Optics in Engineering Physics*, Alpha Science, 2005.

J. E. Greivenkamp, *Field Guide to Geometrical Optics*, SPIE Optical Engineering Press, 2004.

P. Mouroulis and J. Macdonald, *Geometrical Optics and Optical Design*, Oxford University Press, 1997.

Optical System Design

D. Malacara-Hernández and Z. Malacara-Hernández, *Handbook of Optical Design*, CRC Press/Taylor & Francis, 3rd ed. 2013.

J. Sasián, *Introduction to Aberrations in Optical Imaging Systems*, Cambridge University Press, 2013.

K. J. Kasunic, *Optical Systems Engineering*, McGraw–Hill, 2011.

R. E. Fischer, B. Tadic-Galeb, and P. R. Yoder, *Optical System Design*, McGraw–Hill, 2nd ed. 2008.

W. J. Smith, *Modern Optical Engineering*, McGraw–Hill, 1966, 4th ed. 2008.

D. C. O'Shea, *Elements of Modern Optical Design*, Wiley, 1985.

Matrix Optics

A. Gerrard and J. M. Burch, *Introduction to Matrix Methods in Optics*, Wiley, 1975; Dover, reprinted 2012.

J. W. Blaker, *Geometric Optics: The Matrix Theory*, CRC Press, 1971.

Popular and Historical

R. J. Weiss, *A Brief History of Light and Those that Lit the Way*, World Scientific, 1996.

A. R. Hall, *All was Light: An Introduction to Newton's Opticks*, Clarendon Press/Oxford University Press, 1993.

R. Kingslake, *A History of the Photographic Lens*, Academic Press, 1989.

M. I. Sobel, *Light*, University of Chicago Press, 1987.

A. I. Sabra, *Theories of Light from Descartes to Newton*, Cambridge University Press, 1981.

V. Ronchi, *The Nature of Light: An Historical Survey*, Harvard University Press, 1970.

W. H. Bragg, *Universe of Light*, Dover, paperback ed. 1959.

I. Newton, *Opticks: or A Treatise of the Reflections, Refractions, Inflections & Colours of Light*, 4th ed. 1704; Dover, reissued 1979.

PROBLEMS

1.1-2 **Fermat's Principle with Maximum Time.** Consider the elliptical mirror shown in Fig. P1.1-2(a), whose foci are denoted A and B. Geometrical properties of the ellipse dictate that the

pathlength $\overline{APB}$ is identical to the pathlengths $\overline{AP'B}$ and $\overline{AP''B}$ for adjacent points on the ellipse.

(a) Now consider another mirror with a radius of curvature smaller than that of the elliptical mirror, but tangent to it at P, as displayed in Fig. P1.1-2(b). Show that the path $\overline{APB}$ followed by the light ray in traveling between points A and B is a path of *maximum* time, i.e., is greater than the adjacent paths $\overline{AQ'B}$ and $\overline{AQ''B}$.

(b) Finally, consider a mirror that crosses the ellipse, but is tangent to it at P, as illustrated in Fig. P1.1-2(c). Show that the possible ray paths $\overline{AQ'B}$, $\overline{APB}$, and $\overline{AQ''B}$ exhibit a point of inflection.

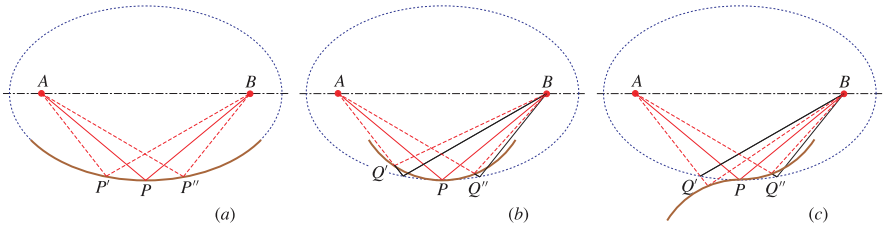


Figure P1.1-2 (a) Reflection from an elliptical mirror. (b) Reflection from an inscribed tangential mirror with greater curvature. (c) Reflection from a tangential mirror with curvature changing from concave to convex.

1.2-7 **Transmission through Planar Plates.**

(a) Use Snell's law to show that a ray entering a planar plate of thickness d and refractive index n_1 (placed in air; $n \approx 1$) emerges parallel to its initial direction. The ray need not be paraxial. Derive an expression for the lateral displacement of the ray as a function of the angle of incidence θ. Explain your results in terms of Fermat's principle.

(b) If the plate instead comprises a stack of N parallel layers stacked against each other with thicknesses $d_1, d_2, \ldots, d_N$ and refractive indices $n_1, n_2, \ldots, n_N$, show that the transmitted ray is parallel to the incident ray. If θ_m is the angle of the ray in the mth layer, show that $n_m \sin\theta_m = \sin\theta$, $m = 1, 2, \ldots$.

1.2-8 **Lens in Water.** Determine the focal length f of a biconvex lens with radii 20 cm and 30 cm and refractive index $n = 1.5$. What is the focal length when the lens is immersed in water $(n = 4/3)$?

1.2-9 **Numerical Aperture of a Cladless Fiber.** Determine the numerical aperture and the acceptance angle of an optical fiber if the refractive index of the core is $n_1 = 1.46$ and the cladding is stripped out (replaced with air $n_2 \approx 1$).

1.2-10 **Fiber Coupling Spheres.** Tiny glass balls are often used as lenses to couple light into and out of optical fibers. The fiber end is located at a distance f from the sphere. For a sphere of radius $a = 1$ mm and refractive index $n = 1.8$, determine f such that a ray parallel to the optical axis at a distance $y = 0.7$ mm is focused onto the fiber, as illustrated in Fig. P1.2-10.

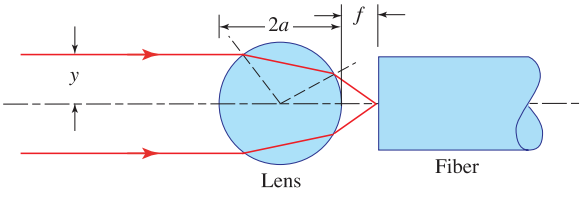


Figure P1.2-10 Focusing light into an optical fiber with a spherical glass ball.

1.2-11 **Extraction of Light from a High-Refractive-Index Medium.** Assume that light is generated isotropically in all directions inside a material of refractive index $n = 3.7$ cut in the shape of a parallelepiped and placed in air $(n = 1)$ (see Exercise 1.2-6).

(a) If a reflective material acting as a perfect mirror is coated on all sides except the front

side, determine the percentage of light that may be extracted from the front side.

(b) If another transparent material of refractive index $n = 1.4$ is placed on the front side, would that help extract some of the trapped light?

1.3-3 **Axially Graded Plate.** A plate of thickness d is oriented normal to the z axis. The refractive index $n(z)$ is graded in the z direction. Show that a ray entering the plate from air at an incidence angle θ_0 in the y–z plane makes an angle $\theta(z)$ at position z in the medium given by $n(z)\sin\theta(z) = \sin\theta_0$. Show that the ray emerges into air parallel to the original incident ray. *Hint:* You may use the results of Prob. 1.2-7. Show that the ray position $y(z)$ inside the plate obeys the differential equation $(dy/dz)^2 = (n^2/\sin^2\theta - 1)^{-1}$.

1.3-4 **Ray Trajectories in GRIN Fibers.** Consider a graded-index optical fiber with cylindrical symmetry about the z axis and refractive index $n(\rho)$, $\rho = \sqrt{x^2 + y^2}$. Let (ρ, ϕ, z) be the position vector in a cylindrical coordinate system. Rewrite the paraxial ray equations, (1.3-4), in a cylindrical system and derive differential equations for ρ and ϕ as functions of z.

1.4-8 **Ray-Transfer Matrix of a Lens System.** Determine the ray-transfer matrix for an optical system made of a thin convex lens of focal length f and a thin concave lens of focal length $-f$ separated by a distance f. Discuss the imaging properties of this composite lens.

1.4-9 **Ray-Transfer Matrix of a GRIN Plate.** Determine the ray-transfer matrix of a SELFOC plate [i.e., a graded-index material with parabolic refractive index $n(y) \approx n_0(1 - \frac{1}{2}\alpha^2 y^2)$] of thickness d.

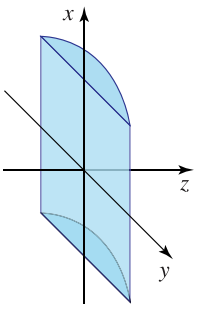
1.4-10 **The GRIN Plate as a Periodic System.** Consider the trajectories of paraxial rays inside a SELFOC plate normal to the z axis. This system may be regarded as a periodic system comprising a sequence of identical contiguous plates, each of thickness d. Using the result of Prob. 1.4-9, determine the stability condition of the ray trajectory. Is this condition dependent on the choice of d?

1.4-11 **Recurrence Relation for a Planar-Mirror Resonator.** Consider a planar-mirror optical resonator, with mirror separation d, as a periodic optical system. Determine the unit ray-transfer matrix for this system, demonstrating that $b = 1$ and $F = 1$. Show that there is then only a single root to the quadratic equation (1.4-28) so that the ray position must then take the form $\alpha + m\beta$, where α and β are constants.

1.4-12 **4 × 4 Ray-Transfer Matrix for Skewed Rays.** Matrix methods may be generalized to describe skewed paraxial rays in circularly symmetric systems, and to astigmatic (non-circularly symmetric) systems. A ray crossing the plane $z = 0$ is generally characterized by four variables — the coordinates (x, y) of its position in the plane, and the angles (θ_x, θ_y) that its projections in the x–z and y–z planes make with the z axis. The emerging ray is also characterized by four variables that are linearly related to the initial four variables. The optical system may then be characterized completely, within the paraxial approximation, by a 4×4 matrix.

(a) Determine the 4×4 ray-transfer matrix of a distance d in free space.

(b) Determine the 4×4 ray-transfer matrix of a thin cylindrical lens with focal length f oriented in the y direction. The cylindrical lens has focal length f for rays in the y–z plane, and no focusing power for rays in the x–z plane.



CHAPTER

2

WAVE OPTICS

2.1 POSTULATES OF WAVE OPTICS 43

2.2 MONOCHROMATIC WAVES 44
- A. Complex Representation and the Helmholtz Equation
- B. Elementary Waves
- C. Paraxial Waves

*2.3 RELATION BETWEEN WAVE OPTICS AND RAY OPTICS 52

2.4 SIMPLE OPTICAL COMPONENTS 53
- A. Reflection and Refraction
- B. Transmission Through Optical Components
- C. Graded-Index Optical Components

2.5 INTERFERENCE 61
- A. Interference of Two Waves
- B. Multiple-Wave Interference

2.6 POLYCHROMATIC AND PULSED LIGHT 71
- A. Temporal and Spectral Description
- B. Light Beating

Christiaan Huygens (1629–1695) advanced a number of novel concepts pertaining to the propagation of light waves.

Thomas Young (1773–1829) championed the wave theory of light and discovered the principle of optical interference.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Light propagates in the form of waves. In free space, light waves travel with a constant speed, $c_o = 3.0 \times 10^8$ m/s (30 cm/ns or 0.3 mm/ps or 0.3 μm/fs or 0.3 nm/as). As illustrated in Fig. 2.0-1, the range of optical wavelengths comprises three principal sub-regions: infrared (0.760 to 300 μm), visible (390 to 760 nm), and ultraviolet (10 to 390 nm). The corresponding range of optical frequencies stretches from 1 THz in the far-infrared to 30 PHz in the extreme ultraviolet.

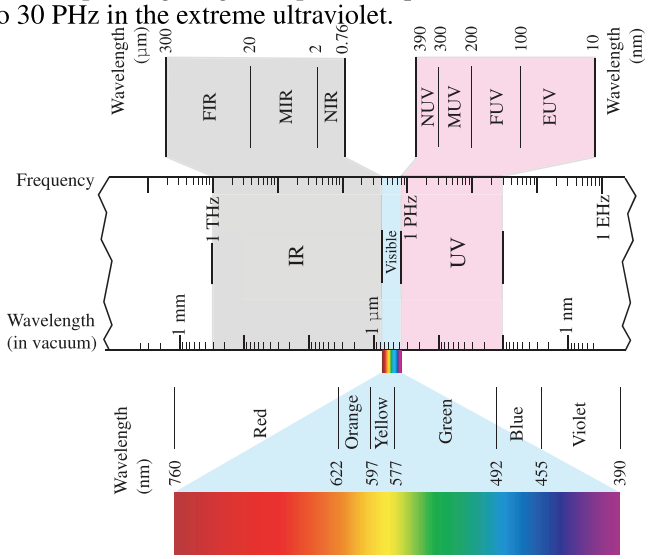


Figure 2.0-1 Optical frequencies and wavelengths. The infrared (**IR**) region of the spectrum comprises the near-infrared (**NIR**), mid-infrared (**MIR**), and far-infrared (**FIR**) bands. The **MWIR** and **LWIR** bands both lie within the MIR band; radiation in these regions can penetrate the atmosphere. The ultraviolet (**UV**) region comprises the near-ultraviolet (**NUV**), mid-ultraviolet (**MUV**) or deep-ultraviolet (**DUV**), far-ultraviolet (**FUV**), and extreme-ultraviolet (**EUV** or **XUV**) bands. The vacuum ultraviolet (**VUV**) consists of the FUV and EUV bands. The ultraviolet region is also divided into the **UVA**, **UVB**, and **UVC** bands, which have chemical and biological significance. The infrared, visible, and ultraviolet regions are gathered under the rubric "optical" since they make use of similar types of components (e.g., lenses and mirrors). The terahertz (**THz**) region occupies frequencies that stretch from 0.3 to 3 THz, corresponding to wavelengths that extend from 1 mm to 100 μm; the THz region partially overlaps the FIR band. For X-ray wavelengths, see Fig. 16.3-7.

The **wave theory** of light encompasses the ray theory (Fig. 2.0-2). Strictly speaking, ray optics is the limit of wave optics when the wavelength is infinitesimally short. However, the wavelength need not actually be zero for the ray-optics theory to be useful. As long as the light waves propagate through and around objects whose dimensions are much greater than the wavelength, the ray theory suffices for describing most optical phenomena. Because the wavelength of visible light is much smaller than the dimensions of the usual objects we encounter on a daily basis, the manifestations of the wave nature of light are usually not apparent without careful observation.

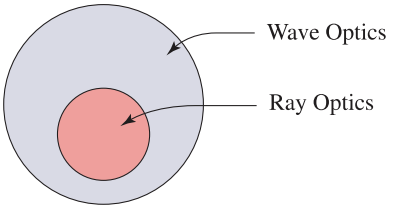


Figure 2.0-2 Wave optics encompasses ray optics. Ray optics is the limit of wave optics when the wavelength is very short.

This Chapter

In the context of wave optics, light is described by a scalar function, called the wavefunction, that obeys a second-order differential equation known as the wave equation. A discussion of the physical significance of the wavefunction is deferred to Chapter 5, where we consider electromagnetic optics; it will become apparent there that the wavefunction represents any of the components of the electric or magnetic fields. The wave equation, together with a relation between the optical power density and the wavefunction, constitute the postulates of the scalar-wave model of light known as **wave optics**. The consequences of these simple postulates are manifold and far reaching. Wave optics constitutes a basis for describing a host of optical phenomena that fall outside the confines of ray optics, including interference and diffraction, as will become clear in this and the following two chapters (Chapters 3 and 4).

Wave optics does have its limitations, however. It is not capable of providing a complete picture of the reflection and refraction of light at the boundaries between various media, nor can it accommodate optical phenomena that require a vector formulation, such as polarization effects. Those issues will be considered from a fundamental perspective in Chapters 5–8, as will the conditions under which scalar wave optics provides a good approximation to electromagnetic optics.

The chapter begins with the postulates of wave optics (Sec. 2.1). In Secs. 2.2–2.5 we consider monochromatic waves. Elementary waves, such as the plane wave, the spherical wave, and paraxial waves are introduced in Sec. 2.2. Section 2.3 establishes how ray optics is formally derived from wave optics. The interaction of optical waves with simple optical components such as mirrors, prisms, lenses, and various graded-index elements is examined in Sec. 2.4. Interference, an important manifestation of the wave nature of light, is the subject of Secs. 2.5 and 2.6, where polychromatic and pulsed light are discussed.

2.1 POSTULATES OF WAVE OPTICS

The Wave Equation

Light propagates in the form of waves. In free space, light waves travel with speed c_o. A homogeneous transparent medium such as glass is characterized by a single constant, its refractive index n (≥ 1). In a medium of refractive index n, light waves travel with a reduced speed

$$\boxed{c = \frac{c_o}{n}\,.} \qquad \text{(2.1-1)}$$

Speed of Light in a Medium

An optical wave is described mathematically by a real function of position $\mathbf{r} = (x, y, z)$ and time t, denoted $u(\mathbf{r}, t)$ and known as the **wavefunction**. It satisfies a partial differential equation called the **wave equation**,

$$\boxed{\nabla^2 u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0\,,} \qquad \text{(2.1-2)}$$

Wave Equation

where ∇^2 is the Laplacian operator, which is $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ in Cartesian coordinates. Any function that satisfies (2.1-2) represents a possible optical wave.

Because the wave equation is linear, the **principle of superposition** applies: if $u_1(\mathbf{r},t)$ and $u_2(\mathbf{r},t)$ represent possible optical waves, then $u(\mathbf{r},t) = u_1(\mathbf{r},t)+u_2(\mathbf{r},t)$ also represents a possible optical wave.

At the boundary between two different media, the wavefunction changes in a way that depends on their refractive indices. However, the laws that govern this change depend on the physical significance assigned to the wavefunction which, as will be seen in Chapter 5, is an electromagnetic-field component. The underlying physical origin of the refractive index derives from electromagnetic optics (Sec. 5.5B).

The wave equation is also approximately applicable for media with refractive indices that are position dependent, provided that the variation is slow within distances of the order of a wavelength. The medium is then said to be locally homogeneous. For such media, n in (2.1-1) and c in (2.1-2) are simply replaced by the appropriate position-dependent functions $n(\mathbf{r})$ and $c(\mathbf{r})$, respectively.

Intensity, Power, and Energy

The optical **intensity** $I(\mathbf{r},t)$, defined as the optical power per unit area (units of watts/cm^2), is proportional to the average of the squared wavefunction:

$$I(\mathbf{r},t) = 2\langle u^2(\mathbf{r},t)\rangle\,. \tag{2.1-3}$$

Optical Intensity

The operation $\langle\cdot\rangle$ denotes averaging over a time interval much longer than the time of an optical cycle, but much shorter than any other time of interest (such as the duration of a pulse of light). The duration of an optical cycle is very short: $2\times10^{-15}\,\text{s} = 2\,\text{fs}$ for light of wavelength 600 nm, as an example. This concept is further elucidated in Sec. 2.6. The quantity $I(\mathbf{r},t)$ is sometimes also called the **irradiance**.

Although the physical significance of the wavefunction $u(\mathbf{r},t)$ has not been explicitly specified, (2.1-3) represents its connection with a physically measurable quantity — the optical intensity. There is some arbitrariness in the definition of the wavefunction and its relation to the intensity. For example, (2.1-3) could have been written without the factor 2 and the wavefunction scaled by a factor $\sqrt{2}$, in which case the intensity would remain the same. The choice of the factor 2 in (2.1-3) will later prove convenient, however.

The optical **power** $P(t)$ (units of watts) flowing into an area A normal to the direction of propagation of light is the integrated intensity

$$P(t) = \int_A I(\mathbf{r},t)\,dA\,. \tag{2.1-4}$$

The optical **energy** E (units of joules) collected in a given time interval is the integral of the optical power over the time interval.

2.2 MONOCHROMATIC WAVES

A monochromatic wave is represented by a wavefunction with harmonic time dependence,

$$u(\mathbf{r},t) = a(\mathbf{r})\cos[2\pi\nu t + \varphi(\mathbf{r})]\,, \tag{2.2-1}$$

as illustrated in Fig. 2.2-1(a), where

$$\begin{aligned} a(\mathbf{r}) &= \text{amplitude} \\ \varphi(\mathbf{r}) &= \text{phase} \\ \nu &= \text{frequency (cycles/s or Hz)} \\ \omega &= 2\pi\nu = \text{angular frequency (radians/s or s}^{-1}\text{)} \\ T &= 1/\nu = 2\pi/\omega = \text{period (s).} \end{aligned}$$

Both the amplitude and phase are generally position dependent, but the wavefunction is a harmonic function of time with frequency ν at all positions. Optical waves have frequencies that lie in the range 3×10^{11} to 3×10^{16} Hz, as depicted in Fig. 2.0-1.

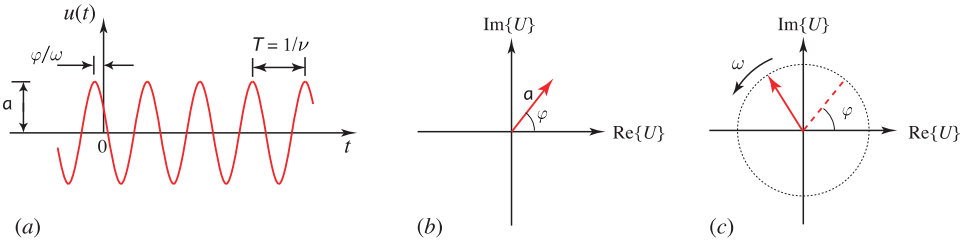


Figure 2.2-1 Representations of a monochromatic wave at a fixed position $\mathbf{r}$: (a) the wavefunction $u(t)$ is a harmonic function of time; (b) the complex amplitude $U = a\exp(j\varphi)$ is a fixed phasor; (c) the complex wavefunction $U(t) = U\exp(j2\pi\nu t)$ is a phasor rotating with angular velocity $\omega = 2\pi\nu$ radians/s.

A. Complex Representation and the Helmholtz Equation

Complex Wavefunction

It is convenient to represent the real wavefunction $u(\mathbf{r}, t)$ in (2.2-1) in terms of a complex function

$$U(\mathbf{r}, t) = a(\mathbf{r})\exp[j\varphi(\mathbf{r})]\exp(j2\pi\nu t)\,, \tag{2.2-2}$$

so that

$$u(\mathbf{r}, t) = \text{Re}\{U(\mathbf{r}, t)\} = \tfrac{1}{2}[U(\mathbf{r}, t) + U^*(\mathbf{r}, t)]\,, \tag{2.2-3}$$

where the symbol * signifies complex conjugation. The function $U(\mathbf{r}, t)$, known as the **complex wavefunction**, describes the wave completely; the **wavefunction** $u(\mathbf{r}, t)$ is simply its real part. Like the wavefunction $u(\mathbf{r}, t)$, the complex wavefunction $U(\mathbf{r}, t)$ must also satisfy the wave equation

$$\boxed{\nabla^2 U - \frac{1}{c^2}\frac{\partial^2 U}{\partial t^2} = 0\,.} \tag{2.2-4}$$

Wave Equation

The two functions satisfy the same boundary conditions.

Complex Amplitude

Equation (2.2-2) may be written in the form

$$U(\mathbf{r},t) = U(\mathbf{r})\exp(j2\pi\nu t)\,, \tag{2.2-5}$$

where the time-independent factor $U(\mathbf{r}) = a(\mathbf{r})\exp[j\varphi(\mathbf{r})]$ is referred to as the **complex amplitude** of the wave. The wavefunction $u(\mathbf{r},t)$ is therefore related to the complex amplitude by

$$u(\mathbf{r},t) = \text{Re}\{U(\mathbf{r})\exp(j2\pi\nu t)\} = \tfrac{1}{2}[U(\mathbf{r})\exp(j2\pi\nu t) + U^*(\mathbf{r})\exp(-j2\pi\nu t)]\,. \tag{2.2-6}$$

At a given position $\mathbf{r}$, the complex amplitude $U(\mathbf{r})$ is a complex variable [depicted in Fig. 2.2-1(*b*)] whose magnitude $|U(\mathbf{r})| = a(\mathbf{r})$ is the amplitude of the wave and whose argument $\arg\{U(\mathbf{r})\} = \varphi(\mathbf{r})$ is the phase. The complex wavefunction $U(\mathbf{r},t)$, shown in Fig. 2.2-1(*c*), is represented graphically by a phasor that rotates with angular velocity $\omega = 2\pi\nu$ radians/s. Its initial value at $t = 0$ is the complex amplitude $U(\mathbf{r})$.

The Helmholtz Equation

Substituting $U(\mathbf{r},t) = U(\mathbf{r})\exp(j2\pi\nu t)$ from (2.2-5) into the wave equation (2.2-4) leads to a differential equation for the complex amplitude $U(\mathbf{r})$:

$$\boxed{\nabla^2 U + k^2 U = 0\,,} \tag{2.2-7}$$

Helmholtz Equation

which is known as the **Helmholtz equation**, where

$$\boxed{k = \frac{2\pi\nu}{c} = \frac{\omega}{c}} \tag{2.2-8}$$

Wavenumber

is referred to as the **wavenumber**. Different solutions are obtained from different boundary conditions.

Optical Intensity

The optical intensity is determined by inserting (2.2-1) into (2.1-3):

$$\begin{aligned} 2u^2(\mathbf{r},t) &= 2a^2(\mathbf{r})\cos^2[2\pi\nu t + \varphi(\mathbf{r})] \\ &= |U(\mathbf{r})|^2\{1 + \cos(2[2\pi\nu t + \varphi(\mathbf{r})])\}\,. \end{aligned} \tag{2.2-9}$$

Averaging (2.2-9) over a time longer than an optical period, $1/\nu$, causes the second term of (2.2-9) to vanish, whereupon

$$\boxed{I(\mathbf{r}) = |U(\mathbf{r})|^2\,.} \tag{2.2-10}$$

Optical Intensity

The optical intensity of a monochromatic wave is the absolute square of its complex amplitude.

The intensity of a monochromatic wave does *not* vary with time.

Wavefronts

The wavefronts are the surfaces of equal phase, $\varphi(\mathbf{r}) =$ constant. The constants are often taken to be multiples of 2π so that $\varphi(\mathbf{r}) = 2\pi q$, where q is an integer. The wavefront normal at position $\mathbf{r}$ is parallel to the gradient vector $\nabla\varphi(\mathbf{r})$ (a vector that has components $\partial\varphi/\partial x$, $\partial\varphi/\partial y$, and $\partial\varphi/\partial z$ in a Cartesian coordinate system). It represents the direction at which the rate of change of the phase is maximum.

Summary

- A monochromatic wave of frequency ν is described by a *complex wavefunction* $U(\mathbf{r},t) = U(\mathbf{r})\exp(j2\pi\nu t)$, which satisfies the wave equation.
- The *complex amplitude* $U(\mathbf{r})$ satisfies the Helmholtz equation; its magnitude $|U(\mathbf{r})|$ and argument $\arg\{U(\mathbf{r})\}$ are the *amplitude* and *phase* of the wave, respectively. The optical *intensity* is $I(\mathbf{r}) = |U(\mathbf{r})|^2$. The *wavefronts* are the surfaces of constant phase, $\varphi(\mathbf{r}) = \arg\{U(\mathbf{r})\} = 2\pi q$ ($q =$ integer).
- The *wavefunction* $u(\mathbf{r},t)$ is the real part of the complex wavefunction, $u(\mathbf{r},t) = \mathrm{Re}\{U(\mathbf{r},t)\}$. The wavefunction also satisfies the wave equation.

B. Elementary Waves

The simplest solutions of the Helmholtz equation in a homogeneous medium are the plane wave and the spherical wave.

The Plane Wave

The plane wave has complex amplitude

$$U(\mathbf{r}) = A\exp(-j\mathbf{k}\cdot\mathbf{r}) = A\exp\left[-j(k_x x + k_y y + k_z z)\right] , \tag{2.2-11}$$

where A is a complex constant called the **complex envelope** that represents the strength of the wave, and $\mathbf{k} = (k_x, k_y, k_z)$ is called the **wavevector**.[†] Substituting (2.2-11) into the Helmholtz equation (2.2-7) yields the relation $k_x^2 + k_y^2 + k_z^2 = k^2$, so that the magnitude of the wavevector $\mathbf{k}$ is the wavenumber k.

Since the phase of the wave is $\arg\{U(\mathbf{r})\} = \arg\{A\} - \mathbf{k}\cdot\mathbf{r}$, the surfaces of constant phase (wavefronts) obey $\mathbf{k}\cdot\mathbf{r} = k_x x + k_y y + k_z z = 2\pi q + \arg\{A\}$ with q integer. This is the equation describing parallel planes perpendicular to the wavevector $\mathbf{k}$ (hence the name "plane wave"). Consecutive planes are separated by a distance $\lambda = 2\pi/k$, so that

$$\boxed{\lambda = \frac{c}{\nu}} , \tag{2.2-12}$$

Wavelength

where λ is called the **wavelength**. The plane wave has a constant intensity $I(\mathbf{r}) = |A|^2$ everywhere in space so that it carries infinite power. This wave is clearly an idealization since it exists everywhere and at all times.

[†] The complex wavefunction for a monochromatic plane wave is written in a form commonly used in electrical engineering: $U(\mathbf{r},t) = A\exp[j(\omega t - \mathbf{k}\cdot\mathbf{r})]$. In the physics literature, this same wave is usually written as $U(\mathbf{r},t) = A\exp[-i(\omega t - \mathbf{k}\cdot\mathbf{r})]$; correspondence is attained by simply replacing i with $-j$, where $i = j = \sqrt{-1}$. This choice has no bearing on the final result, as is evidenced by observing that the wavefunction $u(\mathbf{r},t)$ in (2.2-13) takes the form of a cosine function, for which $\cos(x) = \cos(-x)$.

If the z axis is taken along the direction of the wavevector $\mathbf{k}$, then $U(\mathbf{r}) = A\exp(-jkz)$ and the corresponding wavefunction obtained from (2.2-6) is

$$u(\mathbf{r},t) = |A|\cos\left[2\pi\nu t - kz + \arg\{A\}\right] = |A|\cos\left[2\pi\nu(t - z/c) + \arg\{A\}\right]. \tag{2.2-13}$$

The wavefunction is therefore periodic in time with period $1/\nu$, and periodic in space with period $2\pi/k$, which is equal to the wavelength λ (see Fig. 2.2-2). Since the phase of the complex wavefunction, $\arg\{U(\mathbf{r},t)\} = 2\pi\nu(t - z/c) + \arg\{A\}$, varies with time and position as a function of the variable $t - z/c$ (see Fig. 2.2-2), c is called the **phase velocity** of the wave.

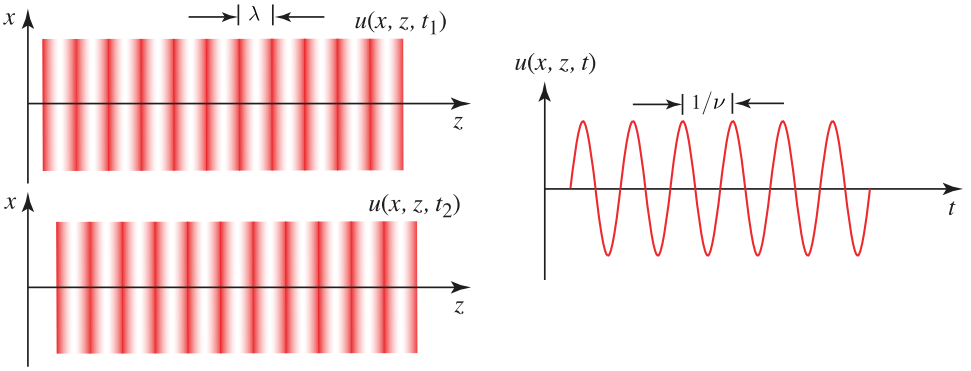


Figure 2.2-2 The wavefunction of a plane wave traveling in the z direction, schematically drawn as a graded red pattern, is a periodic function of z with spatial period λ, and a periodic function of t with temporal period $1/\nu$. The surfaces of constant phase (wavefronts) comprise a parallel set of planes normal to the z axis. The wavelengths displayed in Fig. 2.0-1 are in free space ($\lambda = \lambda_o$).

In a medium of refractive index n, the wave has phase velocity $c = c_o/n$ and wavelength $\lambda = c/\nu = c_o/n\nu$, so that $\lambda = \lambda_o/n$ where $\lambda_o = c_o/\nu$ is the wavelength in free space. Thus, for a given frequency ν, the wavelength in the medium is reduced relative to that in free space by the factor n. As a consequence, the wavenumber $k = 2\pi/\lambda$ is increased relative to that in free space ($k_o = 2\pi/\lambda_o$) by the factor n.

As a monochromatic wave propagates through media of different refractive indices its frequency remains the same, but its velocity, wavelength, and wavenumber are altered:

$$c = \frac{c_o}{n}, \qquad \lambda = \frac{\lambda_o}{n}, \qquad k = nk_o. \tag{2.2-14}$$

The Spherical Wave

Another simple solution of the Helmholtz equation (in spherical coordinates) is the spherical wave complex amplitude

$$U(\mathbf{r}) = \frac{A_0}{r}\exp(-jkr), \tag{2.2-15}$$

where r is the distance from the origin, $k = 2\pi\nu/c = \omega/c$ is the wavenumber, and A_0 is a constant. The intensity $I(\mathbf{r}) = |A_0|^2/r^2$ is inversely proportional to the square of the distance. Taking $\arg\{A_0\} = 0$ for simplicity, the wavefronts are the surfaces $kr = 2\pi q$ or $r = q\lambda$, where q is an integer. These are concentric spheres separated by a radial distance $\lambda = 2\pi/k$ that advance radially at the phase velocity c (Fig. 2.2-3).

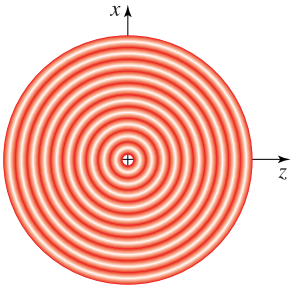


Figure 2.2-3 Cross section of the wavefunction of a spherical wave. The associated wavefronts are a set of concentric spheres.

A spherical wave originating at the position $\mathbf{r}_0$ has a complex amplitude $U(\mathbf{r}) = (A_0/|\mathbf{r}-\mathbf{r}_0|)\exp(-jk\,|\mathbf{r}-\mathbf{r}_0|)$. Its wavefronts are spheres centered about $\mathbf{r}_0$. A wave with complex amplitude $U(\mathbf{r}) = (A_0/r)\exp(+jkr)$ is a spherical wave traveling inwardly (toward the origin) instead of outwardly (away from the origin).

Fresnel Approximation of the Spherical Wave: The Paraboloidal Wave

Let us examine a spherical wave (originating at $\mathbf{r} = 0$) at points $\mathbf{r} = (x, y, z)$ that are sufficiently close to the z axis but far from the origin, so that $\sqrt{x^2+y^2} \ll z$. The paraxial approximation of ray optics (Sec. 1.2) would be applicable were these points the endpoints of rays beginning at the origin. Denoting $\theta^2 = (x^2+y^2)/z^2 \ll 1$, we use an approximation based on the Taylor-series expansion:

$$r = \sqrt{x^2+y^2+z^2} = z\sqrt{1+\theta^2} = z\left(1+\frac{\theta^2}{2}-\frac{\theta^4}{8}+\cdots\right)$$
$$\approx z\left(1+\frac{\theta^2}{2}\right) = z+\frac{x^2+y^2}{2z}. \qquad (2.2\text{-}16)$$

This expression, $r \approx z + (x^2+y^2)/2z$, is now substituted into the phase of $U(\mathbf{r})$ in (2.2-15). A less accurate expression, $r \approx z$, can be substituted for the magnitude since it is less sensitive to errors than is the phase. The result is known as the **Fresnel approximation** of a spherical wave:

$$U(\mathbf{r}) \approx \frac{A_0}{z}\exp(-jkz)\exp\left[-jk\frac{x^2+y^2}{2z}\right]. \qquad (2.2\text{-}17)$$

Fresnel Approximation of a Spherical Wave

This approximation plays an important role in simplifying the theory of optical-wave transmission through apertures (**diffraction**), as discussed in Chapter 4.

The complex amplitude in (2.2-17) may be viewed as representing a plane wave $A_0\exp(-jkz)$ modulated by the factor $(1/z)\exp[-jk(x^2+y^2)/2z]$, which involves the phase $k(x^2+y^2)/2z$. This phase factor serves to bend the planar wavefronts of the plane wave into paraboloidal surfaces (Fig. 2.2-4), since the equation of a paraboloid of revolution is $(x^2+y^2)/z = \text{constant}$. In this region the spherical wave is well approximated by a **paraboloidal wave**. When z becomes very large, the paraboloidal phase factor in (2.2-17) approaches 0 so that the overall phase of the wave becomes kz. Since the magnitude A_0/z varies slowly with z, the spherical wave eventually approaches the plane wave $\exp(-jkz)$, as illustrated in Fig. 2.2-4.

The condition of validity for the Fresnel approximation is *not* simply that $\theta^2 \ll 1$, however. Although the third term of the series expansion, $\theta^4/8$, may be very small

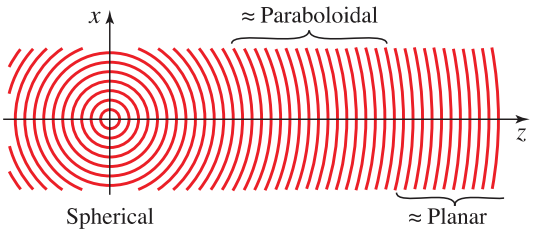


Figure 2.2-4 A spherical wave may be approximated at points near the z axis and sufficiently far from the origin by a paraboloidal wave. For points very far from the origin, the spherical wave approaches a plane wave.

in comparison with the second and first terms, when multiplied by kz it can become comparable to π. The approximation used in the foregoing is therefore valid when $kz\theta^4/8 \ll \pi$, or $(x^2+y^2)^2 \ll 4z^3\lambda$. For points (x, y) lying within a circle of radius a centered about the z axis, the validity condition is thus $a^4 \ll 4z^3\lambda$ or

$$\frac{N_{\mathrm{F}}\theta_m^2}{4} \ll 1\,, \tag{2.2-18}$$

where $\theta_m = a/z$ is the maximum angle and

$$\boxed{N_{\mathrm{F}} = \frac{a^2}{\lambda z}} \tag{2.2-19}$$

Fresnel Number

is known as the **Fresnel number**.

EXERCISE 2.2-1

Validity of the Fresnel Approximation. Determine the radius of a circle within which a spherical wave of wavelength $\lambda = 633$ nm, originating at a distance 1 m away, may be approximated by a paraboloidal wave. Determine the maximum angle θ_m and the Fresnel number N_{F}.

C. Paraxial Waves

A wave is said to be paraxial if its wavefront normals are paraxial rays. One way of constructing a paraxial wave is to start with a plane wave $A\exp(-jkz)$, regard it as a "carrier" wave, and modify or "modulate" its complex envelope A, making it a slowly varying function of position, $A(\mathbf{r})$, so that the complex amplitude of the modulated wave becomes

$$U(\mathbf{r}) = A(\mathbf{r})\exp(-jkz)\,. \tag{2.2-20}$$

The variation of the envelope $A(\mathbf{r})$ and its derivative with position z must be slow within the distance of a wavelength $\lambda = 2\pi/k$ so that the wave approximately maintains its underlying plane-wave nature.

The wavefunction of a paraxial wave, $u(\mathbf{r}, t) = |A(\mathbf{r})|\,\cos[2\pi\nu t - kz + \arg\{A(\mathbf{r})\}]$, is sketched in Fig. 2.2-5(a) as a function of z at $t = 0$ and $x = y = 0$. It is a sinusoidal function of z with amplitude $|A(0, 0, z)|$ and phase $\arg\{A(0, 0, z)\}$, both of which vary slowly with z. Since the phase $\arg\{A(x, y, z)\}$ changes little within the distance of a wavelength, the planar wavefronts $kz = 2\pi q$ of the carrier plane wave bend only slightly, so that their normals form paraxial rays [Fig. 2.2-5(b)].

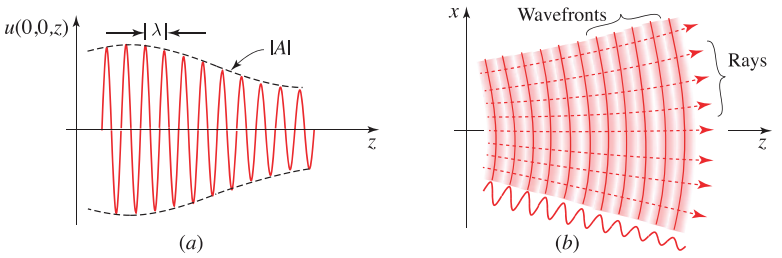


Figure 2.2-5 (a) Wavefunction of a paraxial wave at point on the z axis as a function of the axial distance z. (b) The wavefronts and wavefront normals of a paraxial wave in the x–z plane.

The Paraxial Helmholtz Equation

For the paraxial wave (2.2-20) to satisfy the Helmholtz equation (2.2-7), the complex envelope $A(\mathbf{r})$ must satisfy another partial differential equation that is obtained by substituting (2.2-20) into (2.2-7). The assumption that $A(\mathbf{r})$ varies slowly with respect to z signifies that within a distance $\Delta z = \lambda$, the change ΔA is much smaller than A itself, i.e., $\Delta A \ll A$. This inequality of complex variables applies to the magnitudes of the real and imaginary parts separately. Since $\Delta A = (\partial A/\partial z)\Delta z = (\partial A/\partial z)\lambda$, it follows that $\partial A/\partial z \ll A/\lambda = Ak/2\pi$, so that

$$\frac{\partial A}{\partial z} \ll kA\,. \tag{2.2-21}$$

The derivative $\partial A/\partial z$ itself must also vary slowly within the distance λ, so that $\partial^2 A/\partial z^2 \ll k\,\partial A/\partial z$, which provides

$$\frac{\partial^2 A}{\partial z^2} \ll k^2 A\,. \tag{2.2-22}$$

Substituting (2.2-20) into (2.2-7), and neglecting $\partial^2 A/\partial z^2$ in comparison with $k\,\partial A/\partial z$ or $k^2 A$, leads to a partial differential equation for the complex envelope $A(\mathbf{r})$:

$$\boxed{\nabla_T^2 A - j\,2k\frac{\partial A}{\partial z} = 0\,,} \tag{2.2-23}$$

Paraxial Helmholtz Equation

where $\nabla_T^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the transverse Laplacian operator.

Equation (2.2-23) is the **slowly varying envelope approximation** of the Helmholtz equation. We shall simply call it the **paraxial Helmholtz equation**. It bears some similarity to the Schrödinger equation of quantum physics [see (14.1-1)]. The simplest solution of the paraxial Helmholtz equation is the paraboloidal wave (Exercise 2.2-2), which is the paraxial approximation of a spherical wave. One of the most interesting and useful solutions, however, is the **Gaussian beam**, to which Chapter 3 is devoted.

EXERCISE 2.2-2

The Paraboloidal Wave and the Gaussian Beam. Verify that a paraboloidal wave with the complex envelope $A(\mathbf{r}) = (A_0/z)\exp[-jk(x^2+y^2)/2z]$ [see (2.2-17)] satisfies the paraxial Helmholtz equation (2.2-23). Show that the wave whose complex envelope is given by $A(\mathbf{r}) =$

$[A_1/q(z)]\exp[-jk(x^2+y^2)/2q(z)]$, where $q(z) = z + jz_0$ and z_0 is a constant, also satisfies the paraxial Helmholtz equation. This wave, called the Gaussian beam, is the subject of Chapter 3. Sketch the intensity of the Gaussian beam in the plane $z = 0$.

*2.3 RELATION BETWEEN WAVE OPTICS AND RAY OPTICS

We proceed to show that ray optics emerges as the limit of wave optics when the wavelength $\lambda_o \to 0$. Consider a monochromatic wave of free-space wavelength λ_o in a medium with refractive index $n(\mathbf{r})$ that varies sufficiently slowly with position so that the medium may be regarded as locally homogeneous. We write the complex amplitude in (2.2-5) in the form

$$U(\mathbf{r}) = \mathrm{a}(\mathbf{r})\exp[-jk_o S(\mathbf{r})]\,, \tag{2.3-1}$$

where $\mathrm{a}(\mathbf{r})$ is its magnitude, $-k_o S(\mathbf{r})$ is its phase, and $k_o = 2\pi/\lambda_o$ is the free-space wavenumber. We assume that $\mathrm{a}(\mathbf{r})$ varies sufficiently slowly with $\mathbf{r}$ that it may be regarded as constant within the distance of a wavelength λ_o.

The wavefronts are the surfaces $S(\mathbf{r}) =$ constant and the wavefront normals point in the direction of the gradient vector ∇S. In the neighborhood of a given position $\mathbf{r}_0$, the wave can be locally regarded as a plane wave with amplitude $\mathrm{a}(\mathbf{r}_0)$ and wavevector $\mathbf{k}$ with magnitude $k = n(\mathbf{r}_0)k_o$ and direction parallel to the gradient vector ∇S at $\mathbf{r}_0$. A different neighborhood exhibits a local plane wave of different amplitude and different wavevector.

In ray optics it was shown that the optical rays are normal to the equilevel surfaces of a function $S(\mathbf{r})$ called the eikonal (see Sec. 1.3C). We therefore associate the local wavevectors (wavefront normals) in wave optics with the ray of ray optics and recognize that the function $S(\mathbf{r})$, which is proportional to the phase of the wave, is nothing but the eikonal of ray optics (Fig. 2.3-1). This association has a formal mathematical basis, as will be demonstrated shortly. With this analogy, ray optics can serve to determine the approximate effects of optical components on the wavefront normals, as illustrated in Fig. 2.3-1.

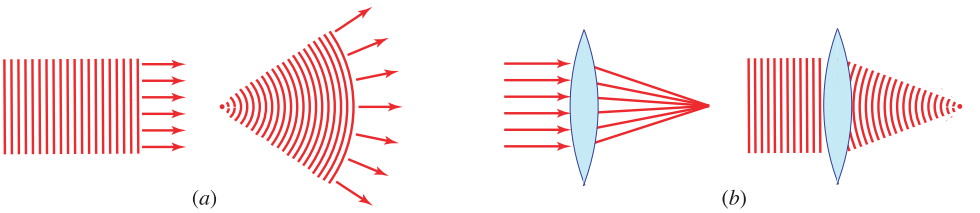


Figure 2.3-1 (*a*) The rays of ray optics are orthogonal to the wavefronts of wave optics (see also Fig. 1.3-10). (*b*) The effect of a lens on rays and wavefronts.

The Eikonal Equation

Substituting (2.3-1) into the Helmholtz equation (2.2-7) provides

$$k_o^2\left[n^2 - |\nabla S|^2\right]\mathrm{a} + \nabla^2\mathrm{a} - jk_o\left[2\,\nabla S\cdot\nabla\mathrm{a} + \mathrm{a}\,\nabla^2 S\right] = 0\,, \tag{2.3-2}$$

where $\mathrm{a} = \mathrm{a}(\mathbf{r})$ and $S = S(\mathbf{r})$. The real and imaginary parts of the left-hand side of (2.3-2) must both vanish. Equating the real part to zero and using $k_o = 2\pi/\lambda_o$, we

obtain

$$|\nabla S|^2 = n^2 + \left(\frac{\lambda_o}{2\pi}\right)^2 \frac{\nabla^2 a}{a}. \tag{2.3-3}$$

The assumption that a varies slowly over the distance λ_o means that $\lambda_o^2 \nabla^2 a / a \ll 1$, so that the second term of the right-hand side may be neglected in the limit $\lambda_o \to 0$, whereupon

$$\boxed{|\nabla S|^2 \approx n^2.} \tag{2.3-4}$$

Eikonal Equation

This is the eikonal equation (1.3-20), which may be regarded as the main postulate of ray optics (Fermat's principle can be derived from the eikonal equation and *vice versa*).

Thus, the scalar function $S(\mathbf{r})$, which is proportional to the phase in wave optics, is the eikonal of ray optics. This is also consistent with the observation that in ray optics $S(\mathbf{r}_B) - S(\mathbf{r}_A)$ equals the optical pathlength between the points $\mathbf{r}_A$ and $\mathbf{r}_B$.

The eikonal equation is the limit of the Helmholtz equation when $\lambda_o \to 0$. Given $n(\mathbf{r})$ we may use the eikonal equation to determine $S(\mathbf{r})$. By equating the imaginary part of (2.3-2) to zero, we obtain a relation between a and S, thereby permitting us to determine the wavefunction.

2.4 SIMPLE OPTICAL COMPONENTS

In this section we examine the effects of optical components, such as mirrors, transparent plates, prisms, and lenses, on optical waves.

A. Reflection and Refraction

Reflection from a Planar Mirror

A plane wave of wavevector $\mathbf{k}_1$ is incident onto a planar mirror located in free space in the $z = 0$ plane. A reflected plane wave of wavevector $\mathbf{k}_2$ is created. The angles of incidence and reflection are θ_1 and θ_2, as illustrated in Fig. 2.4-1. The sum of the two waves satisfies the Helmholtz equation if the wavenumber is the same, i.e., if $k_1 = k_2 = k_o$. Certain boundary conditions must be satisfied at the surface of the mirror. Since these conditions are the same at all points (x, y), it is necessary that the phases of the two waves match, i.e.,

$$\mathbf{k}_1 \cdot \mathbf{r} = \mathbf{k}_2 \cdot \mathbf{r} \qquad \text{for all } \mathbf{r} = (x, y, 0). \tag{2.4-1}$$

This *phase-matching condition* may also be regarded as matching of the tangential components of the two wavevectors in the mirror plane. Substituting $\mathbf{r} = (x, y, 0)$, $\mathbf{k}_1 = (k_o \sin\theta_1, 0, k_o \cos\theta_1)$, and $\mathbf{k}_2 = (k_o \sin\theta_2, 0, -k_o \cos\theta_2)$ into (2.4-1), we obtain $k_o x \sin\theta_1 = k_o x \sin\theta_2$, from which $\theta_1 = \theta_2$, so that the angles of incidence and reflection must be equal. Thus, the law of reflection of optical rays is applicable to the wavevectors of plane waves.

Reflection and Refraction at a Planar Dielectric Boundary

We now consider a plane wave of wavevector $\mathbf{k}_1$ incident on a planar boundary between two homogeneous media of refractive indices n_1 and n_2. The boundary lies in the

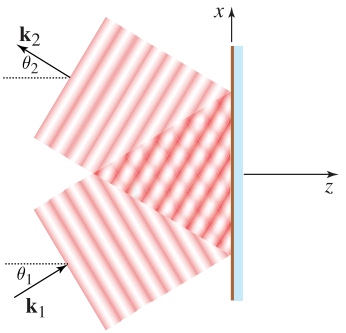


Figure 2.4-1 Reflection of a plane wave from a planar mirror. Phase matching at the surface of the mirror requires that the angles of incidence and reflection be equal.

$z = 0$ plane (Fig. 2.4-2). Refracted and reflected plane waves of wavevectors $\mathbf{k}_2$ and $\mathbf{k}_3$ emerge. The combination of the three waves satisfies the Helmholtz equation everywhere if each of the waves has the appropriate wavenumber in the medium in which it propagates ($k_1 = k_3 = n_1 k_o$ and $k_2 = n_2 k_0$).

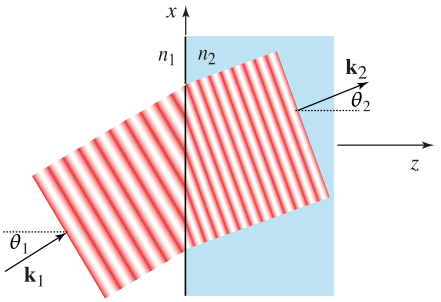


Figure 2.4-2 Refraction of a plane wave at a dielectric boundary. The wavefronts are matched at the boundary so that the distance between wavefronts for the incident wave, $\lambda_1/\sin\theta_1 = \lambda_o/n_1 \sin\theta_1$, equals that for the refracted wave, $\lambda_2/\sin\theta_2 = \lambda_o/n_2 \sin\theta_2$, from which Snell's law follows.

Since the boundary conditions are invariant to x and y, it is necessary that the phases of the three waves match, i.e.,

$$\mathbf{k}_1 \cdot \mathbf{r} = \mathbf{k}_2 \cdot \mathbf{r} = \mathbf{k}_3 \cdot \mathbf{r} \qquad \text{for all } \mathbf{r} = (x, y, 0)\,. \tag{2.4-2}$$

This *phase-matching condition* is tantamount to matching the tangential components of the three wavevectors at the boundary plane. Since $\mathbf{k}_1 = (n_1 k_o \sin\theta_1, 0, n_1 k_o \cos\theta_1)$, $\mathbf{k}_3 = (n_1 k_o \sin\theta_3, 0, -n_1 k_o \cos\theta_3)$, and $\mathbf{k}_2 = (n_2 k_o \sin\theta_2, 0, n_2 k_o \cos\theta_2)$, where θ_1, θ_2, and θ_3 are the angles of incidence, refraction, and reflection, respectively, it follows from (2.4-2) that $\theta_1 = \theta_3$ and $n_1 \sin\theta_1 = n_2 \sin\theta_2$. These are the laws of reflection and refraction (Snell's law) of ray optics, now applicable to the wavevectors.

It is not possible to determine the amplitudes of the reflected and refracted waves using scalar wave optics since the boundary conditions are not completely specified in this theory. This will be achieved in Sec. 6.2 using electromagnetic optics (Chapters 5 and 6).

B. Transmission Through Optical Components

We now proceed to examine the transmission of optical waves through transparent optical components such as plates, prisms, and lenses. The effect of reflection at the surfaces of these components will be ignored, since it cannot be properly accounted for using the scalar wave theory of light. Nor can the effect of absorption in the material, which is relegated to Sec. 5.5. The principal emphasis here is on the phase shift introduced by these components and on the associated wavefront bending.

Transparent Plate

Consider first the transmission of a plane wave through a transparent plate of refractive index n and thickness d surrounded by free space. The surfaces of the plate are the planes $z = 0$ and $z = d$ and the incident wave travels in the z direction (Fig. 2.4-3). Let $U(x, y, z)$ be the complex amplitude of the wave. Since external and internal reflections are ignored, $U(x, y, z)$ is assumed to be continuous at the boundaries. The ratio $\mathrm{t}(x, y) = U(x, y, d)/U(x, y, 0)$ therefore represents the **complex amplitude transmittance** of the plate; it permits us to determine $U(x, y, d)$ for arbitrary $U(x, y, 0)$ at the input. The effect of reflection is considered in Sec. 6.2 and the effect of multiple internal reflections within the plate is examined in Sec. 11.1.

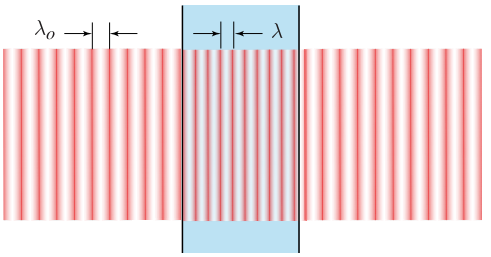


Figure 2.4-3 Transmission of a plane wave through a transparent plate.

Once inside the plate, the wave continues to propagate as a plane wave with wavenumber nk_o, so that $U(x, y, z)$ is proportional to $\exp(-jnk_o z)$. Thus, the ratio $U(x, y, d)/U(x, y, 0) = \exp(-jnk_o d)$, so that

$$\boxed{\mathrm{t}(x, y) = \exp(-jnk_o d)\,.} \qquad (2.4\text{-}3)$$

Transmittance
Transparent Plate

The plate is seen to introduce a phase shift $nk_o d = 2\pi(d/\lambda)$.

If the incident plane wave makes an angle θ with respect to the z axis and has wavevector $\mathbf{k}$ (Fig. 2.4-4), the refracted and transmitted waves are also plane waves with wavevectors $\mathbf{k}_1$ and $\mathbf{k}$ and angles θ_1 and θ, respectively, where θ_1 and θ are related by Snell's law: $\sin\theta = n\sin\theta_1$. The complex amplitude $U(x, y, z)$ inside the plate is now proportional to $\exp(-j\mathbf{k}_1 \cdot \mathbf{r}) = \exp[-jnk_o(z\cos\theta_1 + x\sin\theta_1)]$, so that the complex amplitude transmittance of the plate $U(x, y, d)/U(x, y, 0)$ is

$$\mathrm{t}(x, y) = \exp\left(-jnk_o d\cos\theta_1\right)\,. \qquad (2.4\text{-}4)$$

If the angle of incidence θ is small (i.e., if the incident wave is *paraxial*), then $\theta_1 \approx \theta/n$ is also small and the approximation $\cos\theta_1 \approx 1 - \frac{1}{2}\theta_1^2$ yields $\mathrm{t}(x, y) \approx \exp(-jnk_o d)\exp(jk_o\theta^2 d/2n)$. If the plate is *sufficiently thin*, and the angle θ is *sufficiently small* such that $k_o\theta^2 d/2n \ll 2\pi$ [or $(d/\lambda_o)\theta^2/2n \ll 1$], then the transmittance of the plate may be approximated by (2.4-3). Under these conditions the transmittance of the plate is approximately independent of the angle θ.

Thin Transparent Plate of Varying Thickness

We now determine the amplitude transmittance of a thin transparent plate whose thickness $d(x, y)$ varies smoothly as a function of x and y, assuming that the incident wave is an arbitrary *paraxial* wave. The plate lies between the planes $z = 0$ and $z = d_0$, which are regarded as the boundaries encasing the optical component (Fig. 2.4-5).

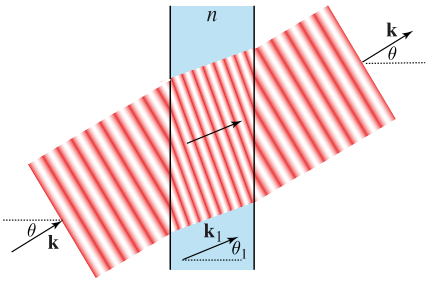


Figure 2.4-4 Transmission of an oblique plane wave through a thin transparent plate.

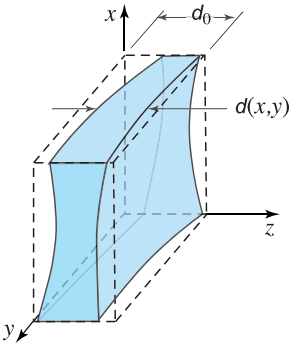


Figure 2.4-5 A transparent plate of varying thickness.

In the vicinity of the position $(x, y, 0)$ the incident paraxial wave may be regarded locally as a plane wave traveling along a direction that makes a small angle with the z axis. It crosses a thin plate of material of thickness $d(x, y)$ surrounded by thin layers of air of total thickness $d_0 - d(x, y)$. In accordance with the approximate relation (2.4-3), the local transmittance is the product of the transmittances of a thin layer of air of thickness $d_0 - d(x, y)$ and a thin layer of material of thickness $d(x, y)$, so that $\mathrm{t}(x, y) \approx \exp[-jnk_o d(x, y)] \exp[-jk_o(d_0 - d(x, y))]$, from which

$$\mathrm{t}(x, y) \approx h_0 \exp[-j(n-1)k_o d(x, y)]\,, \tag{2.4-5}$$

Transmittance
Variable-Thickness Plate

where $h_0 = \exp(-jk_o d_0)$ is a constant phase factor. This relation is valid in the paraxial approximation (where all angles θ are small) and when the thickness d_0 is sufficiently small so that $(d_0/\lambda_o)\theta^2/2n \ll 1$.

EXERCISE 2.4-1

Transmission Through a Prism. Use (2.4-5) to show that the complex amplitude transmittance of a thin inverted prism with small apex angle $\alpha \ll 1$ and thickness d_0 (Fig. 2.4-6) is $\mathrm{t}(x, y) = h_0 \exp[-j\,(n-1)\alpha\, k_o x]$, where $h_0 = \exp(-jk_o d_0)$. The transmittance is independent of y since the prism extends in the y direction. What is the effect of the prism on an incident plane wave traveling in the z direction? Compare your results with that obtained via the ray-optics model, as provided in (1.2-7).

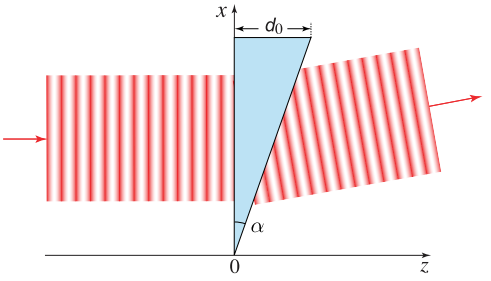


Figure 2.4-6 Transmission of a plane wave through a thin prism.

EXAMPLE 2.4-1. ***Transmission Through a Biprism and an Axicon.*** The biprism depicted in Fig. 1.2-12(a) comprises an inverted prism, such as that illustrated in Fig. 2.4-6, juxtaposed with an identical uninverted prism. Taking its thickness to be d_0 and its edge angle $\alpha \ll 1$, the results of Exercise 2.4-1 generalize to $\mathrm{t}(x,y) = h_0\{\exp[-j\,(n-1)\alpha\,k_o x] + \exp[+j\,(n-1)\alpha\,k_o x]\} = 2h_0\cos[(n-1)\alpha\,k_o x]$, with $h_0 = \exp(-jk_o d_0)$. The biprism thus converts an incident plane wave into a pair of waves that are tilted with respect to each other. The Fresnel biprism portrayed in Fig. 1.2-12(b) behaves in the same way.

The cone-shaped axicon shown in Fig. 1.2-12(c) is constructed by rotating the prism cross section depicted in Fig. 2.4-6 about a horizontal axis located at its top edge, from $\phi = -\pi$ to π. At any angle ϕ, the cross section of this device is an isosceles triangle of thickness d_0 and edge angle $\alpha \ll 1$. Using polar coordinates and integrating the results presented in Exercise 2.4-1 over ϕ provides $\mathrm{t}(x,y) = h_0\int_{-\pi}^{\pi}\exp[-j\,(n-1)\,\alpha\,(k_o\cos\phi)\,x - j\,(n-1)\,\alpha\,(k_o\sin\phi)\,y]\,d\phi = h_0\int_{-\pi}^{\pi}\exp[-j\,(n-1)\,\alpha\,k_o\sqrt{x^2+y^2}\,\sin(\phi+\theta)]\,d\phi$. Since the integration is over 2π, the integral is independent of θ. Given that $\int_{-\pi}^{\pi}\exp(-ju\sin\phi)\,d\phi = 2\pi J_0(u)$, where $J_0(u)$ is the Bessel function of the first kind and zeroth order, the amplitude transmittance may be rewritten as $\mathrm{t}(x,y) = 2\pi h_0 J_0[(n-1)\,\alpha\,k_o\sqrt{x^2+y^2}\,]$. The axicon thus converts an incident plane wave into an infinite number of plane waves, all directed toward its central axis in the form of a cone of half angle $(n-1)\alpha$. This device may be used to convert a plane wave into a Bessel beam (see Sec. 3.5A and Example 4.3-5).

Thin Lens

The general expression (2.4-5) for the complex amplitude transmittance of a thin transparent plate of variable thickness is now applied to the plano-convex thin lens shown in Fig. 2.4-7. Since the lens is the cap of a sphere of radius R, the thickness at the point (x,y) is $d(x,y) = d_0 - \overline{PQ} = d_0 - (R - \overline{QC})$, or

$$d(x,y) = d_0 - \left[R - \sqrt{R^2 - (x^2+y^2)}\right]. \tag{2.4-6}$$

This expression may be simplified by considering only points for which x and y are sufficiently small in comparison with R so that $x^2 + y^2 \ll R^2$. In that case

$$\sqrt{R^2-(x^2+y^2)} = R\sqrt{1-\frac{x^2+y^2}{R^2}} \approx R\left(1-\frac{x^2+y^2}{2R^2}\right), \tag{2.4-7}$$

where we have used the same Taylor-series expansion that led to the Fresnel approximation of a spherical wave in (2.2-17). Using this approximation in (2.4-6) then provides

$$d(x,y) \approx d_0 - \frac{x^2+y^2}{2R}. \tag{2.4-8}$$

Finally, substitution into (2.4-5) yields

$$\boxed{\mathrm{t}(x,y) \approx h_0 \exp\left[jk_o\frac{x^2+y^2}{2f}\right],} \tag{2.4-9}$$

Transmittance Thin Lens

where

$$f = \frac{R}{n-1} \tag{2.4-10}$$

is the focal length of the lens (see Sec. 1.2C) and $h_0 = \exp(-jnk_o d_0)$ is another constant phase factor that is usually of no significance.

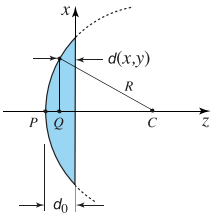


Figure 2.4-7 A plano-convex thin lens. The lens imparts a phase proportional to $x^2 + y^2$ to an incident plane wave, thereby transforming it into a paraboloidal wave centered at a distance f from the lens (see Exercise 2.4-3).

EXERCISE 2.4-2

Double-Convex Lens. Show that the complex amplitude transmittance of the double-convex lens (also called a spherical lens) shown in Fig. 2.4-8 is given by (2.4-9) with

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{2.4-11}$$

You may prove this either by using the general formula (2.4-5) or by regarding the double-convex lens as a cascade of two plano-convex lenses. Recall that, by convention, the radius of a convex/concave surface is positive/negative, so that R_1 is positive and R_2 is negative for the lens displayed in Fig. 2.4-8. The parameter f is recognized as the focal length of the lens [see (1.2-12)].

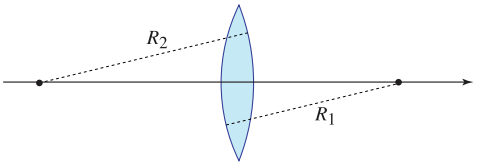


Figure 2.4-8 A double-convex lens.

EXERCISE 2.4-3

Focusing of a Plane Wave by a Thin Lens. Show that when a plane wave is transmitted through a thin lens of focal length f in a direction parallel to the axis of the lens, it is converted into a paraboloidal wave (the Fresnel approximation of a spherical wave) centered about a point at a distance f from the lens, as illustrated in Fig. 2.4-9. What is the effect of the lens on a plane wave incident at a small angle θ?

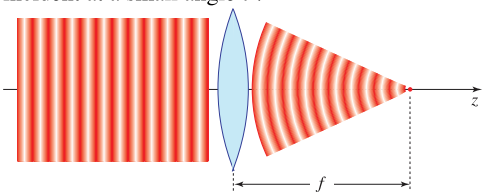


Figure 2.4-9 A thin lens transforms a plane wave into a paraboloidal wave.

EXERCISE 2.4-4

Imaging Property of a Lens. Show that a paraboloidal wave centered at the point P_1 (Fig. 2.4-10) is converted by a lens of focal length f into a paraboloidal wave centered at P_2, where $1/z_1 + 1/z_2 = 1/f$, a formula known as the **imaging equation**.

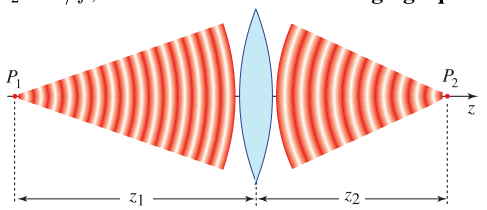


Figure 2.4-10 A lens transforms a paraboloidal wave into another paraboloidal wave. The two waves are centered at distances that satisfy the imaging equation.

Diffraction Gratings

A **diffraction grating** is an optical component that serves to periodically modulate the phase or amplitude of an incident wave. It can be made of a transparent plate with periodically varying thickness or periodically graded refractive index (see Sec. 2.4C). Repetitive arrays of diffracting elements such as apertures, obstacles, or absorbing elements (see Sec. 4.3) can also be used for this purpose. A reflection diffraction grating is often fabricated from a periodically ruled thin film of aluminum that has been evaporated onto a glass substrate.

Consider a diffraction grating made of a thin transparent plate placed in the $z = 0$ plane whose thickness varies periodically in the x direction with period Λ (Fig. 2.4-11). As will be demonstrated in Exercise 2.4-5, this plate converts an incident plane wave of wavelength $\lambda \ll \Lambda$, traveling at a small angle θ_i with respect to the z axis, into several plane waves at small angles with respect to the z axis:

$$\theta_q \approx \theta_i + q\frac{\lambda}{\Lambda}, \tag{2.4-12}$$

Grating Equation

where $q = 0, \pm 1, \pm 2, \ldots$, is called the diffraction order. Successive diffracted waves are separated by an angle $\theta = \lambda/\Lambda$, as shown schematically in Fig. 2.4-11.

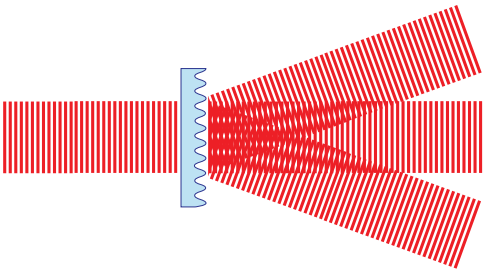


Figure 2.4-11 A thin transparent plate with periodically varying thickness serves as a diffraction grating. It splits an incident plane wave into multiple plane waves traveling in different directions.

EXERCISE 2.4-5

Transmission Through a Diffraction Grating.

(a) The thickness of a thin transparent plate varies sinusoidally in the x direction, $d(x, y) = \frac{1}{2}d_0[1 + \cos(2\pi x/\Lambda)]$, as illustrated in Fig. 2.4-11. Show that the complex amplitude transmittance is $\mathrm{t}(x, y) = h_0 \exp[-j\frac{1}{2}(n-1)k_o d_0 \cos(2\pi x/\Lambda)]$ where $h_0 = \exp[-j\frac{1}{2}(n+1)k_o d_0]$.

(b) Show that an incident plane wave traveling at a small angle θ_i with respect to the z direction is transmitted in the form of a sum of plane waves traveling at angles θ_q given by (2.4-12). *Hint:* Expand the periodic function $\mathrm{t}(x, y)$ in a Fourier series.

Equation (2.4-12) is valid only in the paraxial approximation, when all angles are small, and when the period Λ is much greater than the wavelength λ. A more general analysis of a thin diffraction grating that does not rely on the paraxial approximation reveals that an incident plane wave at an angle θ_i gives rise to a collection of plane waves at angles θ_q that satisfy

$$\sin\theta_q = \sin\theta_i + q\frac{\lambda}{\Lambda}. \tag{2.4-13}$$

This result may be derived by expanding the periodic transmittance $\mathrm{t}(x, y)$ as a sum of Fourier components of the form $\exp(-jq2\pi x/\Lambda)$, where $q = 0, \pm 1, \pm 2, \ldots$ is

the diffraction order. An incident plane wave $\exp(-jkx\sin\theta_i)$, modulated by the harmonic component $\exp(-jq2\pi x/\Lambda)$, generates a transmitted plane wave at the angle θ_q given by $\exp(-jkx\sin\theta_q) \propto \exp(-jkx\sin\theta_i)\exp(-jq2\pi x/\Lambda)$. This leads to the phase-matching condition $k\sin\theta_q = k\sin\theta_o + q2\pi/\Lambda$. Equation (2.4-13) follows since $k = 2\pi/\lambda$; this result is also applicable to waves reflected from the grating.

Diffraction gratings are used as filters and spectrum analyzers. Since the angles θ_q depend on the wavelength λ (and therefore on the frequency ν), an incident polychromatic wave is separated by the grating into its spectral components (Fig. 2.4-12). Diffraction gratings have found numerous applications in spectroscopy.

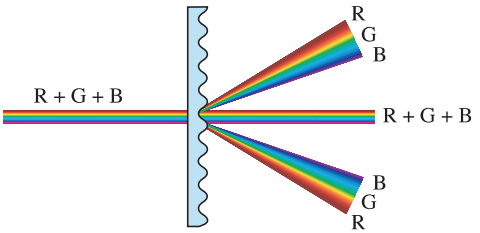


Figure 2.4-12 A diffraction grating directs two waves of different wavelengths, λ_1 and λ_2, into two different directions, θ_1 and θ_2. It therefore serves as a spectrum analyzer or a spectrometer.

C. Graded-Index Optical Components

The effect of a prism, lens, or diffraction grating on an incident optical wave lies in the phase shift it imparts, which serves to bend the wavefront in some prescribed manner. This phase shift is controlled by the variation in the thickness of the material with the transverse distance from the optical axis (linearly, quadratically, or periodically, in the cases of a prism, lens, and diffraction grating, respectively). The same phase shift may instead be introduced by a transparent planar plate of fixed thickness but with varying refractive index. This is a result of the fact that the thickness and refractive index appear as a product in (2.4-3).

The complex amplitude transmittance of a thin transparent planar plate of thickness d_0 and graded refractive index $n(x,y)$ is, from (2.4-3),

$$\mathrm{t}(x,y) = \exp\left[-jn(x,y)k_o d_0\right] . \tag{2.4-14}$$

Transmittance
Graded-Index Thin Plate

By selecting the appropriate variation of $n(x,y)$ with x and y, the action of any constant-index thin optical component can be reproduced, as demonstrated in Exercise 2.4-6.

EXERCISE 2.4-6

Graded-Index Lens. Show that a thin plate of uniform thickness d_0 (Fig. 2.4-13) and quadratically graded refractive index $n(x,y) = n_0[1 - \frac{1}{2}\alpha^2(x^2+y^2)]$, with $\alpha d_0 \ll 1$, acts as a lens of focal length $f = 1/n_0 d_0 \alpha^2$ (see Exercise 1.3-1).

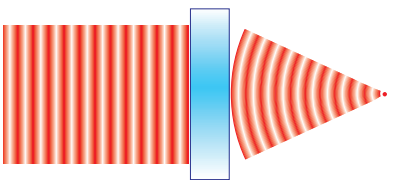


Figure 2.4-13 A graded-index plate acts as a lens.

2.5 INTERFERENCE

When two or more optical waves are simultaneously present in the same region of space and time, the total wavefunction is the sum of the individual wavefunctions. This basic principle of superposition follows from the linearity of the wave equation. For monochromatic waves of the same frequency, the superposition principle carries over to the complex amplitudes, which follows from the linearity of the Helmholtz equation.

The superposition principle does not apply to the optical intensity since the intensity of the sum of two or more waves is not necessarily the sum of their intensities. The disparity is associated with interference. The phenomenon of interference cannot be explained on the basis of ray optics since it is dependent on the phase relationship between the superposed waves.

In this section we examine the interference between two or more monochromatic waves of the same frequency. The interference of waves of different frequencies is discussed in Sec. 2.6.

A. Interference of Two Waves

When two monochromatic waves with complex amplitudes $U_1(\mathbf{r})$ and $U_2(\mathbf{r})$ are superposed, the result is a monochromatic wave of the same frequency that has a complex amplitude

$$U(\mathbf{r}) = U_1(\mathbf{r}) + U_2(\mathbf{r})\,. \tag{2.5-1}$$

In accordance with (2.2-10), the intensities of the constituent waves are $I_1 = |U_1|^2$ and $I_2 = |U_2|^2$, while the intensity of the total wave is

$$I = |U|^2 = |U_1 + U_2|^2 = |U_1|^2 + |U_2|^2 + U_1^* U_2 + U_1 U_2^*. \tag{2.5-2}$$

The explicit dependence on $\mathbf{r}$ has been omitted for convenience. Substituting

$$U_1 = \sqrt{I_1}\,\exp(j\varphi_1) \qquad \text{and} \qquad U_2 = \sqrt{I_2}\,\exp(j\varphi_2) \tag{2.5-3}$$

into (2.5-2), where φ_1 and φ_2 are the phases of the two waves, we obtain

$$\boxed{I = I_1 + I_2 + 2\sqrt{I_1 I_2}\,\cos\varphi\,,} \tag{2.5-4}$$

Interference Equation

with

$$\varphi = \varphi_2 - \varphi_1\,. \tag{2.5-5}$$

This relation, called the **interference equation**, can also be understood in terms of the geometry of the phasor diagram displayed in Fig. 2.5-1(a), which demonstrates that the magnitude of the phasor U is sensitive not only to the magnitudes of the constituent phasors but also to the phase difference φ.

It is clear, therefore, that the intensity of the sum of the two waves is *not* the sum of their intensities [Fig. 2.5-1(b)]; an additional term, attributed to **interference** between the two waves, is present in (2.5-4). This term may be positive or negative, corresponding to constructive or destructive interference, respectively. If $I_1 = I_2 = I_0$, for example, then (2.5-4) yields $I = 2I_0(1 + \cos\varphi) = 4I_0\cos^2(\varphi/2)$, so that for $\varphi = 0$, $I = 4I_0$ (i.e., the total intensity is four times the intensity of each of the

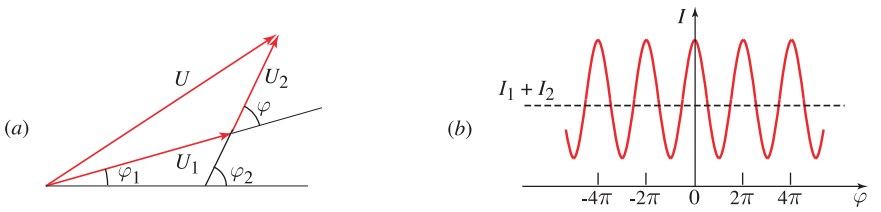


Figure 2.5-1 (a) Phasor diagram for the superposition of two waves of intensities I_1 and I_2 and phase difference $\varphi = \varphi_2 - \varphi_1$. ($b$) Dependence of the total intensity I on the phase difference φ.

superposed waves). For $\varphi = \pi$, on the other hand, the superposed waves cancel one another and the total intensity $I = 0$. Complete cancellation of the intensity in a region of space is generally not possible unless the intensities of the constituent superposed waves are equal. When $\varphi = \pi/2$ or $3\pi/2$, the interference term vanishes and $I = 2I_0$; for these special phase relationships the total intensity is the sum of the constituent intensities. The strong dependence of the intensity I on the phase difference φ permits us to measure phase differences by detecting light intensity. This principle is used in numerous optical systems.

Interference is accompanied by a spatial redistribution of the optical intensity without a violation of power conservation. For example, the two waves may have uniform intensities I_1 and I_2 in a particular plane, but as a result of a position-dependent phase difference φ, the total intensity can be smaller than $I_1 + I_2$ at some positions and larger at others, with the total power (integral of the intensity) conserved.

Interference is not observed under ordinary lighting conditions since the random fluctuations of the phases φ_1 and φ_2 cause the phase difference φ to assume random values that are uniformly distributed between 0 and 2π, so that $\cos\varphi$ averages to 0 and the interference term washes out. Light with such randomness is said to be *partially coherent* and Chapter 12 is devoted to its study. The analysis carried out here, and in subsequent chapters prior to Chapter 12, assume that the light is *coherent*, and therefore deterministic.

Interferometers

Consider the superposition of two plane waves, each of intensity I_0, propagating in the z direction, and assume that one wave is delayed by a distance d with respect to the other so that $U_1 = \sqrt{I_0}\exp(-jkz)$ and $U_2 = \sqrt{I_0}\exp[-jk(z-d)]$. The intensity I of the sum of these two waves can be determined by substituting $I_1 = I_2 = I_0$ and $\varphi = kd = 2\pi d/\lambda$ into the interference equation (2.5-4),

$$I = 2I_0\left[1 + \cos\left(2\pi\frac{d}{\lambda}\right)\right]. \tag{2.5-6}$$

The dependence of I on the delay d is sketched in Fig. 2.5-2. When the delay is an integer multiple of λ, complete constructive interference occurs and the total intensity $I = 4I_0$. On the other hand, when d is an odd integer multiple of $\lambda/2$, complete destructive interference occurs and $I = 0$. The average intensity is the sum of the two intensities, i.e., $2I_0$.

An **interferometer** is an optical instrument that splits a wave into two waves using a beamsplitter, delays them by unequal distances, redirects them using mirrors, recombines them using another (or the same) beamsplitter, and detects the intensity of their superposition. Three important examples are illustrated in Fig. 2.5-3: the **Mach–Zehnder interferometer**, the **Michelson interferometer**, and the **Sagnac interferometer**.

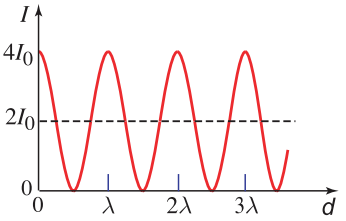
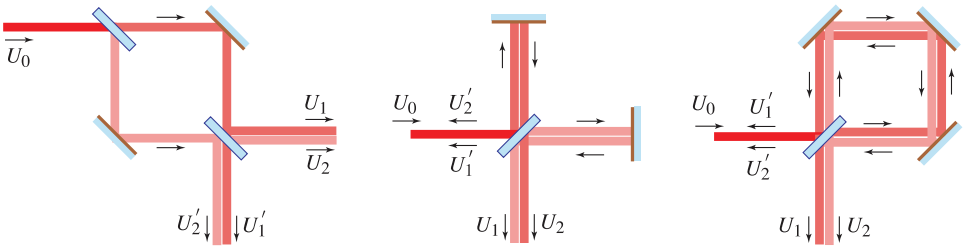


Figure 2.5-2 Dependence of the intensity I of the superposition of two waves, each of intensity I_0, on the delay distance d. When the delay distance is a multiple of λ, the interference is constructive; when it is an odd multiple of $\lambda/2$, the interference is destructive.



(*a*) Mach-Zehnder (*b*) Michelson (*c*) Sagnac

Figure 2.5-3 Interferometers: A wave U_0 is split into two waves U_1 and U_2 (they are shown as shaded light and dark for ease of visualization but are actually congruent). After traveling through different paths, the waves are recombined into a superposition wave $U = U_1 + U_2$ whose intensity is recorded. The waves are split and recombined using beamsplitters. In the Sagnac interferometer the two waves travel through the same path, but in opposite directions.

Since the intensity I is sensitive to the phase $\varphi = 2\pi d/\lambda = 2\pi n d/\lambda_0 = 2\pi n \nu d/c_o$, where d is the difference between the distances traveled by the two waves, the interferometer can be used to measure small changes in the distance d, the refractive index n, or the wavelength λ_o (or frequency ν). For example, if $d/\lambda_o = 10^4$, a change of the refractive index of only $\Delta n = 10^{-4}$ corresponds to an easily observable phase change $\Delta\varphi = 2\pi$. The phase φ also changes by a full 2π if d changes by a wavelength λ. An incremental change of the frequency $\Delta\nu = c/d$ has the same effect.

Interferometers have numerous applications. These include the determination of distance in metrological applications such as strain measurement and surface profiling; refractive-index measurements; and spectrometry for the analysis of polychromatic light (see Sec. 12.2B). In the Sagnac interferometer the optical paths are identical but opposite in direction, so that rotation of the interferometer results in a phase shift φ proportional to the angular velocity of rotation. This system can therefore be used as a gyroscope. Because of its precision, optical interferometry is also being co-opted to detect the passage of gravitational waves, as discussed subsequently.

Finally, we demonstrate that energy conservation in an interferometer requires that the phases of the waves reflected and transmitted at a beamsplitter differ by $\pi/2$. Each of the interferometers considered in Fig. 2.5-3 has an output wave $U = U_1 + U_2$ that exits from one side of the beamsplitter and also another output wave $U' = U_1' + U_2'$ that exits from the opposite side. Energy conservation dictates that the sum of the intensities of these two waves must equal the intensity of the incident wave, so that if one output wave has high intensity by virtue of constructive interference, the other must have low intensity by virtue of destructive interference. This complementarity can only be achieved if the phase differences φ and φ', associated with the components of output waves U and U', respectively, differ by π. Since the components of U and the components of U' experience the same pathlength differences, and the same numbers of reflections from mirrors, the π phase difference must be attributable to different phases introduced by the beamsplitter upon reflection and transmission. Examination

of the three interferometers in Fig. 2.5-3 reveals that for one output wave, each of the components is transmitted through the beamsplitter once and reflected from it once, so that no phase difference is introduced. However, for the other output wave, one component is transmitted twice and the other is reflected twice, thereby introducing the phase difference of π. It follows that the phases of the reflected and transmitted waves at a beamsplitter differ by $\pi/2$. This important property of the beamsplitter is considered in more detail in Example 7.1-6.

Interference of Two Oblique Plane Waves

Consider now the interference of two plane waves of equal intensities: one propagating in the z direction, $U_1 = \sqrt{I_0}\exp(-jkz)$; the other propagating at an angle θ with respect to the z axis, in the x–z plane, $U_2 = \sqrt{I_0}\exp[-j(k\cos\theta\ z + k\sin\theta\ x)]$, as illustrated in Fig. 2.5-4. At the $z = 0$ plane the two waves have a phase difference $\varphi = k\sin\theta\ x$, for which the interference equation (2.5-4) yields a total intensity

$$I = 2I_0\left[1 + \cos(k\sin\theta\ x)\right]. \tag{2.5-7}$$

This pattern varies sinusoidally with x, with period $2\pi/k\sin\theta = \lambda/\sin\theta$, as shown in Fig. 2.5-4. If $\theta = 30°$, for example, the period is 2λ. This suggests a method of printing a sinusoidal pattern of high resolution for use as a diffraction grating. It also suggests a method of monitoring the angle of arrival θ of a wave by mixing it with a reference wave and recording the resultant intensity distribution. As discussed in Sec. 4.5, this is the principle that lies behind holography.

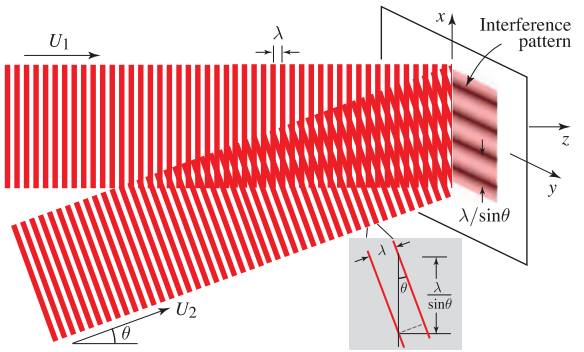


Figure 2.5-4 The interference of two plane waves traveling at an angle θ with respect to each other results in a sinusoidal intensity pattern in the x direction with period $\lambda/\sin\theta$.

EXERCISE 2.5-1

Interference of a Plane Wave and a Spherical Wave. A plane wave traveling along the z direction with complex amplitude $A_1\exp(-jkz)$, and a spherical wave centered at $z = 0$ and approximated by the paraboloidal wave of complex amplitude $(A_2/z)\exp(-jkz)\exp[-jk(x^2 + y^2)/2z]$ [see (2.2-17)], interfere in the $z = d$ plane. Derive an expression for the total intensity $I(x, y, d)$. Assuming that the two waves have the same intensities at the $z = d$ plane, verify that the locus of points of zero intensity is a set of concentric rings, as illustrated in Fig. 2.5-5.

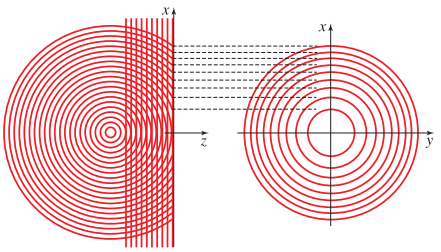


Figure 2.5-5 The interference of a plane wave and a spherical wave creates a pattern of concentric rings (illustrated at the plane $z = d$).

EXERCISE 2.5-2

Interference of Two Spherical Waves. Two spherical waves of equal intensity I_0, originating at the points $(-a, 0, 0)$ and $(a, 0, 0)$, interfere in the plane $z = d$ as illustrated in Fig. 2.5-6. This double-pinhole system is similar to that used by Thomas Young in his celebrated double-slit experiment in which he demonstrated interference. Use the paraboloidal approximation for the spherical waves to show that the intensity at the plane $z = d$ is

$$I(x, y, d) \approx 2I_0 \left(1 + \cos \frac{2\pi x \theta}{\lambda}\right), \tag{2.5-8}$$

where the angle subtended by the centers of the two waves at the observation plane is $\theta \approx 2a/d$. The intensity pattern is periodic with period λ/θ.

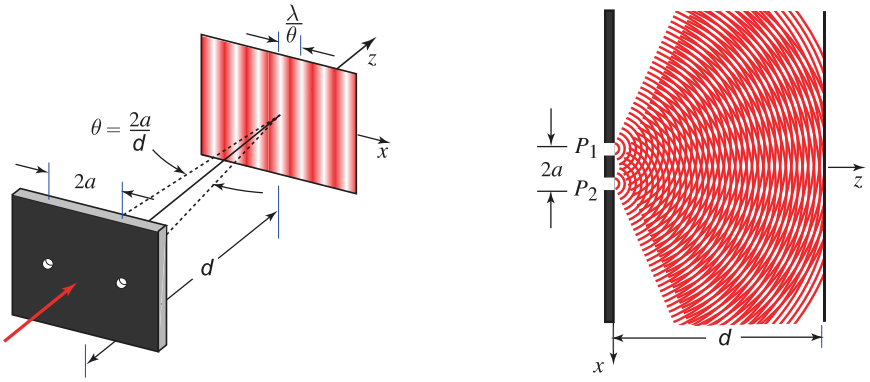


Figure 2.5-6 Interference of two spherical waves of equal intensities originating at the points P_1 and P_2. The two waves can be obtained by permitting a plane wave to impinge on two pinholes in a screen. The light intensity at an observation plane a large distance d from the pinholes takes the form of a sinusoidal interference pattern, with period $\approx \lambda/\theta$, along the direction of the line connecting the pinholes.

B. Multiple-Wave Interference

The superposition of M monochromatic waves of the same frequency, with complex amplitudes $U_1, U_2, \ldots, U_M$, gives rise to a wave whose frequency remains the same and whose complex amplitude is given by $U = U_1 + U_2 + \cdots + U_M$. Knowledge of the intensities of the individual waves, $I_1, I_2, \ldots, I_M$, is not sufficient to determine the total intensity $I = |U|^2$ since the relative phases must also be known. The role played by the phase is dramatically illustrated in the following examples.

Interference of M Waves with Equal Amplitudes and Equal Phase Differences

We first examine the interference of M waves with complex amplitudes

$$U_m = \sqrt{I_0}\,\exp[j(m-1)\varphi], \qquad m = 1, 2, \ldots, M\,. \tag{2.5-9}$$

The waves have equal intensities I_0, and phase difference φ between successive waves, as illustrated in Fig. 2.5-7(a). To derive an expression for the intensity of the superposition, it is convenient to introduce the quantity $h = \exp(j\varphi)$ whereupon $U_m = \sqrt{I_0}\,h^{m-1}$. The complex amplitude of the superposed wave is then

$$\begin{aligned} U &= \sqrt{I_0}\,(1 + h + h^2 + \cdots + h^{M-1}) = \sqrt{I_0}\,\frac{1-h^M}{1-h} \\ &= \sqrt{I_0}\,\frac{1-\exp(jM\varphi)}{1-\exp(j\varphi)}\,, \end{aligned} \tag{2.5-10}$$

which has the corresponding intensity

$$I = |U|^2 = I_0 \left| \frac{\exp(-jM\varphi/2) - \exp(jM\varphi/2)}{\exp(-j\varphi/2) - \exp(j\varphi/2)} \right|^2, \tag{2.5-11}$$

whence

$$\boxed{I = I_0 \frac{\sin^2(M\varphi/2)}{\sin^2(\varphi/2)}\,.} \tag{2.5-12}$$

Interference of M Waves

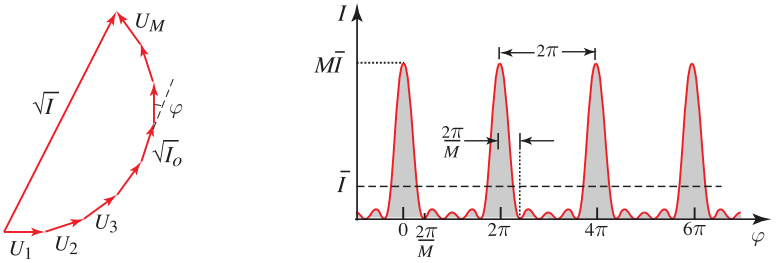


Figure 2.5-7 (a) The sum of M phasors of equal magnitudes and equal phase differences. (b) The intensity I as a function of φ. The peak intensity occurs when all the phasors are aligned; it is then M times greater than the mean intensity $\bar{I} = MI_0$. In this example $M = 5$.

The intensity I is evidently strongly dependent on the phase difference φ, as illustrated in Fig. 2.5-7(b) for $M = 5$. When $\varphi = 2\pi q$, where q is an integer, all the phasors are aligned so that the amplitude of the total wave is M times that of an individual component, and the intensity reaches its peak value of $M^2 I_0$. The mean intensity averaged over a uniform distribution of φ is $\bar{I} = (1/2\pi)\int_0^{2\pi} I\,d\varphi = MI_0$, which is the same as the result obtained in the absence of interference. The peak intensity is therefore M times greater than the mean intensity. The sensitivity of the intensity to the

phase is therefore dramatic for large M. At its peak value, the intensity is magnified by a factor M over the mean but it decreases sharply as the phase difference φ deviates slightly from $2\pi q$. In particular, when $\varphi = 2\pi/M$ the intensity becomes zero. It is instructive to compare Fig. 2.5-7(b) for $M = 5$ with Fig. 2.5-2 for $M = 2$.

EXERCISE 2.5-3

Bragg Reflection. Consider light reflected at an angle θ from M parallel reflecting planes separated by a distance Λ, as shown in Fig. 2.5-8. Assume that only a small fraction of the light is reflected from each plane, so that the amplitudes of the M reflected waves are approximately equal. Show that the reflected waves have a phase difference $\varphi = k(2\Lambda \sin\theta)$ and that the angle θ at which the intensity of the total reflected light is maximum satisfies

$$\sin\theta_{\mathcal{B}} = \frac{\lambda}{2\Lambda}. \tag{2.5-13}$$

Bragg Angle

This equation defines the **Bragg angle** $\theta_{\mathcal{B}}$. Such reflections are encountered when light is reflected from a multilayer structure (see Sec. 7.1) or when X-ray waves are reflected from atomic planes in crystalline structures. It also occurs when light is reflected from a periodic structure created by an acoustic wave (see Chapter 20). An exact treatment of Bragg reflection is provided in Sec. 7.1C.

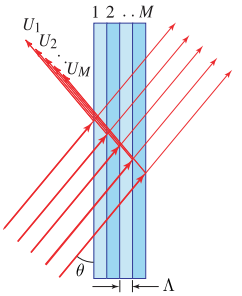


Figure 2.5-8 Reflection of a plane wave from M parallel planes separated from each other by a distance Λ. The reflected waves interfere constructively and yield maximum intensity when the angle θ is the Bragg angle $\theta_{\mathcal{B}}$. Note that θ is defined with respect to the parallel planes.

Fresnel Zone Plate

A **Fresnel zone plate** comprises a set of ring apertures of increasing radii, decreasing widths, and equal areas, as illustrated in Fig. 2.5-9.

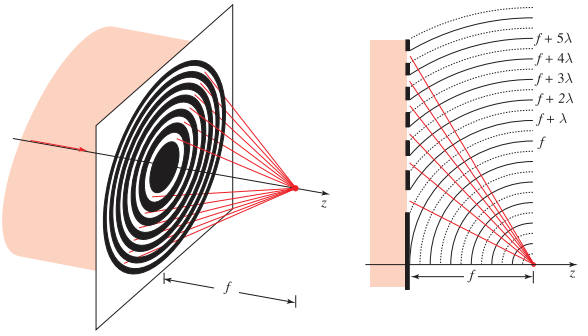


Figure 2.5-9 The Fresnel zone plate serves as a spherical lens with multiple focal lengths.

The structure serves as a spherical lens with multiple focal lengths, as may be understood from the perspective of interference. The center of the mth ring has a radius

ρ_m at the mth peak of the cosine function, i.e., $\pi\rho_m^2/\lambda f = m2\pi$ [see (2.4-9)]. At a focal point $z = f$, the distance R_m to the mth ring is given by $R_m^2 = f^2 + \rho_m^2$, so that $R_m = \sqrt{f^2 + 2m\lambda f}$. If f is sufficiently large so that the angles subtended by the rings are small, then a Taylor-series expansion provides $R_m \approx f + m\lambda$. Thus, the waves transmitted through consecutive rings have pathlengths differing by a wavelength, so that they interfere constructively at the focal point. A similar argument applies for the other foci. The operation of the Fresnel zone plate may also be understood from the perspective of Fourier optics, as explained in Sec. 4.1A.

Interference of an Infinite Number of Waves of Progressively Smaller Amplitudes and Equal Phase Differences

We now examine the superposition of an infinite number of waves with equal phase differences and with amplitudes that decrease at a geometric rate:

$$U_1 = \sqrt{I_0}\,, \quad U_2 = hU_1, \quad U_3 = hU_2 = h^2U_1, \quad \ldots\,, \tag{2.5-14}$$

where $h = |h|e^{j\varphi}$, $|h| < 1$, and I_0 is the intensity of the initial wave. The amplitude of the mth wave is smaller than that of the $(m-1)$st wave by the factor $|h|$ and the phase differs by φ. The phasor diagram is shown in Fig. 2.5-10(a).

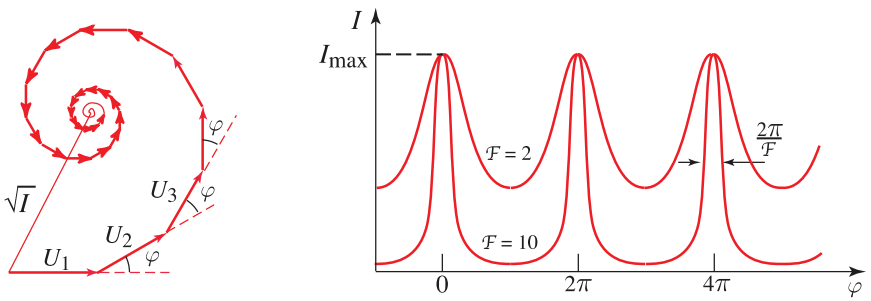


Figure 2.5-10 (a) The sum of an infinite number of phasors whose magnitudes are successively reduced at a geometric rate and whose phase differences φ are equal. (b) Dependence of the intensity I on the phase difference φ for two values of $\mathcal{F}$. Peak values occur at $\varphi = 2\pi q$. The full width at half maximum of each peak is approximately $2\pi/\mathcal{F}$ when $\mathcal{F} \gg 1$. The sharpness of the peaks increases with increasing $\mathcal{F}$.

The superposition wave has a complex amplitude

$$\begin{aligned} U &= U_1 + U_2 + U_3 + \cdots \\ &= \sqrt{I_0}\,(1 + h + h^2 + \cdots) \\ &= \frac{\sqrt{I_0}}{1-h} = \frac{\sqrt{I_0}}{1 - |h|e^{j\varphi}}\,. \end{aligned} \tag{2.5-15}$$

The total intensity is then

$$I = |U|^2 = \frac{I_0}{|1 - |h|e^{j\varphi}|^2} = \frac{I_0}{(1 - |h|\cos\varphi)^2 + |h|^2\sin^2\varphi}\,, \tag{2.5-16}$$

from which

$$I = \frac{I_0}{(1 - |h|)^2 + 4|h|\,\sin^2(\varphi/2)}\,. \tag{2.5-17}$$

It is convenient to write this equation in the form

$$\boxed{I = \frac{I_{\max}}{1 + (2\mathcal{F}/\pi)^2 \sin^2(\varphi/2)}}\,, \qquad I_{\max} = \frac{I_0}{(1-|h|)^2}\,, \tag{2.5-18}$$

Intensity of an Infinite Number of Waves

where the quantity

$$\boxed{\mathcal{F} = \frac{\pi\sqrt{|h|}}{1-|h|}} \tag{2.5-19}$$

Finesse

is a parameter known as the **finesse**.

The intensity I is a periodic function of φ with period 2π, as illustrated in Fig. 2.5-10(b). It reaches its maximum value $I_{\max}$ when $\varphi = 2\pi q$, where q is an integer. This occurs when the phasors align to form a straight line. (This result is not unlike that displayed in Fig. 2.5-7(b) for the interference of M waves of equal amplitudes and equal phase differences.) When the finesse $\mathcal{F}$ is large (i.e., the factor $|h|$ is close to 1), I becomes a sharply peaked function of φ. Consider values of φ near the $\varphi = 0$ peak, as a representative example. For $|\varphi| \ll 1$, $\sin(\varphi/2) \approx \varphi/2$ whereupon (2.5-18) can be written as

$$I \approx \frac{I_{\max}}{1 + (\mathcal{F}/\pi)^2\varphi^2}\,. \tag{2.5-20}$$

The intensity I then decreases to half its peak value when $\varphi = \pi/\mathcal{F}$, so that the full-width at half-maximum (FWHM) of the peak becomes

$$\boxed{\Delta\varphi \approx \frac{2\pi}{\mathcal{F}}\,.} \tag{2.5-21}$$

Width of Interference Pattern

In the regime $\mathcal{F} \gg 1$, we then have $\Delta\varphi \ll 2\pi$ and the assumption that $\varphi \ll 1$ is applicable. The finesse $\mathcal{F}$ is the ratio of the period 2π to the FWHM of the peaks in the interference pattern. It is therefore a measure of the sharpness of the interference function, i.e., the sensitivity of the intensity to deviations of φ from the values $2\pi q$ corresponding to the peaks.

A useful device based on this principle is the Fabry–Perot interferometer. It consists of two parallel mirrors within which light undergoes multiple reflections. In the course of each *round trip*, the light suffers a fixed amplitude reduction $|h| = |\mathrm{r}|$, arising from losses at the mirrors, and a phase shift $\varphi = k2d = 4\pi\nu d/c = 2\pi\nu/(c/2d)$ associated with the propagation, where d is the mirror separation. The total light intensity depends on the phase shift φ in accordance with (2.5-18), attaining maxima when $\varphi/2$ is an integer multiple of π. The proportionality of the phase shift φ to the optical frequency ν shows that the intensity transmission of the Fabry–Perot device will exhibit peaks separated in frequency by $c/2d$. The width of these peaks will be $(c/2d)/\mathcal{F}$, where the finesse $\mathcal{F}$ is governed by the loss via (2.5-19). The Fabry–Perot interferometer, which also serves as a spectrum analyzer, is considered further in Sec. 7.1B. It is commonly used as a resonator for lasers, as discussed in Secs. 11.1 and 16.1A.

EXAMPLE 2.5-1. ***The LIGO Interferometer.*** The LIGO interferometer[†] comprises a Michelson interferometer (MI) with a Fabry–Perot interferometer (FPI) embedded in each of its reflecting arms, as illustrated in Fig. 2.5-11. The MI is sensitive to the phase difference encountered by the optical waves that propagate through its arms; the FPIs serve to amplify the phases in each arm and thereby to significantly increase the sensitivity of the overall instrument.

If the phase shift encountered in a double pass within the FPI is denoted φ, the phase of the overall intracavity reflecting field U is, in accordance with (2.5-15),

$$\arg\{U\} = \arg\left\{\frac{1}{1-|h|e^{j\varphi}}\right\} = \arctan\left\{\frac{|h|\sin\varphi}{1-|h|\cos\varphi}\right\}. \tag{2.5-22}$$

If φ is taken to be an integer multiple of 2π, to which is added a very small double-pass deviation $2\Delta\varphi \ll \pi$, a Taylor-series expansion of (2.5-22) yields $\arg\{U\} \approx 2\Delta\varphi\,|h|/(1-|h|)$. This result is closely related to the finesse of the FPI, $\mathcal{F} = \pi\sqrt{|h|}/(1-|h|)$, as provided in (2.5-19). When $|h| \approx 1$ and the finesse is high, we have $\arg\{U\} \approx 2\Delta\varphi \cdot 1/(1-|h|)$ and $\mathcal{F} \approx \pi/(1-|h|)$, so that $\arg\{U\} \approx (2\mathcal{F}/\pi)\Delta\varphi$. Thus, a very small phase deviation $\Delta\varphi$ imposed on the FPI is amplified by the factor $2\mathcal{F}/\pi$, which is large. This phase amplification results from the many reflections of the light between the mirrors of the FPI, which effectively increases its length and thus its sensitivity.

The interference pattern associated with the Michelson interferometer is characterized by the two-wave interference equation (2.5-4). If the light injected into both of its arms is of equal intensity, i.e., if $I_1 = I_2 = \frac{1}{2}I_0$, (2.5-4) becomes $I = I_0[1+\cos(\varphi_2-\varphi_1)]$. If the interferometer is then operated at a null and the phases for the two arms are taken to be $\varphi_{2,1} = (2\mathcal{F}/\pi)\Delta\varphi_{2,1}$, the LIGO interference pattern is given by

$$1 - \cos\tfrac{2\mathcal{F}}{\pi}\left(\Delta\varphi_2 - \Delta\varphi_1\right). \tag{2.5-23}$$

The LIGO interferometer is thus a factor of $2\mathcal{F}/\pi$ more sensitive to the phase difference $\Delta\varphi_2 - \Delta\varphi_1$ than is a Michelson interferometer with the same arm lengths.

This increased sensitivity is the rationale for using the LIGO interferometer as a gravitational-wave detector. Generated by cataclysmic events in the distant universe, gravitational waves impose a dynamic strain on the fabric of space, which results in differential length variations in the orthogonal arms of the interferometer. This in turn modulates the phase difference $\Delta\varphi_2 - \Delta\varphi_1$, resulting in an overall light intensity whose magnitude is proportional to the gravitational-wave-induced strain. Gravitational waves were first detected by LIGO in 2015, a hundred years after Einstein first predicted their existence.[‡]

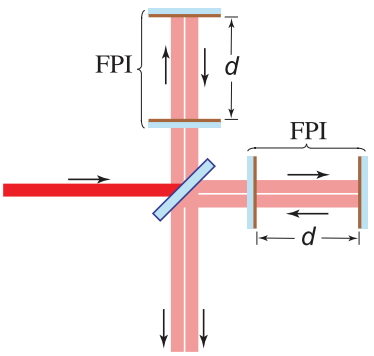


Figure 2.5-11 The LIGO interferometer is a Michelson interferometer (MI) with Fabry–Perot interferometers (FPIs) nested in each of its arms. Each FPI in the advanced-LIGO instrument has a length $d \approx 4$ km and a finesse $\mathcal{F} \approx 450$, so that the enhancement in sensitivity with respect to an ordinary MI is $2\mathcal{F}/\pi \approx 286$. The gravitational wave observed in 2015 imparted to the LIGO interferometer a differential spatial strain $(\Delta d_2 - \Delta d_1)/d \equiv \Delta d/d$ with a magnitude of roughly 5×10^{-22}, which corresponds to a differential length deviation Δd of about 2 am (some 400 times smaller than the radius of a proton). The light source was a 20-W Nd:YAG laser operated at $\lambda_o = c_o/\nu = 1.064$ μm. The corresponding phase difference $\Delta\varphi_2 - \Delta\varphi_1$ thus had a magnitude of $2\pi\nu\Delta d/c_o \approx 1.8 \times 10^{-11}$ rad; its oscillations were in the audio-frequency range.

[†] LIGO is an acronym for Laser Interferometer Gravitational-Wave Observatory, a facility with dual sites in Livingston, Louisiana and Hanford, Washington.

[‡] B. P. Abbott *et al.*, Observation of Gravitational Waves from a Binary Black Hole Merger, *Physical Review Letters*, vol. 116, 061102, 2016. The near-simultaneous detections at both LIGO sites, which are separated by a distance of $\approx$ 3000 km and a time of $\approx$ 10 msec, unequivocally confirmed the cosmological origin of the waves.

2.6 POLYCHROMATIC AND PULSED LIGHT

Since the wavefunction of monochromatic light is a harmonic function of time extending over all time (from $-\infty$ to ∞), it is an idealization that cannot be met in reality. This section is devoted to waves of arbitrary time dependence, including optical pulses of finite time duration. Such waves are polychromatic rather than monochromatic. A more detailed introduction to the optics of pulsed light is provided in Chapter 23.

A. Temporal and Spectral Description

Although a polychromatic wave is described by a wavefunction $u(\mathbf{r}, t)$ with nonharmonic time dependence, it may be expanded as a superposition of harmonic functions, each of which represents a monochromatic wave. Since we already know how monochromatic waves propagate in free space and through various optical components, we can determine the effect of optical systems on polychromatic light by using the principle of superposition.

Fourier methods permit the expansion of an arbitrary function of time $u(t)$, representing the wavefunction $u(\mathbf{r}, t)$ at a fixed position $\mathbf{r}$, as a superposition integral of harmonic functions of different frequencies, amplitudes, and phases:

$$u(t) = \int_{-\infty}^{\infty} v(\nu) \exp(j2\pi\nu t)\, d\nu\,, \tag{2.6-1}$$

where $v(\nu)$ is determined by carrying out the **Fourier transform**

$$v(\nu) = \int_{-\infty}^{\infty} u(t) \exp(-j2\pi\nu t)\, dt\,. \tag{2.6-2}$$

A review of the Fourier transform and its properties is presented in Sec. A.1 of Appendix A. The expansion in (2.6-1) extends over positive and negative frequencies. However, since $u(t)$ is real, $v(-\nu) = v^*(\nu)$ (see Sec. A.1). Thus, the negative-frequency components are not independent; they are simply conjugated versions of the corresponding positive-frequency components.

Complex Representation

It is convenient to represent the real function $u(t)$ in (2.6-1) by a complex function

$$U(t) = 2\int_{0}^{\infty} v(\nu) \exp(j2\pi\nu t)\, d\nu \tag{2.6-3}$$

that includes only the positive-frequency components (multiplied by a factor of 2), and suppresses all the negative frequencies. The Fourier transform of $U(t)$ is therefore a function $V(\nu) = 2v(\nu)$ for $\nu \geq 0$, and 0 for $\nu < 0$.

The real function $u(t)$ can be determined from its complex representation $U(t)$ by simply taking the real part,

$$u(t) = \text{Re}\{U(t)\} = \tfrac{1}{2}[U(t) + U^*(t)]\,. \tag{2.6-4}$$

The complex function $U(t)$ is known as the **complex analytic signal**. The validity of (2.6-4) can be verified by breaking the integral in (2.6-1) into two parts, with limits from 0 to $+\infty$ and from $-\infty$ to 0. The first integral equals $\frac{1}{2}U(t)$ by virtue of (2.6-3), whereas the second is given by

$$\int_{-\infty}^{0} v(\nu)\exp(j2\pi\nu t)\,d\nu = \int_{0}^{\infty} v(-\nu)\exp(-j2\pi\nu t)\,d\nu$$
$$= \int_{0}^{\infty} v^*(\nu)\exp(-j2\pi\nu t)\,d\nu = \tfrac{1}{2}U^*(t)\,. \quad (2.6\text{-}5)$$

The first step above reflects a simple change of variable from ν to $-\nu$, while the second step uses the symmetry relation $v(-\nu) = v^*(\nu)$. The net result is that $u(t)$ can be expressed as a sum of the complex function $\frac{1}{2}U(t)$ and its conjugate, confirming (2.6-4).

As a simple example, the complex representation of the real harmonic function $u(t) = \cos(\omega t)$ is the complex harmonic function $U(t) = \exp(j\omega t)$. This is the complex representation introduced in Sec. 2.2A for monochromatic waves. In fact, the complex representation of a polychromatic wave, as described in this section, is simply a superposition of the complex representations of each of its monochromatic Fourier components.

The complex analytic signal corresponding to the wavefunction $u(\mathbf{r},t)$ is called the **complex wavefunction** $U(\mathbf{r},t)$. Since each of its Fourier components satisfies the wave equation, so too does the complex wavefunction $U(\mathbf{r},t)$,

$$\nabla^2 U - \frac{1}{c^2}\frac{\partial^2 U}{\partial t^2} = 0\,. \quad (2.6\text{-}6)$$
Wave Equation

Figure 2.6-1 shows the magnitudes of the Fourier transforms of the wavefunction $u(\mathbf{r},t)$ and the complex wavefunction $U(\mathbf{r},t)$. In this illustration the optical wave is **quasi-monochromatic**, i.e., it has Fourier components with frequencies confined within a narrow band of width $\Delta\nu$ surrounding a central frequency ν_0, such that $\Delta\nu \ll \nu_0$.

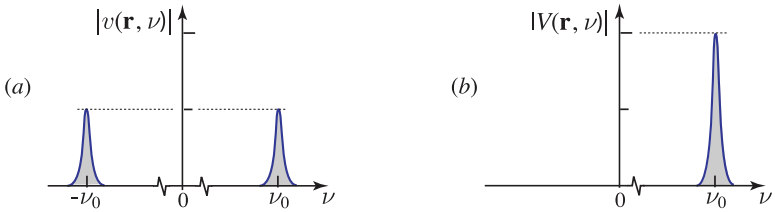


Figure 2.6-1 (*a*) The magnitude $|v(\mathbf{r},\nu)|$ of the Fourier transform of the wavefunction $u(\mathbf{r},t)$. (*b*) The magnitude $|V(\mathbf{r},\nu)|$ of the Fourier transform of the corresponding complex wavefunction $U(\mathbf{r},t)$.

Intensity of a Polychromatic Wave

The optical intensity is related to the wavefunction by (2.1-3):

$$I(\mathbf{r},t) = 2\langle u^2(\mathbf{r},t)\rangle$$
$$= 2\left\langle \left\{\tfrac{1}{2}\left[U(\mathbf{r},t) + U^*(\mathbf{r},t)\right]\right\}^2 \right\rangle$$
$$= \tfrac{1}{2}\left\langle U^2(\mathbf{r},t)\right\rangle + \tfrac{1}{2}\left\langle U^{*2}(\mathbf{r},t)\right\rangle + \left\langle U(\mathbf{r},t)U^*(\mathbf{r},t)\right\rangle . \quad (2.6\text{-}7)$$

For a quasi-monochromatic wave with central frequency ν_0 and spectral width $\Delta\nu \ll \nu_0$, the average $\langle\cdot\rangle$ is taken over a time interval much longer than the time of an optical cycle $1/\nu_0$ but much shorter than $1/\Delta\nu$ (see Sec. 2.1). Since $U(\mathbf{r},t)$ is given by (2.6-4), the term U^2 in (2.6-7) has components oscillating at frequencies $\approx 2\nu_0$. Similarly, the components of U^{*2} oscillate at frequencies $\approx -2\nu_0$. These terms are therefore washed out by the averaging operation. The third term, however, contains only frequency differences, which are of the order of $\Delta\nu \ll \nu_0$. It therefore varies slowly and is unaffected by the time-averaging operation. Thus, the third term in (2.6-7) survives and the light intensity becomes

$$\boxed{I(\mathbf{r},t) = |U(\mathbf{r},t)|^2 \ .} \qquad \begin{matrix}(2.6\text{-}8)\\ \text{Optical Intensity}\end{matrix}$$

The optical intensity of a quasi-monochromatic wave is the absolute square of its complex wavefunction.

The simplicity of this result is, in fact, the rationale for introducing the concept of the complex wavefunction.

Pulsed Plane Wave

The simplest example of pulsed light is a pulsed plane wave. The complex wavefunction has the form

$$U(\mathbf{r},t) = \mathcal{A}\left(t - \frac{z}{c}\right) \exp\left[j2\pi\nu_0\left(t - \frac{z}{c}\right)\right], \qquad (2.6\text{-}9)$$

where the **complex envelope** $\mathcal{A}(t)$ is a time-varying function and ν_0 is the central optical frequency. The monochromatic plane wave is a special case of (2.6-9) for which $\mathcal{A}(t)$ is constant, i.e., $U(\mathbf{r},t) = \mathcal{A}\exp[j2\pi\nu_0(t - z/c)] = \mathcal{A}\exp(-jk_0z)\exp(j\omega_0t)$, where $k_0 = \omega_0/c$ and $\omega_0 = 2\pi\nu_0$.

Since $U(\mathbf{r},t)$ in (2.6-9) is a function of $t - z/c$ it satisfies the wave equation (2.6-6) regardless of the form of the function $\mathcal{A}(\cdot)$ (provided that $d^2\mathcal{A}/dt^2$ exists). This can be verified by direct substitution.

If $\mathcal{A}(t)$ is of finite duration τ, then at any fixed position z the wave lasts for a time period τ, and at any fixed time t it extends over a distance $c\tau$. It is therefore a **wavepacket** of fixed extent traveling in the z direction (Fig. 2.6-2). As an example, a pulse of duration $\tau = 1$ ps extends over a distance $c\tau = 0.3$ mm in free space.

The Fourier transform of the complex wavefunction in (2.6-9) is

$$V(\mathbf{r},\nu) = A(\nu - \nu_0)\exp(-j2\pi\nu z/c)\,, \qquad (2.6\text{-}10)$$

where $A(\nu)$ is the Fourier transform of $\mathcal{A}(t)$. This may be shown by use of the frequency translation property of the Fourier transform (see Sec. A.1 of Appendix A). The complex envelope $\mathcal{A}(t)$ is often slowly varying in comparison with an optical cycle, so that its Fourier transform $A(\nu)$ has a spectral width $\Delta\nu$ much smaller than the central frequency ν_0. The spectral width $\Delta\nu$ is inversely proportional to the temporal width τ. In particular, if $\mathcal{A}(t)$ is Gaussian, then its Fourier transform $A(\nu)$ is also Gaussian. If the temporal and spectral widths are defined as the power-RMS widths, then their product equals $1/4\pi$ (see Sec. A.2 of Appendix A). For example, if $\tau = 1$ ps, then $\Delta\nu = 80$ GHz. If the central frequency ν_0 is 5×10^{14} Hz (corresponding to $\lambda_o = 0.6\ \mu$m), then $\Delta\nu/\nu_0 = 1.6 \times 10^{-4}$, so that the light is quasi-monochromatic. Fig. 2.6-2 illustrates the temporal, spatial, and spectral characteristics of the pulsed plane wave in terms of the wavefunction.

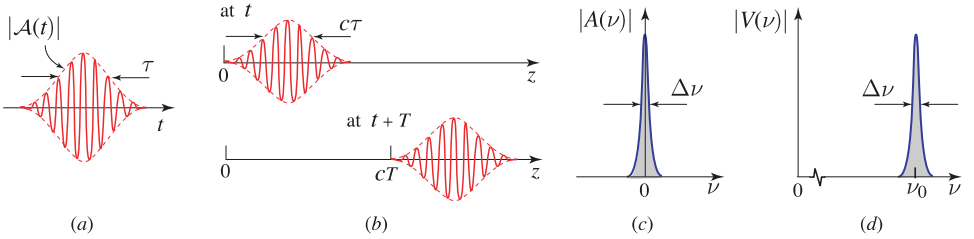


Figure 2.6-2 Temporal, spatial, and spectral characteristics of a pulsed plane wave. (*a*) The wavefunction at a fixed position has duration τ. (*b*) The wavefunction as a function of position at times t and $t + T$. The pulse travels with speed c and occupies a distance $c\tau$. (*c*) The magnitude $|A(\nu)|$ of the Fourier transform of the complex envelope. (*d*) The magnitude $|V(\nu)|$ of the Fourier transform of the complex wavefunction is centered at ν_0.

The propagation of a pulsed plane wave through a medium with frequency-dependent refractive index (i.e., with a frequency-dependent speed of light $c = c_o/n$) is discussed in Sec. 5.7 while other aspects of pulsed optics are considered in Chapter 23.

B. Light Beating

The dependence of the intensity of a polychromatic wave on time may be attributed to interference among the monochromatic components that constitute the wave. This concept is now demonstrated by means of two examples: interference between two monochromatic waves and interference among a finite number of monochromatic waves.

Interference of Two Monochromatic Waves of Different Frequencies

An optical wave composed of two monochromatic waves of frequencies ν_1 and ν_2 and intensities I_1 and I_2 has a complex wavefunction at some location in space

$$U(t) = \sqrt{I_1}\,\exp(j2\pi\nu_1 t) + \sqrt{I_2}\,\exp(j2\pi\nu_2 t)\,, \tag{2.6-11}$$

where the phases are taken to be zero and the $\mathbf{r}$ dependence has been suppressed for convenience. The intensity of the total wave is determined by use of the interference equation (2.5-4),

$$I(t) = I_1 + I_2 + 2\sqrt{I_1 I_2}\,\cos\left[2\pi(\nu_2 - \nu_1)t\right]\,. \tag{2.6-12}$$

The intensity therefore varies sinusoidally at the difference frequency $|\nu_2 - \nu_1|$, which is known as the **beat frequency**. This phenomenon goes by a number of names: **light beating**, **optical mixing**, **photomixing**, **optical heterodyning**, and **coherent detection**.

Equation (2.6-12) is analogous to (2.5-7), which describes the *spatial* interference of two waves of the same frequency traveling in different directions. This can be understood in terms of the phasor diagram in Fig. 2.5-1. The two phasors U_1 and U_2 rotate at angular frequencies $\omega_1 = 2\pi\nu_1$ and $\omega_2 = 2\pi\nu_2$, so that the difference angle is $\varphi = \varphi_2 - \varphi_1 = 2\pi(\nu_2 - \nu_1)t$, in accord with (2.6-12). Waves of different frequencies traveling in different directions exhibit spatiotemporal interference.

In electronics, beating or mixing is said to occur when the sum of two sinusoidal signals is detected by a nonlinear (e.g., quadratic) device called a mixer, producing signals at the difference and sum frequencies. This device is used in heterodyne radio receivers. In optics, photodetectors are responsive to the optical (Sec. 19.1B), or optical intensity which, in accordance with (2.6-8), is proportional to the absolute square of the

complex wavefunction. Optical detectors are therefore sensitive only to the difference frequency.

Much as (2.5-7) provides the basis for determining the direction of a wave via the spatial interference pattern at a screen, (2.6-12) provides a way of determining the frequency of an optical wave by measuring the temporal interference pattern at the output of a photodetector. The use of optical beating in optical heterodyne receivers is discussed in Sec. 25.4. Other forms of optical mixing make use of nonlinear media to generate optical-frequency differences and sums, as described in Chapter 22.

EXERCISE 2.6-1

Optical Doppler Radar. As a result of the **Doppler effect**, a monochromatic optical wave of frequency ν, reflected from an object moving with a velocity component v along the line of sight from an observer, undergoes a frequency shift $\Delta\nu = \pm(2v/c)\,\nu$, depending on whether the object is moving toward (+) or away (−) from the observer. Assuming that the original and reflected waves are superimposed, derive an expression for the intensity of the resultant wave. Suggest a method for measuring the velocity of a target using such an arrangement. If one of the mirrors of a Michelson interferometer [Fig. 2.5-3(*b*)] moves with velocity $\pm v$, use (2.5-6) to show that the beat frequency is $\pm(2v/c)\,\nu$.

Interference of M Monochromatic Waves with Equal Intensities and Equally Spaced Frequencies

The interference of a large number of monochromatic waves with equal intensities, equal phases, and equally spaced frequencies can result in the generation of brief pulses of light. Consider an odd number of waves, $M = 2L + 1$, each with intensity I_0 and zero phase, and with frequencies

$$\nu_q = \nu_0 + q\nu_F, \qquad q = -L, \ldots, 0, \ldots L\,, \tag{2.6-13}$$

centered about frequency ν_0 and spaced by frequency $\nu_F \ll \nu_0$. At a given position, the total wave has a complex wavefunction

$$U(t) = \sqrt{I_0} \sum_{q=-L}^{L} \exp\left[j2\pi(\nu_0 + q\nu_F)t\right]\,. \tag{2.6-14}$$

This represents the sum of M phasors of equal magnitudes and successive phases that differ by $\varphi = 2\pi\nu_F t$. Results for the intensity are immediately available from the analysis carried out in Sec. 2.5B, which is mathematically identical to the case at hand. Referring to (2.5-12) and Fig. 2.5-7, and using the substitution $\varphi = 2\pi t/T_F$ with $T_F = 1/\nu_F$, the total intensity is

$$I(t) = |U(t)|^2 = I_0 \frac{\sin^2(M\pi t/T_F)}{\sin^2(\pi t/T_F)}\,. \tag{2.6-15}$$

As illustrated in Fig. 2.6-3, the intensity $I(t)$ is a periodic sequence of optical pulses with period T_F, peak intensity $M^2 I_0$, and mean intensity $\bar{I} = MI_0$. The peak intensity is therefore M times greater than the mean intensity. The duration of each pulse is approximately T_F/M so that the pulses become very short when M is large. If $\nu_F = 1$ GHz, for example, then $T_F = 1$ ns; for $M = 1000$, pulses of 1-ps duration are generated.

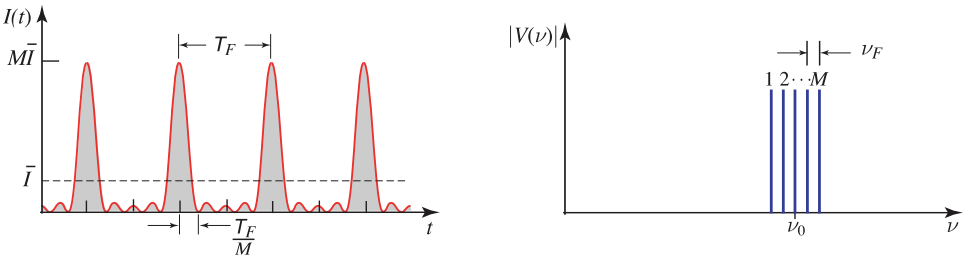


Figure 2.6-3 Time dependence of the optical intensity $I(t)$ of a polychromatic wave comprising M monochromatic waves of equal intensities, equal phases, and successive frequencies that differ by ν_F. The intensity $I(t)$ is a periodic train of pulses of period $T_F = 1/\nu_F$ with a peak that is M times greater than the mean $\bar{I}$. The duration of each pulse is M times shorter than the period. In this example $M = 5$. These graphs should be compared with those in Fig. 2.5-7. The magnitude of the Fourier transform $|V(\nu)|$ is shown in the lower graph.

This example provides a dramatic demonstration of how M monochromatic waves can conspire to produce a train of very short optical pulses. We shall see in Sec. 16.4D that the modes of a laser can be *mode-locked* in the fashion described above to produce a sequence of ultrashort laser pulses.

READING LIST

Wave Optics and Interferometry

See also the reading list on general optics in Chapter 1.

D. Fleisch and L. Kinnaman, *A Student's Guide to Waves*, Cambridge University Press, 2015.

M. Mansuripur, *Classical Optics and Its Applications*, Cambridge University Press, 2nd ed. 2009.

P. Hariharan, *Basics of Interferometry*, Academic Press, 2nd ed. 2006.

J. R. Pierce, *Almost All About Waves*, MIT Press, 1974; Dover, reissued 2006.

H. J. Pain, *The Physics of Vibrations and Waves*, Wiley, 6th ed. 2005.

R. H. Webb, *Elementary Wave Optics*, Academic Press, 1969; Dover, reissued 2005.

E. Hecht and A. Zajac, *Optics*, Addison–Wesley, 2nd ed. 1990.

J. M. Vaughan, *The Fabry–Perot Interferometer*, CRC Press, 1989.

H. D. Young, *Fundamentals of Waves, Optics, and Modern Physics*, McGraw–Hill, paperback 2nd ed. 1976.

M. Françon, N. Krauzman, J. P. Matieu, and M. May, *Experiments in Physical Optics*, CRC Press, 1970.

M. Françon, *Optical Interferometry*, Academic Press, 1966.

Spectroscopy

D. L. Pavia, G. M. Lampman, G. S. Kriz and J. A. Vyvyan, *Introduction to Spectroscopy*, Brooks/Cole, 5th ed. 2014.

B. C. Smith, *Fundamentals of Fourier Transform Infrared Spectroscopy*, CRC Press/Taylor & Francis, 2nd ed. 2011.

J. M. Hollas, *Modern Spectroscopy*, Wiley, paperback 4th ed. 2010.

P. R. Griffiths and J. A. de Haseth, *Fourier Transform Infrared Spectrometry*, Wiley, 2nd ed. 2007.

Diffraction Gratings

C. Palmer, *Diffraction Grating Handbook*, Richardson Gratings (Rochester, NY), 7th ed. 2014.

E. G. Loewen and E. Popov, *Diffraction Gratings and Applications*, CRC Press, 1997.

Interferometry for Gravitational-Wave Detection

S. Wills, Gravitational Waves: The Road Ahead, *Optics & Photonics News*, vol. 29, no. 5, pp. 44–51, 2018.

B. P. Abbott *et al*. (LIGO Scientific Collaboration and Virgo Collaboration), GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, *Physical Review Letters*, vol. 119, 161101, 2017.

B. P. Abbott *et al*. (LIGO Scientific Collaboration and Virgo Collaboration), Observation of Gravitational Waves from a Binary Black Hole Merger, *Physical Review Letters*, vol. 116, 061102, 2016.

B. P. Abbott *et al*., Astrophysical Implications of the Binary Black Hole Merger GW150914, *The Astrophysical Journal Letters*, vol. 818, L22, 2016.

R. W. P. Drever, Fabry–Perot Cavity Gravity-Wave Detectors, in D. G. Blair, ed., *The Detection of Gravitational Waves*, Cambridge University Press, 1991, Chapter 12, pp. 306–328.

A. Brillet, J. Gea-Banacloche, G. Leuchs, C. N. Man, and J. Y. Vinet, Advanced Techniques: Recycling and Squeezing, in D. G. Blair, ed., *The Detection of Gravitational Waves*, Cambridge University Press, 1991, Chapter 15, pp. 369–405.

D. G. Blair, Gravitational Waves in General Relativity, in D. G. Blair, ed., *The Detection of Gravitational Waves*, Cambridge University Press, 1991, Chapter 1, pp. 3–15.

A. Einstein, Die Feldgleichungen der Gravitation (The Field Equations of Gravitation), Sitzungsberichte der Königlich Preussische Akademie der Wissenschaften (Berlin), pp. 844–847 (part 2), 1915.

Popular and Historical

P. Daukantas, 200 Years of Fresnel's Legacy, *Optics & Photonics News*, vol. 26, no. 9, pp. 40–47, 2015.

T. Levitt, *A Short Bright Flash: Augustin Fresnel and the Birth of the Modern Lighthouse*, Norton, 2013.

F. J. Dijksterhuis, *Lenses and Waves: Christiaan Huygens and the Mathematical Science of Optics in the Seventeenth Century*, 2004, Springer-Verlag, paperback ed. 2011.

J. Z. Buchwald, *The Rise of the Wave Theory of Light: Optical Theory and Experiment in the Early Nineteenth Century*, University of Chicago Press, paperback ed. 1989.

W. E. Kock, *Sound Waves and Light Waves: The Fundamentals of Wave Motion*, Doubleday/Anchor, 1965.

C. Huygens, *Treatise on Light*, 1690, University of Chicago Press, 1945; Echo Library, reprinted 2007.

Seminal Articles

G. W. Kamerman, ed., *Selected Papers on Laser Radar*, SPIE Optical Engineering Press (Milestone Series Volume 133), 1997.

P. Hariharan and D. Malacara-Hernandez, eds., *Selected Papers on Interference, Interferometry, and Interferometric Metrology*, SPIE Optical Engineering Press (Milestone Series Volume 110), 1995.

D. Maystre, ed., *Selected Papers on Diffraction Gratings*, SPIE Optical Engineering Press (Milestone Series Volume 83), 1993.

P. Hariharan, ed., *Selected Papers on Interferometry*, SPIE Optical Engineering Press (Milestone Series Volume 28), 1991.

PROBLEMS

2.2-3 **Spherical Waves.** Use a spherical coordinate system to verify that the complex amplitude of the spherical wave (2.2-15) satisfies the Helmholtz equation (2.2-7).

2.2-4 **Intensity of a Spherical Wave.** Derive an expression for the intensity I of a spherical wave at a distance r from its center in terms of the optical power P. What is the intensity at $r = 1\,\mathrm{m}$ for $P = 100\,\mathrm{W}$?

2.2-5 **Cylindrical Waves.** Derive expressions for the complex amplitude and intensity of a monochromatic wave whose wavefronts are cylinders centered about the y axis.

2.2-6 **Paraxial Helmholtz Equation.** Derive the paraxial Helmholtz equation (2.2-23) using the approximations in (2.2-21) and (2.2-22).

2.2-7 **Conjugate Waves.** Compare a monochromatic wave with complex amplitude $U(\mathbf{r})$ to a monochromatic wave of the same frequency but with complex amplitude $U^*(\mathbf{r})$, with respect to intensity, wavefronts, and wavefront normals. Use the plane wave $U(\mathbf{r}) = A\exp[-jk(x+y)/\sqrt{2}]$ and the spherical wave $U(\mathbf{r}) = (A/r)\exp(-jkr)$ as examples.

2.3-1 **Wave in a GRIN Slab.** Sketch the wavefronts of a wave traveling in the graded-index SELFOC slab described in Example 1.3-1.

2.4-7 **Reflection of a Spherical Wave from a Planar Mirror.** A spherical wave is reflected from a planar mirror sufficiently far from the wave origin so that the Fresnel approximation is satisfied. By regarding the spherical wave locally as a plane wave with slowly varying direction, use the law of reflection of plane waves to determine the nature of the reflected wave.

2.4-8 **Optical Pathlength.** A plane wave travels in a direction normal to a thin plate made of N thin parallel layers of thicknesses d_q and refractive indices n_q, $q = 1, 2, \ldots, N$. If all reflections are ignored, determine the complex amplitude transmittance of the plate. If the plate is replaced with a distance d of free space, what should d be so that the same complex amplitude transmittance is obtained? Show that this distance is the optical pathlength defined in Sec. 1.1.

2.4-9 **Diffraction Grating.** Repeat Exercise 2.4-5 for a thin transparent plate whose thickness $d(x, y)$ is a square (instead of sinusoidal) periodic function of x of period $\Lambda \gg \lambda$. Show that the angle θ between the diffracted waves is still given by $\theta \approx \lambda/\Lambda$. If a plane wave is incident in a direction normal to the grating, determine the amplitudes of the different diffracted plane waves.

2.4-10 **Reflectance of a Spherical Mirror.** Show that the complex amplitude reflectance $\mathrm{r}(x, y)$ (the ratio of the complex amplitudes of the reflected and incident waves) of a thin spherical mirror of radius R is given by $\mathrm{r}(x, y) = h_0 \exp[-jk_o(x^2 + y^2)/R]$, where h_0 is a constant. Compare this to the complex amplitude transmittance of a lens of focal length $f = -R/2$.

2.5-4 **Standing Waves.** Derive an expression for the intensity I of the superposition of two plane waves of wavelength λ traveling in opposite directions along the z axis. Sketch I versus z.

2.5-5 **Fringe Visibility.** The visibility of an interference pattern such as that described by (2.5-4) and plotted in Fig. 2.5-1 is defined as the ratio $\mathcal{V} = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$, where $I_{\max}$ and $I_{\min}$ are the maximum and minimum values of I. Derive an expression for $\mathcal{V}$ as a function of the ratio I_1/I_2 of the two interfering waves and determine the ratio I_1/I_2 for which the visibility is maximum.

2.5-6 **Michelson Interferometer.** If one of the mirrors of the Michelson interferometer [Fig. 2.5-3(*b*)] is misaligned by a small angle $\Delta\theta$, describe the shape of the interference pattern in the detector plane. What happens to this pattern as the other mirror moves?

2.6-2 **Pulsed Spherical Wave.**

(a) Show that a pulsed spherical wave has a complex wavefunction of the form $U(\mathbf{r}, t) = (1/r)\mathfrak{a}(t - r/c)$, where $\mathfrak{a}(t)$ is an arbitrary function.

(b) An ultrashort optical pulse has a complex wavefunction with central frequency corresponding to a wavelength $\lambda_o = 585$ nm and a Gaussian envelope of RMS width of $\sigma_t = 6$ fs (1 fs$= 10^{-15}$ s). How many optical cycles are contained within the pulse width? If the pulse propagates in free space as a spherical wave initiated at the origin at $t = 0$, describe the spatial distribution of the intensity as a function of the radial distance at time $t = 1$ ps.

CHAPTER

3

BEAM OPTICS

3.1 THE GAUSSIAN BEAM 80
- A. Complex Amplitude
- B. Properties
- C. Beam Quality

3.2 TRANSMISSION THROUGH OPTICAL COMPONENTS 91
- A. Transmission Through a Thin Lens
- B. Beam Shaping
- C. Reflection from a Spherical Mirror
- *D. Transmission Through an Arbitrary Optical System

3.3 HERMITE–GAUSSIAN BEAMS 99

3.4 LAGUERRE–GAUSSIAN BEAMS 102

3.5 NONDIFFRACTING BEAMS 105
- A. Bessel Beams
- *B. Airy Beams

The Gaussian beam, named after the German mathematician **Carl Friedrich Gauss (1777–1855)**, is circularly symmetric and has a radial intensity that follows the form of a Gaussian distribution.

Edmond Nicolas Laguerre (1834–1886), a French mathematician, devised a set of polynomials useful for describing circularly symmetric light beams with helical wavefronts and orbital angular momentum.

Friedrich Wilhelm Bessel (1784–1846), a noted German astronomer, established a set of functions that characterize the radial intensity of circularly symmetric, planar-wavefront, non-diffracting optical beams.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Can light be spatially confined and transported in free space without angular spread? Although the wave nature of light precludes the possibility of such idealized transport, light can, in fact, be confined in the form of beams that come as close as possible to waves that are spatially localized and nondiverging.

The two extremes of angular and spatial confinement are the plane wave and the spherical wave, respectively. The wavefront normals (rays) of a plane wave coincide with the direction of travel of the wave so that there is no angular spread, but its energy extends spatially over all space. The spherical wave, in contrast, originates from a single spatial point, but its wavefront normals (rays) diverge in all angular directions.

Waves whose wavefront normals make small angles with the z axis are called paraxial waves. They must satisfy the paraxial Helmholtz equation, which was derived in Sec. 2.2C. The **Gaussian beam** is an important solution of this equation that exhibits the characteristics of an optical beam, as attested to by a number of its properties. The beam power is principally concentrated within a small cylinder that surrounds the beam axis. The intensity distribution in any transverse plane is a circularly symmetric Gaussian function centered about the beam axis. The width of this function is minimum at the beam waist and gradually becomes larger as the distance from the waist increases in both directions. The wavefronts are approximately planar near the beam waist, then gradually curve as the distance from the waist increases, and ultimately become approximately spherical far from the beam waist. The angular divergence of the wavefront normals assumes the minimum value permitted by the wave equation for a given beam width. The wavefront normals are therefore much like a thin pencil of rays. Under ideal conditions, the light from many types of lasers takes the form of a Gaussian beam.

This Chapter

An expression for the complex amplitude of the Gaussian beam is set forth in Sec. 3.1 and a detailed discussion of its physical properties (intensity, power, beam width, beam divergence, depth of focus, and phase) is provided. The shaping of Gaussian beams (focusing, relaying, collimating, and expanding) via the use of various optical components is the subject of Sec. 3.2. In Secs. 3.3 and 3.4 we introduce more general families of optical beams, known as Hermite–Gaussian and Laguerre–Gaussian beams, respectively, of which the simple Gaussian beam is a member. Finally, in Sec. 3.5 we discuss nondiffracting beams, including Bessel, Bessel–Gaussian, and Airy beams.

3.1 THE GAUSSIAN BEAM

A. Complex Amplitude

The concept of paraxial waves was introduced in Sec. 2.2C. A monochromatic paraxial wave is a plane wave traveling along the z direction e^{-jkz} (with wavenumber $k = 2\pi/\lambda$ and wavelength λ), modulated by a complex envelope $A(\mathbf{r})$ that is a slowly varying function of position (see Fig. 2.2-5), so that its complex amplitude is

$$U(\mathbf{r}) = A(\mathbf{r}) \exp(-jkz). \tag{3.1-1}$$

The envelope is taken to be approximately constant within a neighborhood of size λ, so that the wave locally maintains its plane-wave nature but exhibits wavefront normals that are paraxial rays.

In order that the complex amplitude $U(\mathbf{r})$ satisfy the Helmholtz equation, $\nabla^2 U + k^2 U = 0$, the complex envelope $A(\mathbf{r})$ must satisfy the paraxial Helmholtz equation (2.2-23)

$$\nabla_T^2 A - j\,2k\frac{\partial A}{\partial z} = 0, \tag{3.1-2}$$

where $\nabla_T^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the transverse Laplacian operator. A simple solution to the paraxial Helmholtz equation yields the paraboloidal wave (see Exercise 2.2-2), for which

$$A(\mathbf{r}) = \frac{A_1}{z}\exp\left(-jk\frac{\rho^2}{2z}\right), \qquad \rho^2 = x^2 + y^2, \tag{3.1-3}$$

where A_1 is a constant. The paraboloidal wave is the paraxial approximation of the spherical wave $U(r) = (A_1/r)\exp(-jkr)$ when x and y are much smaller than z (see Sec. 2.2B).

Another solution of the paraxial Helmholtz equation leads to the Gaussian beam. It is obtained from the paraboloidal wave by use of a simple transformation. Since the complex envelope of the paraboloidal wave (3.1-3) is a solution of the paraxial Helmholtz equation (3.1-2), so too is a shifted version of it, with $z - \xi$ replacing z where ξ is a constant:

$$A(\mathbf{r}) = \frac{A_1}{q(z)}\exp\left[-jk\,\frac{\rho^2}{2q(z)}\right], \qquad q(z) = z - \xi. \tag{3.1-4}$$

This represents a paraboloidal wave centered about the point $z = \xi$ instead of about $z = 0$. Equation (3.1-4) remains a solution of (3.1-2) even when ξ is complex, but the solution acquires dramatically different properties. In particular, when ξ is purely imaginary, say $\xi = -jz_0$ where z_0 is real, (3.1-4) yields the complex envelope of the Gaussian beam

$$\boxed{A(\mathbf{r}) = \frac{A_1}{q(z)}\exp\left[-jk\frac{\rho^2}{2q(z)}\right], \qquad q(z) = z + jz_0.} \tag{3.1-5}$$

Complex Envelope

The quantity $q(z)$ is called the **q-parameter** of the beam and the parameter z_0 is known as the **Rayleigh range**.

To separate the amplitude and phase of this complex envelope, we write the complex function $1/q(z) = 1/(z + jz_0)$ in terms of its real and imaginary parts by defining two new real functions, $R(z)$ and $W(z)$, such that

$$\boxed{\frac{1}{q(z)} = \frac{1}{R(z)} - j\,\frac{\lambda}{\pi W^2(z)}.} \tag{3.1-6}$$

It will be shown subsequently that $W(z)$ and $R(z)$ are measures of the beam width and wavefront radius of curvature, respectively. Expressions for $W(z)$ and $R(z)$ as functions of z and z_0 are provided in (3.1-8) and (3.1-9). Substituting (3.1-6) into (3.1-5) and using (3.1-1) leads directly to an expression for the complex amplitude $U(\mathbf{r})$ of

the Gaussian beam:

$$U(\mathbf{r}) = A_0 \frac{W_0}{W(z)} \exp\left[-\frac{\rho^2}{W^2(z)}\right] \exp\left[-jkz - jk\frac{\rho^2}{2R(z)} + j\zeta(z)\right] \tag{3.1-7}$$

Complex Amplitude

$$W(z) = W_0 \sqrt{1 + \left(\frac{z}{z_0}\right)^2} \tag{3.1-8}$$

$$R(z) = z\left[1 + \left(\frac{z_0}{z}\right)^2\right] \tag{3.1-9}$$

$$\zeta(z) = \tan^{-1}\frac{z}{z_0} \tag{3.1-10}$$

$$W_0 = \sqrt{\frac{\lambda z_0}{\pi}}\,. \tag{3.1-11}$$

Beam Parameters

A new constant $A_0 = A_1/jz_0$ has been defined for convenience.

The expression for the complex amplitude of the Gaussian beam provided above is central to this chapter. It is described by two independent parameters, A_0 and z_0, which are determined from the boundary conditions. All other parameters are related to the z_0 and the wavelength λ by (3.1-8) to (3.1-11). The significance of these parameters will become clear in the sequel.

B. Properties

Equations (3.1-7)–(3.1-11) will now be used to determine the properties of the Gaussian beam.

Intensity

The optical intensity $I(\mathbf{r}) = |U(\mathbf{r})|^2$ is a function of the axial and radial positions, z and $\rho = \sqrt{x^2 + y^2}$, respectively

$$I(\rho, z) = I_0 \left[\frac{W_0}{W(z)}\right]^2 \exp\left[-\frac{2\rho^2}{W^2(z)}\right], \tag{3.1-12}$$

where $I_0 = |A_0|^2$. At any value of z the intensity is a Gaussian function of the radial distance ρ — hence the appellation "Gaussian beam." The Gaussian function has its peak on the z axis, at $\rho = 0$, and decreases monotonically as ρ increases. The beam width $W(z)$ of the Gaussian distribution increases with the axial distance z as illustrated in Fig. 3.1-1.

On the beam axis ($\rho = 0$) the intensity in (3.1-12) reduces to

$$I(0, z) = I_0 \left[\frac{W_0}{W(z)}\right]^2 = \frac{I_0}{1 + (z/z_0)^2}\,, \tag{3.1-13}$$

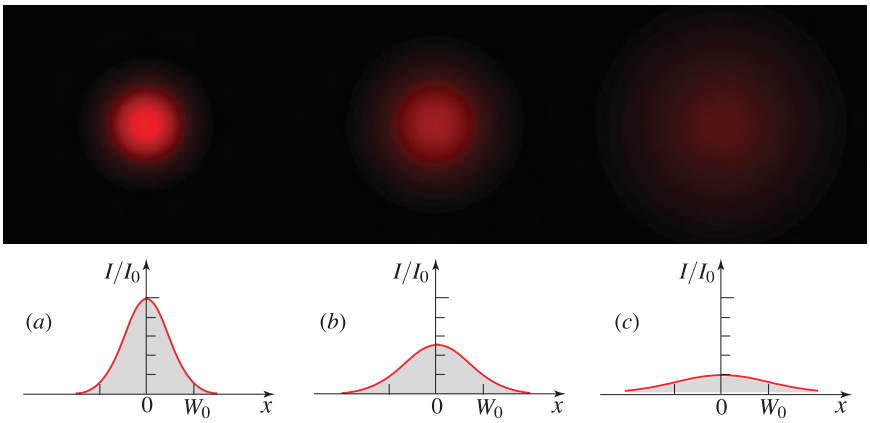


Figure 3.1-1 Normalized Gaussian beam intensity I/I_0 as a function of the radial distance ρ at different axial distances: (a) $z = 0$; (b) $z = z_0$; (c) $z = 2z_0$.

which has its maximum value I_0 at $z = 0$ and decays gradually with increasing z, reaching half its peak value at $z = \pm z_0$ (Fig. 3.1-2). When $|z| \gg z_0$, $I(0, z) \approx I_0 z_0^2/z^2$, so that the intensity decreases with distance in accordance with an inverse-square law, as for spherical and paraboloidal waves. Overall, the beam center ($z = 0, \rho = 0$) is the location of the greatest intensity: $I(0, 0) = I_0$.

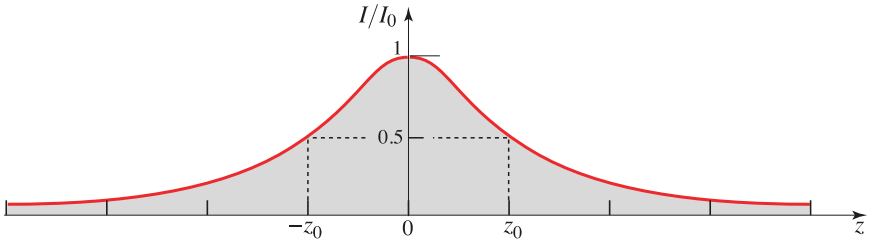


Figure 3.1-2 The normalized beam intensity I/I_0 at points on the beam axis ($\rho = 0$) as a function of distance along the beam axis, z.

Power

The total optical power carried by the beam is the integral of the optical intensity over any transverse plane (say at position z),

$$P = \int_0^\infty I(\rho, z)\, 2\pi\rho\, d\rho, \tag{3.1-14}$$

which yields

$$P = \tfrac{1}{2} I_0 \left(\pi W_0^2\right). \tag{3.1-15}$$

The beam power is thus half the peak intensity multiplied by the beam area. The result is independent of z, as expected. Since optical beams are often described by their power

P, it is useful to express I_0 in terms of P via (3.1-15), whereupon (3.1-12) can be rewritten in the form

$$I(\rho, z) = \frac{2P}{\pi W^2(z)} \exp\left[-\frac{2\rho^2}{W^2(z)}\right]. \tag{3.1-16}$$

Beam Intensity

The ratio of the power carried within a circle of radius ρ_0 in the transverse plane to the total power, at position z, is

$$\frac{1}{P}\int_0^{\rho_0} I(\rho, z)\, 2\pi\rho\, d\rho = 1 - \exp\left[-\frac{2\rho_0^2}{W^2(z)}\right]. \tag{3.1-17}$$

The power contained within a circle of radius $\rho_0 = W(z)$ is therefore approximately 86% of the total power. About 99% of the power is contained within a circle of radius $1.5\,W(z)$.

Beam Width

At any transverse plane, the beam intensity assumes its peak value on the beam axis, and decreases by the factor $1/e^2 \approx 0.135$ at the radial distance $\rho = W(z)$. Since 86% of the power is carried within a circle of radius $W(z)$, we regard $W(z)$ as the beam radius (or beam width). The RMS width of the intensity distribution, on the other hand, is $\sigma = \frac{1}{2}W(z)$ (see Appendix A, Sec. A.2, for the different definitions of width).

The dependence of the beam width on z is governed by (3.1-8),

$$W(z) = W_0\sqrt{1 + \left(\frac{z}{z_0}\right)^2}. \tag{3.1-18}$$

Beam Width (Beam Radius)

It assumes its minimum value, W_0, at the plane $z = 0$. This is the beam waist and W_0 is thus known as the **waist radius**. The waist diameter $2W_0$ is also called the **spot size**. The beam width increases monotonically with z, and assumes the value $\sqrt{2}W_0$ at $z = \pm z_0$ (Fig. 3.1-3).

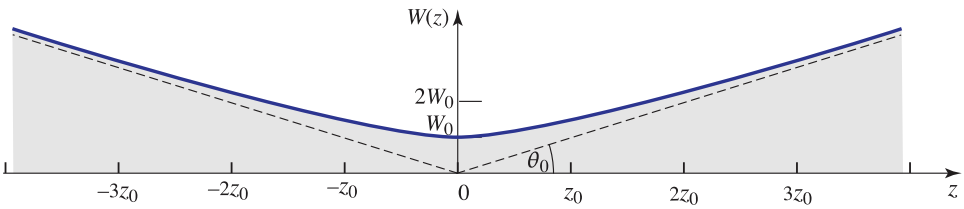


Figure 3.1-3 The beam width $W(z)$ assumes its minimum value W_0 at the beam waist ($z = 0$), reaches $\sqrt{2}W_0$ at $z = \pm z_0$, and increases linearly with z for large z.

Beam Divergence

For $z \gg z_0$ the first term of (3.1-18) may be neglected, which results in the linear relation

$$W(z) \approx \frac{W_0}{z_0} z = \theta_0 z. \tag{3.1-19}$$

As illustrated in Fig. 3.1-3, the beam then diverges as a cone of half-angle

$$\theta_0 = \frac{W_0}{z_0} = \frac{\lambda}{\pi W_0}, \tag{3.1-20}$$

where we have made use of (3.1-11). Approximately 86% of the beam power is confined within this cone, as indicated following (3.1-17).

Rewriting (3.1-20) in terms of the spot size, the angular divergence of the beam becomes

$$2\theta_0 = \frac{4}{\pi} \frac{\lambda}{2W_0}. \tag{3.1-21}$$

Divergence Angle

The divergence angle is directly proportional to the wavelength λ and inversely proportional to the spot size $2W_0$. Squeezing the spot size (beam-waist diameter) therefore leads to increased beam divergence. It is clear that a highly directional beam is constructed by making use of a short wavelength and a thick beam waist.

Depth of Focus

Since the beam has its minimum width at $z = 0$, as shown in Fig. 3.1-3, it achieves its best focus at the plane $z = 0$. In either direction, the beam gradually grows "out of focus." The axial distance within which the beam width is no greater than a factor $\sqrt{2}$ times its minimum value, so that its area is within a factor of 2 of the minimum, is known as the **depth-of-focus** or **confocal parameter** (Fig. 3.1-4). It is evident from (3.1-18) and (3.1-11) that the actual depth of focus is twice the Rayleigh range:

$$2z_0 = \frac{2\pi W_0^2}{\lambda}. \tag{3.1-22}$$

Depth of Focus

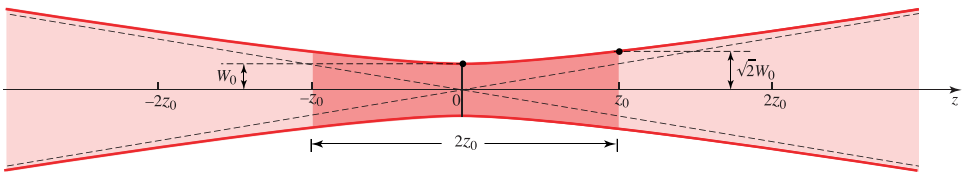


Figure 3.1-4 Depth of focus of a Gaussian beam.

The depth of focus is therefore directly proportional to the area of the beam at its waist, πW_0^2, and inversely proportional to the wavelength, λ. A beam focused to a

small spot size thus has a short depth of focus; locating the plane of focus thus requires increased accuracy. Small spot size and long depth of focus can be simultaneously attained only for short wavelengths. As an example, at $\lambda_o = 633$ nm (a common He–Ne laser-line wavelength), a spot size $2W_0 = 2$ cm corresponds to a depth of focus $2z_0 \approx 1$ km. A much smaller spot size of 20 μm corresponds to a much shorter depth of focus of 1 mm.

Phase

The phase of the Gaussian beam is, from (3.1-7),[†]

$$\varphi(\rho, z) = kz - \zeta(z) + \frac{k\rho^2}{2R(z)}. \tag{3.1-23}$$

On the beam axis ($\rho = 0$) the phase comprises two components:

$$\varphi(0, z) = kz - \zeta(z). \tag{3.1-24}$$

The first, kz, is the phase of a plane wave. The second represents a phase retardation $\zeta(z)$ given by (3.1-10), which ranges from $-\pi/2$ at $z = -\infty$ to $+\pi/2$ at $z = \infty$, as illustrated in Fig. 3.1-5. This phase retardation corresponds to an excess delay of the wavefront in relation to a plane wave (see also Fig. 3.1-8). The total accumulated excess retardation as the wave travels from $z = -\infty$ to $z = \infty$ is π. This phenomenon is known as the **Gouy effect**. It arises from the transverse spatial confinement of the beam, which is accompanied by a spread in its transverse wavevector components by virtue of the Fourier transform. This results in a reduction in the axial component of the wavevector k_z from its plane-wave value $k_z = \sqrt{k^2 - k_x^2 - k_y^2}$ (see Sec. 2.2B).[‡]

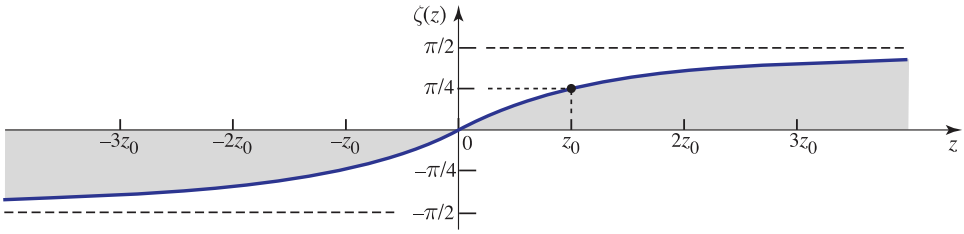


Figure 3.1-5 The function $\zeta(z)$ represents the phase retardation of the Gaussian beam relative to a uniform plane wave at points on the beam axis.

Wavefronts

The third component in (3.1-23) is responsible for wavefront bending. It represents the deviation of the phase at off-axis points in a given transverse plane from that at the axial point. The surfaces of constant phase satisfy $k[z + \rho^2/2R(z)] - \zeta(z) = 2\pi q$. Since $\zeta(z)$ and $R(z)$ are relatively slowly varying functions, they are effectively constant at points within the beam width on each wavefront. We may therefore write $z + \rho^2/2R \approx q\lambda + \zeta\lambda/2\pi$, where $R = R(z)$ and $\zeta = \zeta(z)$. This is the equation of a paraboloidal surface with **radius of curvature** R. Thus, $R(z)$, plotted in Fig. 3.1-6, is the radius of curvature of the wavefront at position z along the beam axis.

[†] The phase $\varphi(\rho, z)$ in (3.1-23), and throughout this chapter, is related to the phase factor specified in (3.1-7) by $\exp(-j\varphi)$.

[‡] See S. Feng and H. G. Winful, Physical Origin of the Gouy Phase Shift, *Optics Letters*, vol. 26, pp. 485–487, 2001.

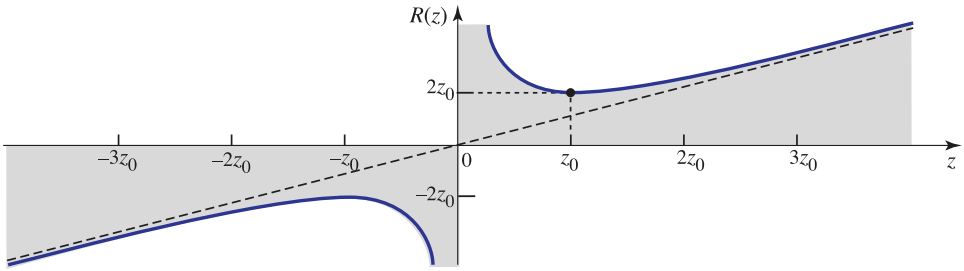


Figure 3.1-6 The radius of curvature $R(z)$ of the wavefronts of a Gaussian beam as a function of position along the beam axis. The dashed line is the radius of curvature of a spherical wave.

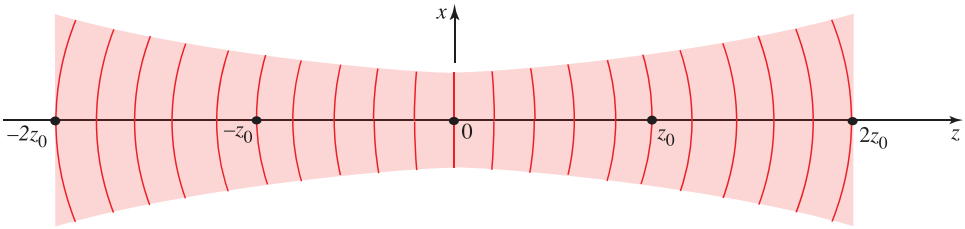


Figure 3.1-7 Wavefronts of a Gaussian beam.

As illustrated in Fig. 3.1-6, the radius of curvature $R(z)$ is infinite at $z = 0$, so that the wavefronts are planar, i.e., they have no curvature. The radius decreases to a minimum value of $2z_0$ at $z = z_0$, where the wavefront has the greatest curvature (Fig. 3.1-7). The radius of curvature subsequently increases as z increases further until $R(z) \approx z$ for $z \gg z_0$. The wavefronts are then approximately the same as those of a spherical wave. The pattern of the wavefronts is identical for negative z, except for a change in sign (Fig. 3.1-8). We have adopted the convention that a diverging wavefront has a positive radius of curvature whereas a converging wavefront has a negative radius of curvature.

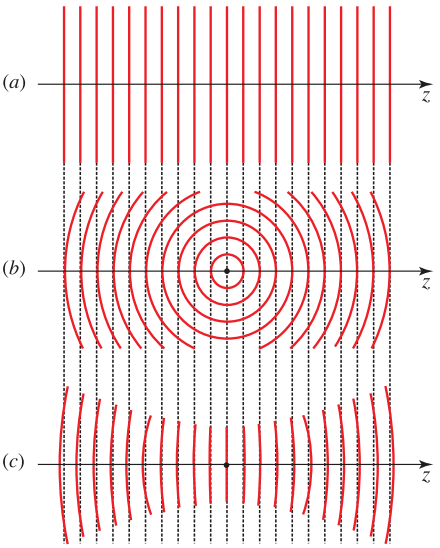


Figure 3.1-8 Wavefronts of (a) a uniform plane wave; (b) a spherical wave; (c) a Gaussian beam. At points near the beam center, the Gaussian beam resembles a plane wave. At large z the beam behaves like a spherical wave except that its phase is retarded by $\pi/2$ (a quarter of the distance between two adjacent wavefronts).

Parameters Required to Characterize a Gaussian Beam

Assuming that the wavelength λ is known, how many parameters are required to describe a plane wave, a spherical wave, and a Gaussian beam? The plane wave is completely specified by its complex amplitude and direction. The spherical wave is specified by its complex amplitude and the location of its origin. The Gaussian beam, in contrast, requires more parameters for its characterization — its peak amplitude [determined by A_0 in (3.1-7)], its direction (the beam axis), the location of its waist, *and* one additional parameter, such as the waist radius W_0 *or* the Rayleigh range z_0. Thus, if the beam peak amplitude and the axis are known, two additional parameters are required for full specification.

If the complex q-parameter, $q(z) = z + jz_0$, is known, the distance to the beam waist z and the Rayleigh range z_0 are readily identified as the real and imaginary parts thereof. As an example, if $q(z)$ is $3 + j4$ cm at some point on the beam axis, we infer that the beam waist lies at a distance $z = 3$cm to the left of that point and that the depth of focus is $2z_0 = 8$ cm. The waist radius W_0 may then be determined via (3.1-11). The quantity $q(z)$ is therefore sufficient for characterizing a Gaussian beam of known peak amplitude and beam axis. Given $q(z)$ at a single point, the linear dependence of q on z permits it to be determined at all points: if $q(z) = q_1$ and $q(z + d) = q_2$, then $q_2 = q_1 + d$. Using the example provided immediately above, at $z = 13$ cm it is evident that $q = 13 + j4$.

If the beam width $W(z)$ and the radius of curvature $R(z)$ are known at an arbitrary point on the beam axis, the beam can be fully identified by solving (3.1-8), (3.1-9), and (3.1-11) for z, z_0, and W_0. Alternatively, the beam can be identified by determining $q(z)$ from $W(z)$ and $R(z)$ using (3.1-6).

Summary: Properties of the Gaussian Beam at Special Locations

- *At the location $z = z_0$.* At an axial distance z_0 from the beam waist, the wave has the following properties:
  - The intensity on the beam axis is $\frac{1}{2}$ the peak intensity.
  - The beam width is a factor of $\sqrt{2}$ greater than the width at the beam waist, and the beam area is larger by a factor of 2.
  - The phase on the beam axis is retarded by an angle $\pi/4$ relative to the phase of a plane wave.
  - The radius of curvature of the wavefront achieves its minimum value, $R = 2z_0$, so that the wavefront has the greatest curvature.
- *Near the beam center.* At locations for which $|z| \ll z_0$ and $\rho \ll W_0$, the quantity $\exp[-\rho^2/W^2(z)] \approx \exp(-\rho^2/W_0^2) \approx 1$, so that the beam intensity, which is proportional to the square of this quantity, is approximately constant. Also, $R(z) \approx z_0^2/z$ and $\zeta(z) \approx 0$, so that the phase $k[z + \rho^2/2R(z)] \approx kz(1 + \rho^2/2z_0^2) \approx kz$, by virtue of (3.1-11) when $z_0 \gg \lambda$. The Gaussian beam may therefore be approximated near its center by a plane wave.
- *Far from the beam waist.* At transverse locations within the waist radius ($\rho < W_0$), but far from the beam waist ($z \gg z_0$), the wave behaves approximately like a spherical wave. In this domain $W(z) \approx W_0 z/z_0 \gg W_0$ and $\rho < W_0$, so that $\exp[-\rho^2/W^2(z)] \approx 1$ and the beam intensity is approximately uniform. Since $R(z) \approx z$ in this regime, the wavefronts are approximately spherical. Thus, except for the Gouy phase retardation $\zeta(z) \approx \pi/2$, the complex amplitude of the Gaussian beam approaches that of the paraboloidal wave, which in turn approaches that of the spherical wave in the paraxial approximation.

EXERCISE 3.1-1

Parameters of a Gaussian Laser Beam. A 1-mW He–Ne laser produces a Gaussian beam at a wavelength of $\lambda = 633$ nm with a spot size $2W_0 = 0.1$ mm.

(a) Determine the angular divergence of the beam, its depth of focus, and its diameter at $z = 3.5 \times 10^5$ km (approximately the distance to the moon).
(b) What is the radius of curvature of the wavefront at $z = 0$, $z = z_0$, and $z = 2z_0$?
(c) What is the optical intensity (in W/cm^2) at the beam center ($z = 0$, $\rho = 0$) and at the axial point $z = z_0$? Compare this with the intensity at $z = z_0$ of a 100-W spherical wave produced by a small isotropically emitting light source located at $z = 0$.

EXERCISE 3.1-2

Validity of the Paraxial Approximation for a Gaussian Beam. The complex envelope $A(\mathbf{r})$ of a Gaussian beam is an exact solution of the paraxial Helmholtz equation (3.1-2), but its corresponding complex amplitude $U(\mathbf{r}) = A(\mathbf{r})\exp(-jkz)$ is only an approximate solution of the Helmholtz equation (2.2-7). This is because the paraxial Helmholtz equation is itself approximate. The approximation is satisfactory if the condition (2.2-21) is satisfied. Show that if the divergence angle θ_0 of a Gaussian beam is small ($\theta_0 \ll 1$), the necessary condition (2.2-21) for the validity of the paraxial Helmholtz equation is indeed satisfied.

EXERCISE 3.1-3

Determination of a Beam with Given Width and Curvature. Consider a Gaussian beam whose width W and radius of curvature R are known at a particular point on the beam axis (Fig. 3.1-9). Show that the beam waist is located to the left at a distance

$$z = \frac{R}{1 + (\lambda R/\pi W^2)^2} \tag{3.1-25}$$

and that the waist radius is

$$W_0 = \frac{W}{\sqrt{1 + (\pi W^2/\lambda R)^2}}\,. \tag{3.1-26}$$

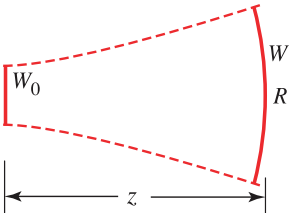


Figure 3.1-9 Given W and R, determine z and W_0.

EXERCISE 3.1-4

Determination of the Width and Curvature at One Point Given the Width and Curvature at Another Point. Assume that the width and radius of curvature of a Gaussian beam of wavelength $\lambda = 1$ μm at some point on the beam axis are $W_1 = 1$ mm and $R_1 = 1$ m, respectively (Fig. 3.1-10). Determine the beam width W_2 and radius of curvature R_2 at a distance $d = 10$ cm to the right.

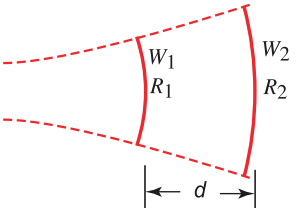


Figure 3.1-10 Given W_1, R_1 and d, determine W_2 and R_2.

EXERCISE 3.1-5

Identification of a Beam with Known Curvatures at Two Points. A Gaussian beam has radii of curvature R_1 and R_2 at two points on the beam axis separated by a distance d, as illustrated in Fig. 3.1-11. Verify that the location of the beam center and its depth of focus may be determined from the relations

$$z_1 = \frac{-d(R_2 - d)}{R_2 - R_1 - 2d} \tag{3.1-27}$$

$$z_0^2 = \frac{-d(R_1 + d)(R_2 - d)(R_2 - R_1 - d)}{(R_2 - R_1 - 2d)^2} \tag{3.1-28}$$

$$W_0 = \sqrt{\frac{\lambda z_0}{\pi}}.$$

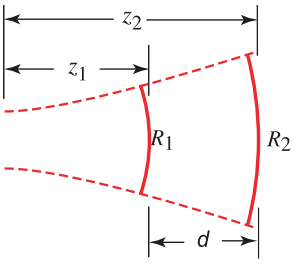


Figure 3.1-11 Given R_1, R_2, and d, determine z_1, z_2, z_0, and W_0.

C. Beam Quality

The Gaussian beam is an idealization that is only approximately met, even in well-designed laser systems. A measure of the quality of an optical beam is the deviation of its profile from Gaussian form. For a beam of waist diameter $2W_m$ and angular divergence $2\theta_m$, a useful numerical measure of the beam quality is provided by the $\mathbb{M}^2$-factor, which is defined as the ratio of the waist-diameter–divergence product, $2W_m \cdot 2\theta_m$ (usually measured in units of mm·mrad), to that expected for a Gaussian beam, which is $2W_0 \cdot 2\theta_0 = 4\lambda/\pi$. Thus,

$$\mathbb{M}^2 = \frac{2W_m \cdot 2\theta_m}{4\lambda/\pi}. \tag{3.1-29}$$

If the two beams have the same waist diameter, the $\mathbb{M}^2$-factor is simply the ratio of their angular divergences,

$$\mathbb{M}^2 = \theta_m/\theta_0, \tag{3.1-30}$$

where $\theta_0 = \lambda/\pi W_0 = \lambda/\pi W_m$ [see (3.1-21)]. Since the Gaussian beam enjoys the smallest possible divergence angle of all beams with the same waist diameter, $\mathbb{M}^2 \geq 1$. The specification of the $\mathbb{M}^2$-factor of an optical beam thus signifies a divergence angle that is $\mathbb{M}^2$ times greater than that of a Gaussian beam of the same waist diameter.

Optical beams produced by commonly available Helium–Neon lasers usually exhibit $\mathbb{M}^2 < 1.1$. For ion lasers, $\mathbb{M}^2$ is typically in the range 1.1–1.3. Collimated TEM_{00} diode-laser beams usually exhibit $\mathbb{M}^2 \approx$1.1–1.7, whereas high-energy multimode lasers display $\mathbb{M}^2$ factors as high as 3 or 4.

For an optical beam that is approximately Gaussian, the $\mathbb{M}^2$-factor may be determined by making use of a charge-coupled device (CCD) camera to measure the

intensity profile of the beam at various locations along the axis of the beam. The beam is focused, by a high-quality lens with a long focal length and large $F_{\#}$, to a size that is roughly the same as that of the CCD array [see (3.2-17)]. First, the beam center is located by finding the plane at which the spot size is minimized; the waist diameter $2W_m$ is then measured. The axial distance from the beam center to the plane at which the beam diameter increases by a factor of $\sqrt{2}$ provides the Rayleigh range z_m. An estimate of the angular divergence $2\theta_m$ is obtained by using the Gaussian-beam relation $\theta_m = \sqrt{\lambda/\pi z_m}$, which is obtained from (3.1-11) and (3.1-20). Finally, the $\mathbb{M}^2$-factor is computed by means of (3.1-29).

3.2 TRANSMISSION THROUGH OPTICAL COMPONENTS

We proceed now to a discussion of the effects of various optical components on a Gaussian beam. We demonstrate that if a Gaussian beam is transmitted through a set of circularly symmetric optical components aligned with the beam axis, *the Gaussian beam remains a Gaussian beam*, provided that the overall system maintains the paraxial nature of the wave. The beam is reshaped, however — its waist and curvature are altered. The results of this section are of importance in the design of optical instruments that rely on Gaussian beams.

A. Transmission Through a Thin Lens

The complex amplitude transmittance of a thin lens of focal length f is proportional to $\exp(jk\rho^2/2f)$ [see (2.4-9)]. When a Gaussian beam traverses such a component, its complex amplitude, given in (3.1-7), is multiplied by this phase factor. As a result, although the beam width is not altered ($W' = W$), the wavefront is.

To be specific consider a Gaussian beam centered at $z = 0$, with waist radius W_0, transmitted through a thin lens located at position z, as illustrated in Fig. 3.2-1. The phase of the incident wave at the plane of the lens is $kz + k\rho^2/2R - \zeta$, as prescribed by (3.1-23), where $R = R(z)$ and $\zeta = \zeta(z)$ are given in (3.1-9) and (3.1-10), respectively. The phase of the emerging wave therefore becomes

$$kz + k\frac{\rho^2}{2R} - \zeta - k\frac{\rho^2}{2f} = kz + k\frac{\rho^2}{2R'} - \zeta, \tag{3.2-1}$$

where

$$\frac{1}{R'} = \frac{1}{R} - \frac{1}{f}. \tag{3.2-2}$$

We conclude that the transmitted wave is itself a Gaussian beam with width $W' = W$ and radius of curvature R', where R' satisfies the imaging equation $1/R - 1/R' = 1/f$. The sign of R is positive since the wavefront of the incident beam is diverging whereas the opposite is true of R'.

The parameters of the emerging beam are determined by referring to the outcome of Exercise 3.1-3, in which the parameters of a Gaussian beam are determined from its width and curvature at a given point. Equation (3.1-26) provides that the waist radius is

$$W_0' = \frac{W}{\sqrt{1 + (\pi W^2/\lambda R')^2}} \tag{3.2-3}$$

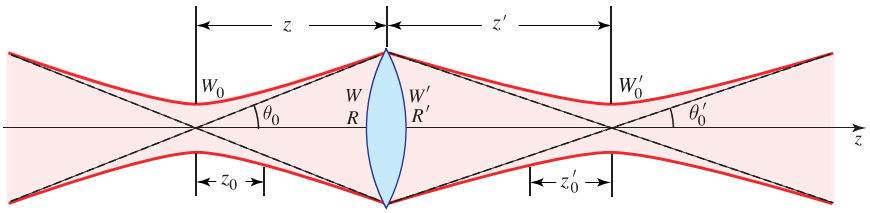


Figure 3.2-1 Transmission of a Gaussian beam through a thin lens.

whereas (3.1-25) provides that the beam center is located at a distance from the lens given by

$$-z' = \frac{R'}{1 + (\lambda R'/\pi W^2)^2}\,. \tag{3.2-4}$$

The minus sign in (3.2-4) indicates that the beam waist lies to the right of the lens. Substituting $W = W_0\sqrt{1 + (z/z_0)^2}$ and $R = z[1 + (z_0/z)^2]$ from (3.1-8) and (3.1-9) into (3.2-2) to (3.2-4) yields a set of formulas that relate the unprimed parameters of the Gaussian beam incident on the lens to the primed parameters of the Gaussian beam that emerges from the lens, as represented in Fig. 3.2-1:

Waist radius	$W_0' = MW_0$	(3.2-5)
Waist location	$(z' - f) = M^2(z - f)$	(3.2-6)
Depth of focus	$2z_0' = M^2(2z_0)$	(3.2-7)
Divergence angle	$2\theta_0' = \dfrac{2\theta_0}{M}$	(3.2-8)
Magnification	$M = \dfrac{M_r}{\sqrt{1 + r^2}}$	(3.2-9)
$r = \dfrac{z_0}{z - f}$,	$M_r = \left\lvert \dfrac{f}{z - f} \right\rvert$.	(3.2-9a)

Parameter Transformation by a Lens

The magnification factor M evidently plays an important role. The waist radius is magnified by M, the depth of focus is magnified by M^2, and the divergence angle is minified by M.

Limit of Ray Optics

Consider the limiting case in which $(z - f) \gg z_0$, so that the lens is well outside the depth of focus of the incident beam (Fig. 3.2-2). The beam may then be approximated by a spherical wave, and, in accordance with (3.2-9) and (3.2-9a), $r \ll 1$ so that $M \approx M_r$. In this case (3.2-5)–(3.2-9a) reduce to

$$W_0' \approx MW_0 \tag{3.2-10}$$

$$\frac{1}{z'} + \frac{1}{z} \approx \frac{1}{f} \tag{3.2-11}$$

$$M \approx M_r = \left| \frac{f}{z - f} \right|. \tag{3.2-12}$$

Equations (3.2-10)–(3.2-12) are precisely the relations provided by ray optics for the location and size of a patch of light of diameter $2W_0$ located at a distance z to the left of a thin lens (see Sec. 1.2C). Indeed, the magnification factor M_r is identically that based on ray optics. Since (3.2-9) provides that $M < M_r$, the maximum Gaussian-beam magnification attainable is the ray-optics limit M_r. As r^2 increases, the magnification is reduced and the deviation from ray optics widens. Equations (3.2-10)–(3.2-12) also correspond to the results obtained from wave optics for the focusing of a spherical wave in the paraxial approximation (see Sec. 2.4B).

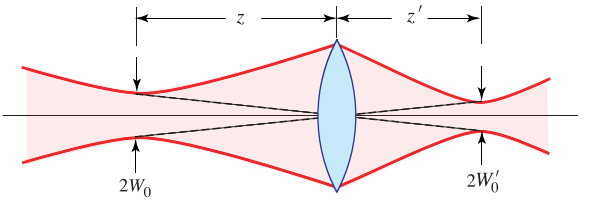


Figure 3.2-2 Beam imaging in the ray-optics limit.

.

B. Beam Shaping

A lens, or sequence of lenses, may be used to reshape a Gaussian beam without compromising its Gaussian nature. Of course, graded-index components can serve this purpose as well.

Beam Focusing

For a lens placed at the waist of a Gaussian beam, as illustrated in Fig. 3.2-3, the appropriate parameter-transformation formulas are obtained by simply substituting $z = 0$ in (3.2-5) to (3.2-9a). The transmitted beam is then focused to a waist radius W_0' at a distance z' given by

$$W_0' = \frac{W_0}{\sqrt{1 + (z_0/f)^2}} \tag{3.2-13}$$

$$z' = \frac{f}{1 + (f/z_0)^2}. \tag{3.2-14}$$

In the special case when the depth of focus of the incident beam $2z_0$ is much longer than the focal length f of the lens, as illustrated in Fig. 3.2-4, (3.2-13) reduces to $W_0' \approx (f/z_0)W_0$. Using $z_0 = \pi W_0^2/\lambda$ from (3.1-11), along with (3.1-20), then leads to the simple result

$$W_0' \approx \frac{\lambda}{\pi W_0} f = \theta_0 f \tag{3.2-15}$$

$$z' \approx f. \tag{3.2-16}$$

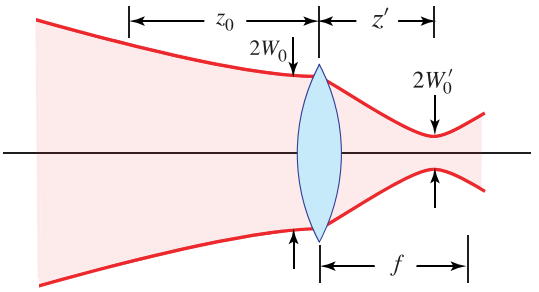


Figure 3.2-3 Focusing a Gaussian beam with a lens at the beam waist.

The transmitted beam is then focused in the focal plane of the lens as would be expected for a collimated beam of parallel rays impinging on the lens. This result emerges because, at its waist, the incident Gaussian beam is well approximated by a plane wave. Wave optics provides that the focused waist radius W_0' is directly proportional to the wavelength and the focal length, and inversely proportional to the radius of the incident beam. The spot size expected from ray optics is, of course, zero, a result that is indeed obtained from the wave-optics formulas as $\lambda \to 0$.

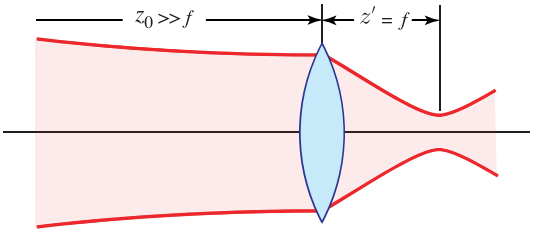


Figure 3.2-4 Focusing a collimated beam.

In many applications, such as laser scanning, laser printing, compact-disc (CD) burning, and laser fusion, it is desired to generate the smallest possible spot size. It is clear from (3.2-15) that this is achieved by making use of the shortest possible wavelength, the thickest incident beam, and the shortest focal-length lens. Since the lens must intercept the incident beam, its diameter D should be at least $2W_0$. Taking $D = 2W_0$, and making use of (3.2-15), the diameter of the focused spot is given by

$$\boxed{2W_0' \approx \frac{4}{\pi}\lambda F_\#} \qquad F_\# = \frac{f}{D}, \tag{3.2-17}$$

Focused Spot Size

where the F-number of the lens is denoted $F_\#$. A microscope objective with small F-number is often used for this purpose. A caveat is in order: since (3.2-15) and (3.2-16) are approximate their validity must always be confirmed before use.

EXERCISE 3.2-1

Beam Relaying. A Gaussian beam of radius W_0 and wavelength λ is repeatedly focused by a sequence of identical lenses, each of focal length f and separated by a distance d (Fig. 3.2-5). The focused waist radius is equal to the incident waist radius, i.e., $W_0' = W_0$. Using (3.2-6), (3.2-9), and (3.2-9a) show that this condition can arise only if the inequality $d \leq 4f$ is satisfied. Note that this is the same as the ray-confinement condition for a sequence of lenses derived in Example 1.4-1 using ray optics.

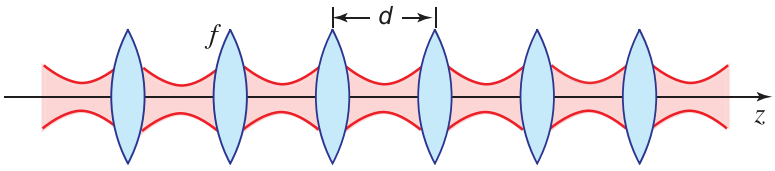


Figure 3.2-5 Beam relaying.

EXERCISE 3.2-2

Beam Collimation. A Gaussian beam is transmitted through a thin lens of focal length f.

(a) Show that the locations of the waists of the incident and transmitted beams, z and z', respectively, are related by

$$\frac{z'}{f} - 1 = \frac{z/f - 1}{(z/f - 1)^2 + (z_0/f)^2}. \tag{3.2-18}$$

This relation is plotted in Fig. 3.2-6.

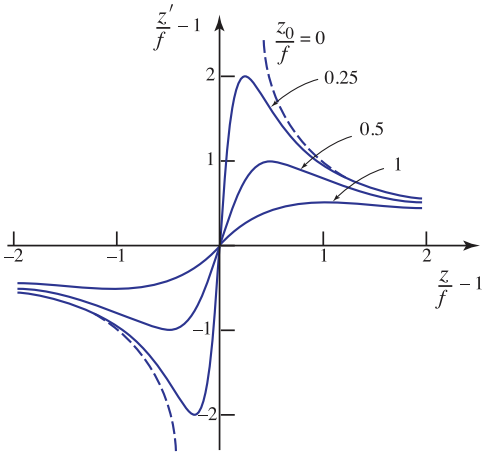


Figure 3.2-6 Relation between the waist locations of the incident and transmitted beams.

(b) The beam is collimated by making the location of the new waist z' as distant as possible from the lens. This is achieved by using the smallest possible ratio z_0/f (short depth of focus and long focal length). For a given ratio z_0/f, show that the optimal value of z for collimation is $z = f + z_0$.

(c) Given $\lambda = 1$ μm, $z_0 = 1$ cm, and $f = 50$ cm, determine the optimal value of z for collimation, and the corresponding magnification M, distance z', and width W_0' of the collimated beam.

EXERCISE 3.2-3

Beam Expansion. A Gaussian beam may be expanded and collimated by using two lenses of focal lengths f_1 and f_2, as illustrated in Fig. 3.2-7. Parameters of the initial beam (W_0, z_0) are modified by the first lens to (W_0'', z_0'') and subsequently altered by the second lens to (W_0', z_0'). The first lens, which has a short focal length, serves to reduce the depth of focus $2z_0''$ of the beam. This prepares it for collimation by the second lens, which has a long focal length. The system functions as an inverse Keplerian telescope.

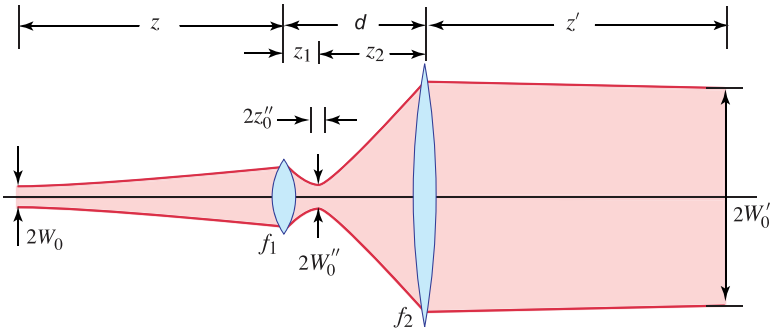


Figure 3.2-7 Beam expansion using a two-lens system.

(a) Assuming that $f_1 \ll z$ and $z - f_1 \gg z_0$, use the results of Exercise 3.2-2 to determine the optical distance d between the lenses such that the distance z' to the waist of the final beam is as large as possible.

(b) Determine an expression for the overall magnification $M = W_0'/W_0$ of the system.

C. Reflection from a Spherical Mirror

We now examine the reflection of a Gaussian beam from a spherical mirror. The complex amplitude reflectance of the mirror is proportional to $\exp(-jk\rho^2/R)$ (see Prob. 2.4-10), where by convention $R > 0$ for convex mirrors and $R < 0$ for concave mirrors. The action of the mirror on a Gaussian beam of width W_1 and radius of curvature R_1 is therefore to reflect the beam and to modify its phase by the factor $-k\rho^2/R$, while leaving the beam width unaltered. The reflected beam therefore remains Gaussian, with parameters W_2 and R_2 given by

$$W_2 = W_1 \tag{3.2-19}$$

$$\frac{1}{R_2} = \frac{1}{R_1} + \frac{2}{R}. \tag{3.2-20}$$

Equation (3.2-20) is identical to (3.2-2) provided $f = -R/2$. Thus, the Gaussian beam is modified in precisely the same way as it is by a lens, except for a reversal of the direction of propagation.

Three special cases, illustrated in Fig. 3.2-8, are of interest:

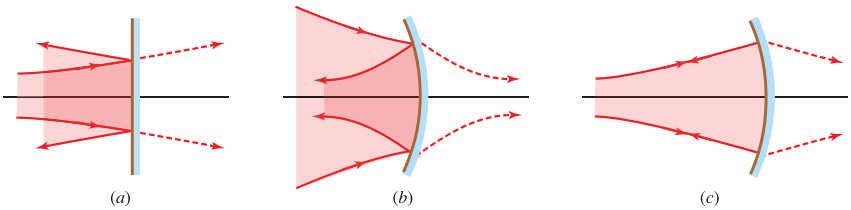


Figure 3.2-8 Reflection of a Gaussian beam with radius of curvature R_1 from a mirror with radius of curvature R: (*a*) $R = \infty$; (*b*) $R_1 = \infty$; (*c*) $R_1 = -R$. The dashed curves show the effects of replacing the mirror by a lens of focal length $f = -R/2$.

- If the *mirror is planar*, i.e., $R = \infty$, then $R_2 = R_1$, so that the mirror reverses the direction of the beam without altering its curvature, as illustrated in Fig. 3.2-8(a).
- If $R_1 = \infty$, i.e., if the *beam waist lies on the mirror*, then $R_2 = R/2$. If the mirror is concave ($R < 0$), $R_2 < 0$ so that the reflected beam acquires a negative curvature and the wavefronts converge. The mirror then focuses the beam to a smaller spot size, as illustrated in Fig. 3.2-8(b).
- If $R_1 = -R$, i.e., if the incident *beam has the same curvature as the mirror*, then $R_2 = R$. The wavefronts of both the incident and reflected waves then coincide with the mirror and the wave retraces its path as shown in Fig. 3.2-8(c). This is expected since the wavefront normals are also normal to the mirror so that the mirror reflects the wave back onto itself. In the illustration in Fig. 3.2-8(c) the mirror is concave ($R < 0$); the incident wave is diverging ($R_1 > 0$) and the reflected wave is converging ($R_2 < 0$).

EXERCISE 3.2-4

Variable-Reflectance Mirrors. A spherical mirror of radius R has a variable power reflectance characterized by $\mathcal{R}(\rho) = |\mathrm{r}(\rho)|^2 = \exp(-2\rho^2/W_m^2)$, which is a Gaussian function of the radial distance ρ. The reflectance is unity on axis and falls by a factor $1/e^2$ when $\rho = W_m$. Determine the effect of the mirror on a Gaussian beam with radius of curvature R_1 and beam width W_1 at the mirror.

*D. Transmission Through an Arbitrary Optical System

In the paraxial ray-optics approximation, an optical system is completely characterized by the 2×2 ray-transfer matrix relating the position and inclination of the transmitted ray to those of the incident ray (see Sec. 1.4). We now consider how an arbitrary paraxial optical system, characterized by a matrix $\mathbf{M}$ of elements (A, B, C, D), modifies a Gaussian beam (Fig. 3.2-9).

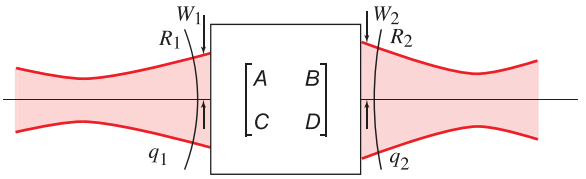


Figure 3.2-9 Modification of a Gaussian beam by an arbitrary paraxial system described by an $ABCD$ matrix.

The ABCD Law

The q-parameters, q_1 and q_2, of the incident and transmitted Gaussian beams at the input and output planes of a paraxial optical system described by the (A, B, C, D) matrix are related by

$$q_2 = \frac{Aq_1 + B}{Cq_1 + D}. \tag{3.2-21}$$

The $ABCD$ Law

Because the complex q-parameter identifies the width W and radius of curvature R of the Gaussian beam (see Exercise 3.1-3), this simple expression, called the $ABCD$ **law**, governs the effect of an arbitrary paraxial system on a Gaussian beam. The $ABCD$ law will be established by verification in special cases; its generality will ultimately be proved by induction.

Transmission Through Free Space

When the optical system is a distance d of free space (or of any homogeneous medium), the elements of the ray-transfer matrix $\mathbf{M}$ are $A = 1$, $B = d$, $C = 0$, $D = 1$ [see (1.4-4)]. Since it has been established earlier that $q = z + jz_0$ in free space, the q-parameter is modified by the optical system in accordance with $q_2 = q_1 + d$. This is, in fact, is equal to $(1 \cdot q_1 + d)/(0 \cdot q_1 + 1)$ so that the $ABCD$ law is seen to apply.

Transmission Through a Thin Optical Component

An arbitrary thin optical component does not affect the ray position so that

$$y_2 = y_1, \tag{3.2-22}$$

but does alter the inclination angle in accordance with

$$\theta_2 = Cy_1 + D\theta_1, \tag{3.2-23}$$

as illustrated in Fig. 3.2-10. Thus, $A = 1$ and $B = 0$, but C and D are arbitrary. However, in all of the thin optical components described in Sec. 1.4B, $D = n_1/n_2$. By virtue of the vanishing thickness of the component, the beam width does not change, i.e.,

$$W_2 = W_1. \tag{3.2-24}$$

Moreover, if the beams at the input and output planes of the component are approximated by spherical waves of radii R_1 and R_2, respectively, then in the paraxial approximation, when θ_1 and θ_2 are small, $\theta_1 \approx y_1/R_1$ and $\theta_2 \approx y_2/R_2$. Substituting these expressions into (3.2-23), with the help of (3.2-22) we obtain

$$\frac{1}{R_2} = C + \frac{D}{R_1}. \tag{3.2-25}$$

Using (3.1-6), which is the expression for q as a function of R and W, and noting that $D = n_1/n_2 = \lambda_2/\lambda_1$, (3.2-24) and (3.2-25) can be combined into a single equation,

$$\frac{1}{q_2} = C + \frac{D}{q_1}, \tag{3.2-26}$$

from which $q_2 = (1 \cdot q_1 + 0)/(Cq_1 + D)$, so that the $ABCD$ law again applies.

Invariance of the ABCD Law to Cascading

If the $ABCD$ law is applicable to each of two optical systems with matrices $\mathbf{M}_i = (A_i, B_i, C_i, D_i)$, $i = 1, 2$, it must also apply to a system comprising their cascade (a system with matrix $\mathbf{M} = \mathbf{M}_2\mathbf{M}_1$). This may be shown by straightforward substitution.

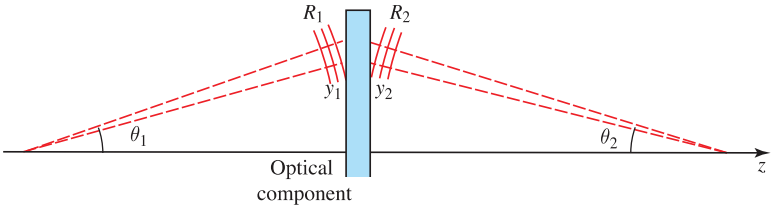


Figure 3.2-10 Modification of a Gaussian beam by a thin optical component.

Generality of the ABCD Law

Since the $ABCD$ law applies to thin optical components as well as to propagation in a homogeneous medium, it also applies to any combination thereof. All of the paraxial optical systems of interest are combinations of propagation in homogeneous media and thin optical components such as thin lenses and mirrors. It is therefore apparent that the $ABCD$ law is applicable to all of these systems. Furthermore, since an inhomogeneous continuously varying medium may be regarded as a cascade of incremental thin elements followed by incremental distances, we conclude that the $ABCD$ law applies to these systems as well, provided that all rays (wavefront normals) remain paraxial.

EXERCISE 3.2-5

Transmission of a Gaussian Beam Through a Transparent Plate. Use the $ABCD$ law to examine the transmission of a Gaussian beam from air, through a transparent plate of refractive index n and thickness d, and again into air. Assume that the beam axis is normal to the plate.

3.3 HERMITE–GAUSSIAN BEAMS

The Gaussian beam is not the only beam-like solution of the paraxial Helmholtz equation (3.1-2). Of particular interest are solutions that exhibit non-Gaussian intensity distributions but share the wavefronts of the Gaussian beam. Such beams have the salutary feature of being able to match the curvatures of spherical mirrors of large radius, such as those that form an optical resonator, and reflect between them without being altered. Such self-reproducing waves are called the **modes** of the resonator (see Appendix C). The optics of resonators is discussed in Chapter 11.

Consider a Gaussian beam of complex envelope [see (3.1-5)]

$$A_G(x,y,z) = \frac{A_1}{q(z)} \exp\left[-jk\frac{x^2+y^2}{2q(z)}\right], \tag{3.3-1}$$

where $q(z) = z + jz_0$. Expressions for the beam width $W(z)$ and the wavefront radius of curvature $R(z)$ are provided in (3.1-8) and (3.1-9), respectively. Now consider a second wave whose complex envelope is a modulated version of the Gaussian beam,

$$A(x,y,z) = \mathcal{X}\left[\sqrt{2}\frac{x}{W(z)}\right] \mathcal{Y}\left[\sqrt{2}\frac{y}{W(z)}\right] \exp[j\mathcal{Z}(z)]\, A_G(x,y,z), \tag{3.3-2}$$

where $\mathcal{X}(\cdot)$, $\mathcal{Y}(\cdot)$, and $\mathcal{Z}(\cdot)$ are real functions. This wave, should it be shown to exist, has the following two properties:

1. The phase is the same as that of the underlying Gaussian wave, except for an excess phase $\mathcal{Z}(z)$ that is independent of x and y. If $\mathcal{Z}(z)$ is a slowly varying function of z, both waves have wavefronts with the same radius of curvature $R(z)$. These two waves are therefore focused by thin lenses and mirrors in precisely the same manner.
2. The magnitude

$$A_0 \mathcal{X}\left[\sqrt{2}\frac{x}{W(z)}\right] \mathcal{Y}\left[\sqrt{2}\frac{y}{W(z)}\right] \left[\frac{W_0}{W(z)}\right] \exp\left[-\frac{x^2+y^2}{W^2(z)}\right], \tag{3.3-3}$$

where $A_0 = A_1/jz_0$, is a function of $x/W(z)$ and $y/W(z)$ whose widths in the x and y directions vary with z in accordance with the same scaling factor $W(z)$. As z increases, the intensity distribution in the transverse plane remains fixed, except for a magnification factor $W(z)$. This distribution is a Gaussian function modulated in the x and y directions by the functions $\mathcal{X}^2(\cdot)$ and $\mathcal{Y}^2(\cdot)$, respectively.

The modulated wave therefore represents a beam of non-Gaussian intensity distribution, but it shares the same wavefronts and angular divergence as the underlying Gaussian wave.

The existence of this wave is assured if three real functions $\mathcal{X}(\cdot)$, $\mathcal{Y}(\cdot)$, and $\mathcal{Z}(z)$ can be found such that (3.3-2) satisfies the paraxial Helmholtz equation (3.1-2). Substituting (3.3-2) into (3.1-2), using the fact that A_G itself satisfies (3.1-2), and defining two new variables $u = \sqrt{2}\,x/W(z)$ and $v = \sqrt{2}\,y/W(z)$, we obtain

$$\frac{1}{\mathcal{X}}\left(\frac{\partial^2 \mathcal{X}}{\partial u^2} - 2u\frac{\partial \mathcal{X}}{\partial u}\right) + \frac{1}{\mathcal{Y}}\left(\frac{\partial^2 \mathcal{Y}}{\partial v^2} - 2v\frac{\partial \mathcal{Y}}{\partial v}\right) + kW^2(z)\frac{\partial \mathcal{Z}}{\partial z} = 0. \tag{3.3-4}$$

Since the left-hand side of this equation is the sum of three terms, each of which is a function of a single independent variable, u, v, and z, respectively, each of these terms must be constant. Equating the first term to the constant $-2\mu_1$ and the second to $-2\mu_2$, the third must be equal to $2(\mu_1 + \mu_2)$. This technique of "separation of variables" permits us to reduce the partial differential equation (3.3-4) into three ordinary differential equations, for $\mathcal{X}(u)$, $\mathcal{Y}(v)$, and $\mathcal{Z}(z)$, respectively:

$$-\frac{1}{2}\frac{d^2\mathcal{X}}{du^2} + u\frac{d\mathcal{X}}{du} = \mu_1 \mathcal{X} \tag{3.3-5a}$$

$$-\frac{1}{2}\frac{d^2\mathcal{Y}}{dv^2} + v\frac{d\mathcal{Y}}{dv} = \mu_2 \mathcal{Y} \tag{3.3-5b}$$

$$z_0\left[1 + \left(\frac{z}{z_0}\right)^2\right]\frac{d\mathcal{Z}}{dz} = \mu_1 + \mu_2, \tag{3.3-5c}$$

where we have made use of the expression for $W(z)$ given in (3.1-8) and (3.1-11).

Equation (3.3-5a) represents an eigenvalue problem (see Appendix C) whose eigenvalues are $\mu_l = l$, where $l = 0, 1, 2, \ldots$ and whose eigenfunctions are the **Hermite polynomials** $\mathcal{X}(u) = \mathbb{H}_l(u)$, $l = 0, 1, 2, \ldots$. These polynomials are defined by the recurrence relation

$$\mathbb{H}_{l+1}(u) = 2u\,\mathbb{H}_l(u) - 2l\,\mathbb{H}_{l-1}(u) \tag{3.3-6}$$

with

$$\mathbb{H}_0(u) = 1, \qquad \mathbb{H}_1(u) = 2u. \tag{3.3-7}$$

Thus,

$$\mathbb{H}_2(u) = 4u^2 - 2, \qquad \mathbb{H}_3(u) = 8u^3 - 12u, \quad \ldots . \tag{3.3-8}$$

Similarly, the solutions of (3.3-5b) are $\mu_2 = m$ and $\mathcal{Y}(v) = \mathbb{H}_m(v)$, where $m = 0, 1, 2, \ldots$. There is therefore a family of solutions labeled by the indices (l, m). Substituting $\mu_1 = l$ and $\mu_2 = m$ in (3.3-5c), and integrating, we obtain

$$\mathcal{Z}(z) = (l + m)\,\zeta(z), \tag{3.3-9}$$

where $\zeta(z) = \tan^{-1}(z/z_0)$. The excess phase $\mathcal{Z}(z)$ thus varies slowly between $-(l + m)\,\pi/2$ and $+(l + m)\,\pi/2$, as z varies between $-\infty$ and ∞ (see (3.1-10) and Fig. 3.1-5).

Complex Amplitude

Finally, substitution into (3.3-2) yields an expression for the complex envelope of the beam labeled by the indices (l, m). Rearranging terms and multiplying by $\exp(-jkz)$ provides the complex amplitude

$$U_{l,m}(x, y, z) = A_{l,m} \left[\frac{W_0}{W(z)}\right] \mathbb{G}_l\left[\frac{\sqrt{2}\,x}{W(z)}\right] \mathbb{G}_m\left[\frac{\sqrt{2}\,y}{W(z)}\right] \times \exp\left[-jkz - jk\frac{x^2 + y^2}{2R(z)} + j(l + m + 1)\,\zeta(z)\right] \tag{3.3-10}$$

(3.3-10) Hermite–Gaussian Beam

where

$$\mathbb{G}_l(u) = \mathbb{H}_l(u)\,\exp\left(\frac{-u^2}{2}\right), \qquad l = 0, 1, 2, \ldots \tag{3.3-11}$$

is known as the **Hermite–Gaussian function** of order l, and $A_{l,m}$ is a constant.

Since $\mathbb{H}_0(u) = 1$, the Hermite–Gaussian function of order 0 is simply the Gaussian function. Continuing to higher order, $\mathbb{G}_1(u) = 2u\exp(-u^2/2)$ is an odd function, $\mathbb{G}_2(u) = (4u^2 - 2)\exp(-u^2/2)$ is even, $\mathbb{G}_3(u) = (8u^3 - 12u)\exp(-u^2/2)$ is odd, and so on. These functions are displayed schematically in Fig. 3.3-1.

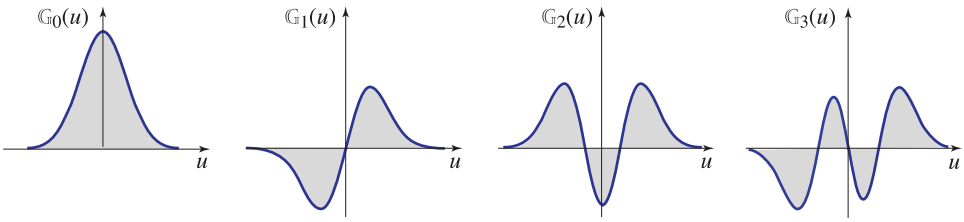


Figure 3.3-1 Low-order Hermite–Gaussian functions: (*a*) $\mathbb{G}_0(u)$; (*b*) $\mathbb{G}_1(u)$; (*c*) $\mathbb{G}_2(u)$; (*d*) $\mathbb{G}_3(u)$.

An optical wave with complex amplitude given by (3.3-10) is known as a **Hermite–Gaussian beam** of order (l, m), which is often denoted HG_{lm}. The Hermite–Gaussian beam of order $(0, 0)$, namely HG_{00}, is the simple Gaussian beam.

Intensity Distribution

The optical intensity of the HG_{lm} Hermite–Gaussian beam, $I_{l,m} = |U_{l,m}|^2$, is given by

$$I_{l,m}(x, y, z) = |A_{l,m}|^2 \left[\frac{W_0}{W(z)}\right]^2 \mathbb{G}_l^2\left[\frac{\sqrt{2}\,x}{W(z)}\right] \mathbb{G}_m^2\left[\frac{\sqrt{2}\,y}{W(z)}\right]. \tag{3.3-12}$$

Figure 3.3-2 illustrates the dependence of the intensity on the normalized transverse distances $u = \sqrt{2}\,x/W(z)$ and $v = \sqrt{2}\,y/W(z)$ for several values of l and m. Beams of higher order have larger widths than those of lower order, as is evident in Fig. 3.3-1. Regardless of the order, however, the width of the beam is proportional to $W(z)$; thus, as z increases, the transverse spatial extent of the intensity pattern is magnified by the factor $W(z)/W_0$ but otherwise maintains its profile. The only circularly symmetric member of the family of Hermite–Gaussian beams is the elementary Gaussian beam itself.

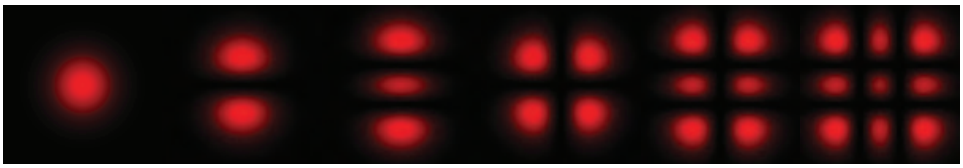


Figure 3.3-2 Intensity distributions of several low-order Hermite–Gaussian beams, HG_{lm}, in the transverse plane. The HG_{00} beam is the elementary Gaussian beam displayed in Fig. 3.1-1.

The Hermite–Gaussian beam defined in (3.3-10) may be generalized by ascribing different beam widths to its x and y components, $W_x(z)$ and $W_y(z)$, respectively, thereby defining the **elliptic Hermite–Gaussian beam**. Because (3.3-10) is a separable function of x and y, this constitutes yet another exact solution of the paraxial Helmholtz equation. A special case is the **elliptic Gaussian beam** that appears in Prob. 3.1-8; it exhibits elliptical, rather than circular, contours of constant intensity.

3.4 LAGUERRE–GAUSSIAN BEAMS

Laguerre–Gaussian Beams

The Hermite–Gaussian beams form a complete set of solutions to the paraxial Helmholtz equation. Any other solution can be written as a superposition of these beams. An alternate complete set of solutions, known as **Laguerre–Gaussian beams**, is obtained by writing the paraxial Helmholtz equation in cylindrical coordinates (ρ, ϕ, z) and then using the separation-of-variables technique in ρ and ϕ, rather than in x and y. The complex amplitude of the Laguerre–Gaussian beam, denoted LG_{lm}, can be expressed as

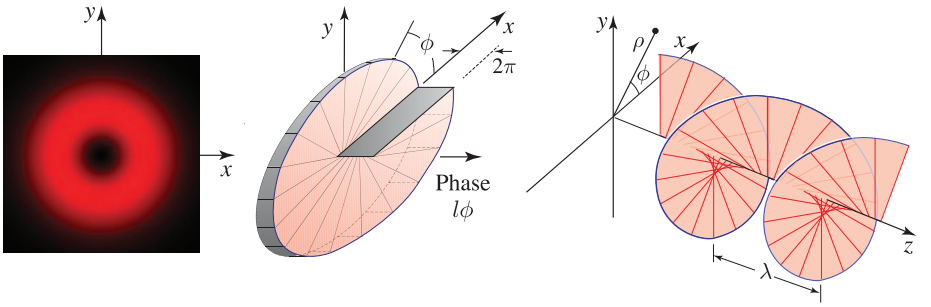
$$U_{l,m}(\rho, \phi, z) = A_{l,m}\left[\frac{W_0}{W(z)}\right]\left(\frac{\rho}{W(z)}\right)^l \mathbb{L}_m^l\left(\frac{2\rho^2}{W^2(z)}\right) \exp\left(-\frac{\rho^2}{W^2(z)}\right)$$
$$\times \exp\left[-jkz - jk\frac{\rho^2}{2R(z)} \mp jl\phi + j(l + 2m + 1)\,\zeta(z)\right], \tag{3.4-1}$$

where the $\mathbb{L}_m^l(\cdot)$ represent generalized Laguerre polynomials,[†] and where $W(z)$, $R(z)$, $\zeta(z)$, and W_0 are given by (3.1-8)–(3.1-11). The integers $l = 0, 1, 2, \ldots$ and $m = 0, 1, 2, \ldots$ are azimuthal and radial indices, respectively. The lowest-order Laguerre–Gaussian beam LG_{00}, like the lowest-order Hermite–Gaussian beam HG_{00}, is the simple Gaussian beam.

EXERCISE 3.4-1

Laguerre–Gaussian Beam as a Superposition of Hermite–Gaussian Beams. Demonstrate that the Laguerre–Gaussian beam LG_{10} is equivalent to the superposition of two Hermite–Gaussian beams, HG_{01} and HG_{10}, with equal amplitudes and a phase shift of $\pi/2$, i.e., $\text{LG}_{10} \equiv \frac{1}{\sqrt{2}}(\text{HG}_{01} + j\text{HG}_{10})$.

The intensity of the Laguerre–Gaussian beam, which is proportional to the absolute square of (3.4-1), is a function of ρ and z, but not of ϕ, so that it is circularly symmetric. As illustrated in Fig. 3.4-1(a), the transverse intensity distribution for the LG_{10} beam assumes a toroidal shape. Its peak value is attained at a radius of $\rho = \sqrt{1/2}\, W(z)$, which increases with the distance z from the beam center (much as for the Gaussian beam). Beams of any order $l \neq 0$ are also toroidal when $m = 0$, and attain their peak values at radii $\sqrt{l/2}W(z)$. All beams with $l \neq 0$ have zero intensity at the beam center ($\rho = 0$); those with a radial index $m > 0$ take the form of multiple rings.



(a) Intensity (b) Phase (c) Wavefront

Figure 3.4-1 The Laguerre–Gaussian beam LG_{10}. (a) The transverse intensity distribution takes the form of a toroid. (b) The phase component $l\phi$, plotted for $l = 1$, is a linear function of the azimuthal angle ϕ. (c) The wavefront is a left-handed helical surface that undergoes corkscrew-like motion as it travels in the z direction.

The phase behavior of the Laguerre–Gaussian beam has the same dependence on ρ and z as does the Gaussian beam [see (3.1-7)], with two notable exceptions: (1) the Gouy phase is enhanced by the factor $(l + 2m + 1)$, and (2) there is an additional phase factor $e^{\mp jl\phi}$ that is proportional to the azimuthal angle ϕ. The phase component $l\phi$, illustrated in Fig. 3.4-1(b) for $l = 1$, is associated with the phase factor $\exp(-jl\phi)$ [see (3.1-23) and associated footnote]. It results in the wavefront assuming the form of a left-handed helix that undergoes corkscrew-like motion as the wave advances in the z direction, as shown in Fig. 3.4-1(c). Beams with $l > 1$ have wavefronts comprising l distinct but intertwined helices. The pitch of each helix is $l\lambda$ and the $\mp$ sign determines its handedness.

[†] The generalized Laguerre polynomials are expressible as $\mathbb{L}_m^l(x) = (m+l)! \sum_{i=0}^{m} (-x)^i / [i!(m-i)!(l+i)!]$. A few elementary examples are $\mathbb{L}_0^l(x) = 1$; $\mathbb{L}_1^l(x) = 1 - x + l$; $\mathbb{L}_2^l(x) = \frac{1}{2}[x^2 - 2(l+2)x + (l+1)(l+2)]$. The generalized Laguerre polynomials $\mathbb{L}_m^l(x)$ reduce to the simple Laguerre polynomials $\mathbb{L}_m(x)$ when $l = 0$.

As discussed in Sec. 2.5A, the phase of an optical beam may be determined by detecting its interference with an auxiliary optical field of known form (e.g., a plane wave). The phase of a Laguerre–Gaussian beam can be readily observed by detecting its superposition with another Laguerre–Gaussian beam of the same order but opposite handedness. Such a superposition, which constitutes a form of standing wave, has an intensity proportional to $|\exp(-jl\phi)+\exp(jl\phi)|^2 = 4\cos^2(l\phi)$, explicitly illustrating that the resulting intensity is sensitive to $l\phi$ as shown in Fig. 3.4-2. The number of angular interference fringes is equal to $2l$.

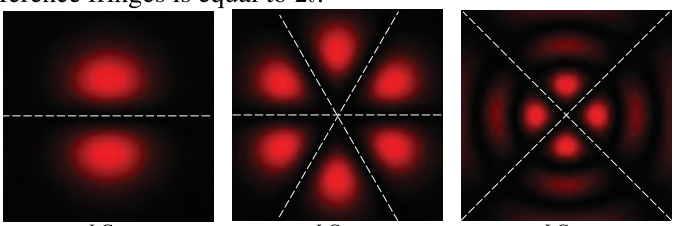


Figure 3.4-2 Transverse intensity distributions of the superposition of two Laguerre–Gaussian beams of the same order LG_{lm} but opposite handedness. The dashed white lines signify the loci of zero intensity (the nodes of the standing waves). The number of such lines is equal to the azimuthal order l. As the azimuthal angle ϕ moves from one node to the next, the phase changes by 2π.

Laguerre–Gaussian beams may be directly generated as laser modes or as combinations of Hermite–Gaussian laser modes, as discussed in Exercise 3.4-1. A Gaussian beam may be converted into an Laguerre–Gaussian beam by imparting to it the phase factor $\exp(-jl\phi)$ with the help of a spiral phase plate, a dielectric slab whose optical thickness increases linearly with ϕ [see Fig. 3.4-1(b)]. However, the method of choice for converting a Gaussian beam to a Laguerre–Gaussian beam is to make use of a diffractive optical element, or hologram, endowed with a fork dislocation centered on the beam axis, as exhibited in Example 4.5-3.

Beams with spiral phase carry orbital angular momentum. This may be understood by observing that an optical wave carries linear momentum that points along the direction orthogonal to its wavefronts (see Secs. 5.1 and 13.1D), which is also the direction of the optical rays. Since rays orthogonal to the helical wavefront of a Laguerre–Gaussian beam have azimuthal components that revolve about the beam axis, their linear momentum is accompanied by orbital angular momentum. This can also be visualized by considering that refracted optical rays incident on the surface of a spiral phase plate acquire azimuthal components [see Fig. 3.4-1(b)]. By virtue of their orbital angular momentum, Laguerre–Gaussian beams can exert a mechanical torque on micro-objects and can thus be used to manipulate microparticles.

Optical Vortices

An **optical vortex** is an optical field that exhibits a *line* of zero optical intensity, such as the line along the axis of a Laguerre–Gaussian beam with $l \neq 0$. It is also called a *screw dislocation* since the phase of the field is twisted like a corkscrew about the axis of travel. An optical vortex in a plane is a *point* at which the optical field vanishes; it is also called a *phase singularity*. An example of the latter is the point $(x, y) = (0, 0)$ in the transverse plane of the Laguerre–Gaussian beam illustrated in Fig. 3.4-1(a).

The strength of a vortex is indicated by its **topological charge**, which is determined by the number of full twists that the phase undergoes in a distance of one wavelength. For the Laguerre–Gaussian beam, the topological charge is the azimuthal index l, which is indicated by the number of lines of zero intensity that appear in the standing wave generated by the combination of two beams of the same order but opposite handedness, as illustrated in Fig. 3.4-2. This number also determines the orbital angular momentum of the associated photon, as will be discussed in Sec. 13.1D.

Optical vortex beams can assume forms that are far more complex than the simple Laguerre–Gaussian beam. Interference among three or more randomly directed plane-wave components of similar intensities always results in a field cross-section that contains many vortices. Such beams often exhibit unusual and dramatic properties — the field surrounding a vortex can, for example, tangle and form links and knots.[†]

Ince–Gaussian Beams

As discussed in Sec. 3.3 and at the beginning of this section, Hermite–Gaussian and Laguerre–Gaussian beams form complete sets of exact solutions to the paraxial Helmholtz equation, in Cartesian and cylindrical coordinates, respectively. A third complete set of exact solutions, known as Ince–Gaussian (IG) beams,[‡] exists in elliptic cylindrical coordinates, another three-dimensional orthogonal coordinate system. The transverse structure of these beams is characterized by Ince polynomials, which have an intrinsic elliptical character. Laguerre–Gaussian and Hermite–Gaussian beams are limiting forms of Ince–Gaussian beams when the ellipticity parameter is 0 and ∞, respectively.

3.5 NONDIFFRACTING BEAMS

A. Bessel Beams

In the search for beam-like waves, it is natural to attempt to construct waves whose wavefronts are planar but whose intensity distributions are nonuniform in the transverse plane. Consider, for example, a wave with complex amplitude

$$U(\mathbf{r}) = A(x, y)\, e^{-j\beta z}. \tag{3.5-1}$$

In order that this wave satisfy the Helmholtz equation (2.2-7), $\nabla^2 U + k^2 U = 0$, the quantity $A(x, y)$ must satisfy

$$\nabla_T^2 A + k_T^2 A = 0, \tag{3.5-2}$$

where $k_T^2 + \beta^2 = k^2$ and $\nabla_T^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the transverse Laplacian operator. Equation (3.5-2), known as the **two-dimensional Helmholtz equation**, may be solved by employing the method of separation of variables. Using polar coordinates ($x = \rho\cos\phi,\ y = \rho\sin\phi$), the result turns out to be

$$A(x, y) = A_m\, J_m(k_T\rho)\, e^{jm\phi}, \qquad m = 0, \pm 1, \pm 2, \ldots, \tag{3.5-3}$$

where $J_m(\cdot)$ is the Bessel function of the first kind and mth order, and A_m is a constant. Solutions of (3.5-3) that are singular at $\rho = 0$ are not included.

For $m = 0$, the wave has complex amplitude

$$U(\mathbf{r}) = A_0\, J_0(k_T\rho)\, e^{-j\beta z}, \tag{3.5-4}$$

and therefore has planar wavefronts — the wavefront normals (rays) are all parallel to the z axis. The intensity distribution $I(\rho, \phi, z) = |A_0|^2 J_0^2(k_T\rho)$ is circularly symmetric

[†] See M. R. Dennis, R. P. King, B. Jack, K. O'Holleran, and M. J. Padgett, Isolated Optical Vortex Knots, *Nature Physics*, vol. 6, pp. 118–121, 2010.

[‡] See M. A. Bandres and J. C. Gutiérrez-Vega, Ince–Gaussian Beams, *Optics Letters*, vol. 29, pp. 144–146, 2004.

and varies with ρ as illustrated in Fig. 3.5-1(a); it is independent of z so that there is no spread of the optical power. This wave is known as the **Bessel beam**.

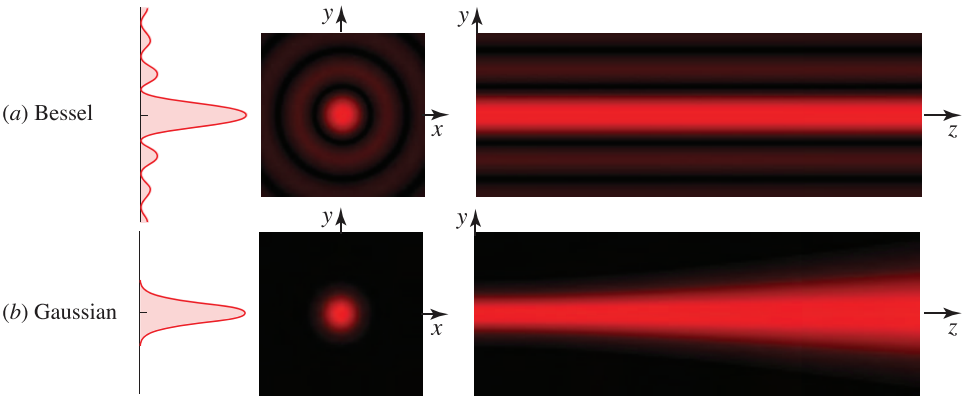


Figure 3.5-1 (a) The intensity distribution of the Bessel beam in the transverse plane is independent of z. The beam is nondiffracting and therefore does not diverge. (b) Transverse intensity distribution of a Gaussian beam for comparison with the Bessel beam. Parameters are selected such that the peak intensities and $1/e^2$ widths are identical in both cases.

It is useful to compare the Bessel beam with the Gaussian beam. Whereas the complex amplitude of the Bessel beam is an *exact* solution of the Helmholtz equation, the complex amplitude of the Gaussian beam is only an *approximate* solution thereof (since its complex envelope is an exact solution of the paraxial Helmholtz equation). The intensity distributions of these two beams are compared graphically in Fig. 3.5-1.

It is apparent that the asymptotic behavior of these distributions in the limit of large radial distances is significantly different. The intensity of the Gaussian beam decreases exponentially with ρ as $\exp[-2\rho^2/W^2(z)]$. The intensity of the Bessel beam, on the other hand, decreases as $J_0^2(k_T\rho) \approx (2/\pi k_T\rho)\cos^2(k_T\rho - \pi/4)$, which is an oscillatory function superimposed on a slow inverse-power-law decay with ρ. As a consequence, the transverse RMS width of the Gaussian beam, $\sigma = \frac{1}{2}W(z)$, is finite, while the transverse RMS width of the ideal Bessel beam is infinite for all z (see Appendix A, Sec. A.2 for the definition of RMS width). This is a manifestation of the tradeoff between beam size and divergence: the RMS width of the ideal Bessel beam is infinite and its divergence is zero, just as for the ideal plane wave.

As shown in Examples 2.4-1 and 4.3-5, the Bessel beam is associated with a continuum of plane waves whose directions form a cone of fixed half angle with respect to the propagation direction. It can be implemented by use of an axicon [Fig. 1.2-12(c)]. A derivation of the complex amplitude for the Bessel beam along with a general discussion of nondiffracting beams from the perspective of Fourier optics is provided in Sec. 4.3C.

A hybrid beam, called a **Bessel–Gaussian beam**, is a Bessel beam modulated by a Gaussian function of the radial coordinate ρ. The Gaussian serves as a window function that accelerates the slow radial decay of the Bessel beam (see Fig. 3.5-1). The Bessel–Gaussian beam can be generated by illuminating an axicon with a Gaussian beam.

*B. Airy Beams

In analogy with the Bessel beam, the Airy beam arises as a diffraction-free exact solution to the paraxial Helmholtz equation (2.2-23). Although the shape of its transverse

intensity distribution is maintained, the beam center is transversally displaced in an accelerated manner as it propagates along the axial direction,[†] as shown in Fig. 3.5-2.

The complex envelope of the Airy beam in one dimension is expressed as

$$A(x,z) = \mathtt{Ai}\left(\frac{x}{W_0} - \frac{z^2}{(4z_0)^2}\right) \exp\left(-j\frac{x}{W_0}\frac{z}{2z_0}\right) \exp\left(j\frac{z^3}{(24z_0)^3}\right), \qquad (3.5\text{-}5)$$

where $\mathtt{Ai}(x)$ is the Airy function, a special function that is the solution of the Airy differential equation $d^2y/dx^2 = xy$. The parameters W_0 and z_0 are, respectively, transverse and axial scaling factors that obey the relation $W_0^2 = \lambda z_0/\pi$, which also applies to the Gaussian beam [see (3.1-11)]. At $z = 0$ the transverse intensity of the Airy beam, $I(x,0) = \mathtt{Ai}^2(x/W_0)$, is distinctly asymmetrical as illustrated on the left-hand side of Fig. 3.5-2. At an arbitrary value of z, the intensity has the same transverse distribution except that it exhibits an axially dependent transverse shift $x = W_0 z^2/(4z_0)^2$ that follows a parabolic trajectory, $x = z^2/4a$ with $a = 4z_0^2/W_0$, thereby mimicking the path of a ballistic projectile. At $z = 4z_0$, for example, the transverse shift is W_0 while at $z = 20\,z_0$ it grows to $25\,W_0$, which provides the rationale for the appellation **accelerating beam**.

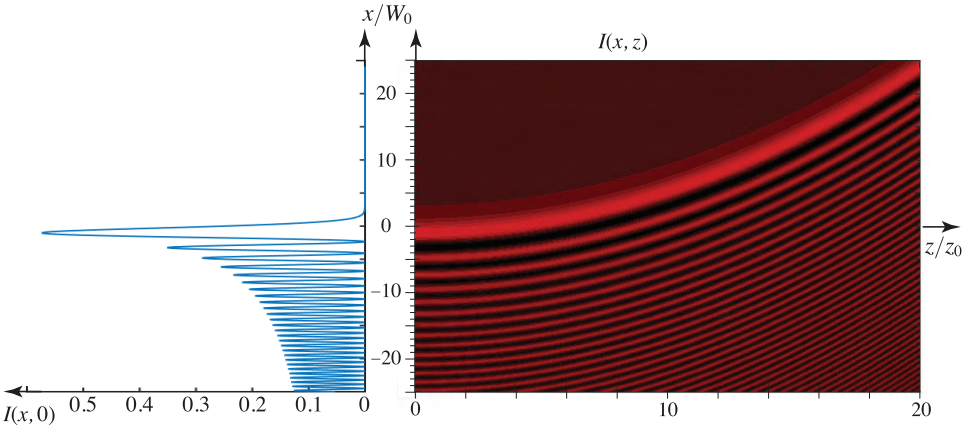


Figure 3.5-2 Transverse intensity distribution for the Airy beam $I(x,0)$ (left) and $I(x,z)$ (right).

The Airy beam may be generated by making use of an optical Fourier-transform system, as described in Prob. 4.2-6. Applications of the Airy beam include microscopy and prodding small particles along curved trajectories.

Other Bessel-like and Airy-like beams with main-lobe intensity distributions that remain nearly invariant and symmetrical as they travel can be engineered to propagate along arbitrary trajectories in free space (including 3D spirals).[‡] These nondiffracting beams, which can be partially obstructed and yet recover further down the beam axis (so-called "self-healing"), are useful for applications such as optical trapping and precision drilling.

[†] See G. A. Siviloglou, J. Broky, A. Dogariu, and D. N. Christodoulides, Observation of Accelerating Airy Beams, *Physical Review Letters*, vol. 99, 213901, 2007.

[‡] See J. Zhao, P. Zhang, D. Deng, J. Liu, Y. Gao, I. D. Chremmos, N. K. Efremidis, D. N. Christodoulides, and Z. Chen, Observation of Self-Accelerating Bessel-Like Optical Beams Along Arbitrary Trajectories, *Optics Letters*, vol. 38, pp. 498–500, 2013.

READING LIST

Books

See also the reading list on lasers in Chapter 16.

J. Secor, R. Alfano, and S. Ashrafi, *Complex Light*, IOP Publishing, 2017.

F. M. Dickey, ed., *Laser Beam Shaping: Theory and Techniques*, CRC Press/Taylor & Francis, 2nd ed. 2014.

A. N. Oraevsky, *Gaussian Beams and Optical Resonators*, Nova Science, 1996.

Articles

M. J. Padgett, Orbital Angular Momentum 25 Years On, *Optics Express*, vol. 25, pp. 11265–11274, 2017.

J. Zhao, P. Zhang, D. Deng, J. Liu, Y. Gao, I. D. Chremmos, N. K. Efremidis, D. N. Christodoulides, and Z. Chen, Observation of Self-Accelerating Bessel-Like Optical Beams Along Arbitrary Trajectories, *Optics Letters*, vol. 38, pp. 498–500, 2013.

A. Dudley, M. Lavery, M. Padgett, and A. Forbes, Unraveling Bessel Beams, *Optics & Photonics News*, vol. 24, no. 6, pp. 22–29, 2013.

A. M. Yao and M. J. Padgett, Orbital Angular Momentum: Origins, Behavior and Applications, *Advances in Optics and Photonics*, vol. 3, pp. 161–204, 2011.

M. R. Dennis, R. P. King, B. Jack, K. O'Holleran, and M. J. Padgett, Isolated Optical Vortex Knots, *Nature Physics*, vol. 6, pp. 118–121, 2010.

M. R. Dennis, K. O'Holleran, and M. J. Padgett, Singular Optics: Optical Vortices and Polarization Singularities, in E. Wolf, ed., *Progress in Optics*, Elsevier, 2009, vol. 53, pp. 293–363.

J. Vickers, M. Burch, R. Vyas, and S. Singh, Phase and Interference Properties of Optical Vortex Beams, *Journal of the Optical Society of America A*, vol. 25, pp. 823–827, 2008.

M. Martinelli and P. Martelli, Laguerre Mathematics in Optical Communications, *Optics & Photonics News*, vol. 19, no. 2, pp. 30–35, 2008.

G. A. Siviloglou, J. Broky, A. Dogariu, and D. N. Christodoulides, Observation of Accelerating Airy Beams, *Physical Review Letters*, vol. 99, 213901, 2007.

M. A. Bandres and J. C. Gutiérrez-Vega, Ince–Gaussian Modes of the Paraxial Wave Equation and Stable Resonators, *Journal of the Optical Society of America*, vol. 21, pp. 873–880, 2004.

F. Gori, G. Guattari, and C. Padovani, Bessel–Gauss Beams, *Optics Communications*, vol. 64, pp. 491–495, 1987.

J. Durnin, J. J. Miceli, Jr., and J. H. Eberly, Diffraction-Free Beams, *Physical Review Letters*, vol. 58, pp. 1499–1501, 1987.

Special issue on propagation and scattering of beam fields, *Journal of the Optical Society of America A*, vol. 3, no. 4, 1986.

H. Kogelnik and T. Li, Laser Beams and Resonators, *Proceedings of the IEEE*, vol. 54, pp. 1312–1329, 1966.

G. D. Boyd and J. P. Gordon, Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, *Bell System Technical Journal*, vol. 40, pp. 489–508, 1961.

A. G. Fox and T. Li, Resonant Modes in a Maser Interferometer, *Bell System Technical Journal*, vol. 40, pp. 453–488, 1961.

PROBLEMS

3.1-6 **Beam Parameters.** The light emitted from a Nd:YAG laser at a wavelength of 1.06 μm is a Gaussian beam of 1-W optical power and beam divergence $2\theta_0 = 1$ mrad. Determine the beam waist radius, the depth of focus, the maximum intensity, and the intensity on the beam axis at a distance $z = 100$ cm from the beam waist.

3.1-7 **Beam Identification by Two Widths.** A Gaussian beam of wavelength $\lambda_o = 10.6$ μm (emitted by a CO_2 laser) has widths $W_1 = 1.699$ mm and $W_2 = 3.380$ mm at two points separated by a distance $d = 10$ cm. Determine the location of the waist and the waist radius.

3.1-8 **The Elliptic Gaussian Beam.** The paraxial Helmholtz equation admits a Gaussian beam with intensity $I(x, y, 0) = |A_0|^2 \exp[-2(x^2/W_{0x}^2 + y^2/W_{0y}^2)]$ in the $z = 0$ plane, with the beam waist radii W_{0x} and W_{0y} in the x and y directions, respectively. The contours of constant intensity are therefore ellipses instead of circles. Write expressions for the beam depth of focus, angular divergence, and radii of curvature in the x and y directions, as functions of W_{0x}, W_{0y}, and the wavelength λ. If $W_{0x} = 2W_{0y}$, sketch the shape of the beam spot in the $z = 0$ plane and in the far field (z much greater than the depths of focus in both transverse directions).

3.2-6 **Beam Focusing.** An argon-ion laser produces a Gaussian beam of wavelength $\lambda = 488$ nm with waist radius $W_0 = 0.5$ mm. Design a single-lens optical system for focusing the light to a spot of diameter 100 μm. What is the shortest focal-length lens that may be used?

3.2-7 **Spot Size.** A Gaussian beam of Rayleigh range $z_0 = 50$ cm and wavelength $\lambda = 488$ nm is converted into a Gaussian beam of waist radius W_0' using a lens of focal length $f = 5$ cm at a distance z from its waist, as illustrated in Fig. 3.2-2. Plot W_0' as a function of z. Verify that in the limit $z - f \gg z_0$, (3.2-10) and (3.2-12) hold; and that in the limit $z \ll z_0$, (3.2-13) holds.

3.2-8 **Beam Refraction.** A Gaussian beam is incident from air ($n = 1$) into a medium with a planar boundary and refractive index $n = 1.5$. The beam axis is normal to the boundary and the beam waist lies at the boundary. Sketch the transmitted beam. If the angular divergence of the beam in air is 1 mrad, what is the angular divergence in the medium?

*3.2-9 **Transmission of a Gaussian Beam Through a Graded-Index Slab.** The $ABCD$ matrix of a SELFOC graded-index slab with quadratic refractive index $n(y) \approx n_0(1 - \frac{1}{2}\alpha^2 y^2)$ (see Sec. 1.3B) and length d is $A = \cos\alpha d$, $B = (1/\alpha)\sin\alpha d$, $C = -\alpha \sin\alpha d$, $D = \cos\alpha d$ for paraxial rays along the z direction. A Gaussian beam of wavelength λ_o, waist radius W_0 in free space, and axis in the z direction enters the slab at its waist. Use the $ABCD$ law to determine an expression for the beam width in the y direction as a function of d. Sketch the shape of the beam as it travels through the medium.

3.3-2 **Power Confinement in Hermite–Gaussian Beams.** Determine the ratio of the power contained within a circle of radius $W(z)$ in the transverse plane, to the total power, for the Hermite–Gaussian beams HG_{00}, HG_{01}, HG_{10}, and HG_{11}. What is the ratio of the power contained within a circle of radius $\frac{1}{10}W(z)$ to the total power for the HG_{00} and HG_{11} beams?

3.3-3 **The Donut Beam.** Consider a wave that is a superposition of two Hermite–Gaussian beams, HG_{01} and HG_{10}, with equal intensities. The two beams have independent and random phases so that their intensities add with no interference. Show that the total intensity is described by a donut-shaped (toroidal) circularly symmetric function. Assuming that $W_0 = 1$ mm, determine the radius of the circle of peak intensity and the radii of the two circles of $1/e^2$ times the peak intensity at the beam waist.

3.3-4 **Axial Phase.** Consider the Hermite–Gaussian beams of all orders, HG_{lm}, with Rayleigh range $z_0 = 30$ cm in a medium of refractive index $n = 1$. Determine the frequencies within the band $\nu = 10^{14} \pm 2 \times 10^9$ Hz for which the phase retardation between the planes $z = -z_0$ and $z = z_0$ is an integer multiple of π along the beam axis. These frequencies are the modes of a resonator comprising two spherical mirrors placed at the $z = \pm z_0$ planes, as described in Sec. 11.2D.

CHAPTER

4

FOURIER OPTICS

4.1 PROPAGATION OF LIGHT IN FREE SPACE 113
- A. Spatial Harmonic Functions and Plane Waves
- B. Transfer Function of Free Space
- C. Impulse Response Function of Free Space
- D. Huygens–Fresnel Principle

4.2 OPTICAL FOURIER TRANSFORM 124
- A. Fourier Transform in the Far Field
- B. Fourier Transform Using a Lens

4.3 DIFFRACTION OF LIGHT 129
- A. Fraunhofer Diffraction
- *B. Fresnel Diffraction
- *C. Nondiffracting Waves

4.4 IMAGE FORMATION 137
- A. Ray Optics of a Single-Lens Imaging System
- B. Wave Optics of a 4-*f* Imaging System
- C. Wave Optics of a Single-Lens Imaging System
- D. Near-Field Imaging

4.5 HOLOGRAPHY 147

Josef von Fraunhofer (1787–1826) developed the diffraction grating and contributed to our understanding of diffraction. His epitaph reads *Approximavit sidera* (he brought the stars closer).

Jean-Baptiste Joseph Fourier (1768–1830) demonstrated that periodic functions could be constructed from sums of sinusoids. Harmonic analysis is the basis of Fourier optics; it has many applications.

Dennis Gabor (1900–1979) invented holography and contributed to its development. He made the first hologram in 1947 and received the Nobel Prize in 1971 for carrying out this body of work.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Fourier optics provides a description of the propagation of light waves based on harmonic analysis (the Fourier transform) and linear systems. The methods of harmonic analysis have proved useful for describing signals and systems in many disciplines. Harmonic analysis is based on the expansion of an arbitrary function of time $f(t)$ in terms of a superposition (a sum or integral) of harmonic functions of time of different frequencies (see Appendix A, Sec. A.1). The harmonic function $F(\nu)\exp(j2\pi\nu t)$, which has frequency ν and complex amplitude $F(\nu)$, is the building block of the theory. Several of these functions, each with its own amplitude $F(\nu)$, are added to construct the function $f(t)$, as illustrated in Fig. 4.0-1. The complex amplitude $F(\nu)$, as a function of frequency, is called the Fourier transform of $f(t)$. This approach is highly useful for analyzing linear systems (see Appendix B, Sec. B.1). If the response of the system to each harmonic function is known, the response to an arbitrary input function is readily determined by the use of harmonic analysis at the input of the system and superposition at the output.

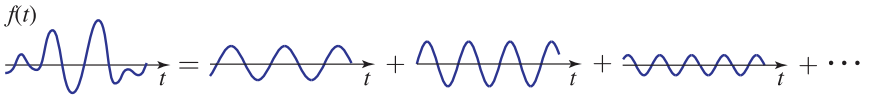


Figure 4.0-1 An arbitrary function $f(t)$ may be analyzed as a sum of harmonic functions of different frequencies and complex amplitudes.

An arbitrary complex function $f(x, y)$ of two variables that represent spatial coordinates in a plane, say x and y, may similarly be written as a superposition of harmonic functions of x and y, each of the form $F(\nu_x, \nu_y)\exp[-j2\pi(\nu_x x + \nu_y y)]$, where $F(\nu_x, \nu_y)$ is the complex amplitude and ν_x and ν_y are the **spatial frequencies** (cycles per unit length; typically cycles/mm) in the x and y directions, respectively.[†] The harmonic function $F(\nu_x, \nu_y)\exp[-j2\pi(\nu_x x + \nu_y y)]$ is the two-dimensional building block of the theory. It can be used to generate an arbitrary function of two variables $f(x, y)$, as depicted in Fig. 4.0-2 and explained in Appendix A, Sec. A.3.

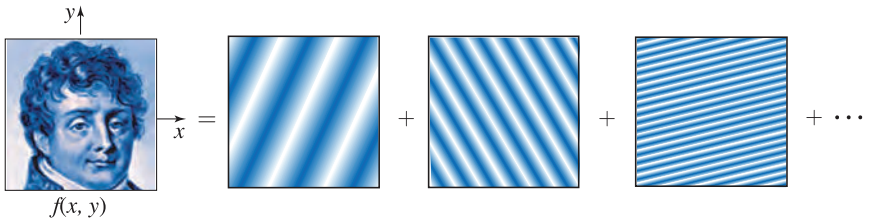


Figure 4.0-2 An arbitrary function $f(x, y)$ may be analyzed in terms of a sum of harmonic functions of different spatial frequencies and complex amplitudes, drawn here schematically as graded blue lines.

The monochromatic plane wave $U(x, y, z) = A\exp[-j(k_x x + k_y y + k_z z)]$ plays an important role in wave optics. The coefficients (k_x, k_y, k_z) are the components of the wavevector $\mathbf{k}$, and A is a complex constant. $U(x, y, z)$ reduces to a spatial harmonic function of the points in an arbitrary plane. At the $z = 0$ plane, for example, $U(x, y, 0)$ becomes the harmonic function $f(x, y) = A\exp[-j2\pi(\nu_x x + \nu_y y)]$,

[†] The spatial harmonic function is defined with a minus sign in the exponent, in contrast to the plus sign used in the definition of the temporal harmonic function (compare (A.1-1) and (A.3-1) in Appendix A). This sign convention is chosen to match that for the forward-traveling plane wave set forth in (2.2-11).

where $\nu_x = k_x/2\pi$ and $\nu_y = k_y/2\pi$ are the spatial frequencies (cycles/mm), and k_x and k_y are the spatial angular frequencies (radians/mm). There is a one-to-one correspondence between the plane wave $U(x, y, z)$ and the spatial harmonic function $f(x, y) = U(x, y, 0)$ since knowledge of k_x and k_y is sufficient to determine k_z via the relation $k_x^2 + k_y^2 + k_z^2 = \omega^2/c^2 = (2\pi/\lambda)^2$. As will be explained subsequently, k_x and k_y may not exceed ω/c under usual circumstances; i.e., the spatial frequencies ν_x and ν_y may not exceed the inverse wavelength $1/\lambda$.

Since an arbitrary function $f(x, y)$ can be analyzed as a superposition of harmonic functions, an arbitrary traveling wave $U(x, y, z)$ may be analyzed in terms of a sum of plane waves (Fig. 4.0-3). The plane wave is thus the building block used to construct a wave of arbitrary complexity. Furthermore, if it can be determined how a linear optical system modifies plane waves, the principle of superposition can be used to establish the effect of the system on an arbitrary wave.

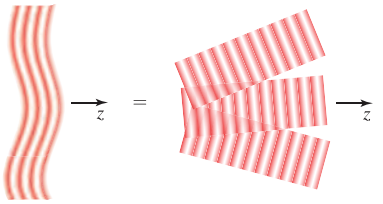


Figure 4.0-3 The principle of Fourier optics: An arbitrary wave in free space can be analyzed in terms of a superposition of plane waves.

Because of the important role that Fourier analysis plays in describing linear systems, it is useful to consider the propagation of light through linear optical components, including free space, in terms of a linear-systems approach. The complex amplitudes at two planes normal to the optic (z) axis are regarded as the input and output of the system (Fig. 4.0-4). A linear system may be characterized by either its **impulse response function**, which is the response of the system to a point (i.e., an impulse) at its input, or by its **transfer function**, which is the response of the system to a set of spatial harmonic functions (as described in Appendix B).

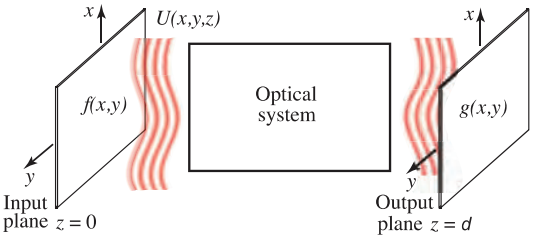


Figure 4.0-4 The transmission of an optical wave $U(x, y, z)$ through an optical system located between an input plane $z = 0$ and an output plane $z = d$. This configuration is regarded as a linear system whose input and output are the functions of $f(x, y) = U(x, y, 0)$ and $g(x, y) = U(x, y, d)$, respectively.

This Chapter

The chapter begins with a Fourier description of the propagation of monochromatic light in free space (Sec. 4.1). The transfer function and impulse response function of the free-space propagation system are determined. In Sec. 4.2 we show that a lens may be used to carry out the spatial Fourier-transform operation. The transmission of light through apertures is discussed in Sec. 4.3; this section comprises a Fourier-optics approach to the diffraction of light, a subject usually presented in introductory textbooks from the perspective of the Huygens principle. Section 4.4 is devoted to image formation and spatial filtering in the context of both ray and wave optics. Subwavelength imaging, in the form of near-field optical microscopy, is also considered. Finally, an introduction to holography, the recording and reconstruction of optical waves, is presented in Sec. 4.5. It is important to understand the basic properties of Fourier transforms and linear systems in one and two dimensions (as reviewed in Appendices A and B, respectively) to follow this chapter.

4.1 PROPAGATION OF LIGHT IN FREE SPACE

A. Spatial Harmonic Functions and Plane Waves

A *monochromatic* plane wave of complex amplitude $U(x,y,z) = A\exp[-j(k_x x + k_y y + k_z z)]$ has wavevector $\mathbf{k} = (k_x, k_y, k_z)$, wavelength λ, wavenumber $k = \sqrt{k_x^2 + k_y^2 + k_z^2} = 2\pi/\lambda$, and complex envelope A. The vector $\mathbf{k}$ makes angles $\theta_x = \sin^{-1}(k_x/k)$ and $\theta_y = \sin^{-1}(k_y/k)$ with the y–z and x–z planes, respectively, as illustrated in Fig. 4.1-1. Thus, if $\theta_x = 0$, there is no component of $\mathbf{k}$ in the x direction. The complex amplitude at the $z = 0$ plane, $U(x,y,0)$, is a spatial harmonic function $f(x,y) = A\exp[-j2\pi(\nu_x x + \nu_y y)]$ with spatial frequencies $\nu_x = k_x/2\pi$ and $\nu_y = k_y/2\pi$. The angles of the wavevector are therefore related to the spatial frequencies of the harmonic function by

$$\theta_x = \sin^{-1}\lambda\nu_x, \quad \theta_y = \sin^{-1}\lambda\nu_y. \tag{4.1-1}$$

Spatial Frequencies and Angles

The spatial frequency $\nu = k/2\pi$ is specified in cycles/mm, whereas the optical frequency $\nu = kc/2\pi = c/\lambda$ is specified in cycles/sec or Hz, as shown in Sec. 2.2.

Recognizing $\Lambda_x = 1/\nu_x$ and $\Lambda_y = 1/\nu_y$ as the periods of the harmonic functions in the x and y directions (mm/cycle), we see that the angles $\theta_x = \sin^{-1}(\lambda/\Lambda_x)$ and $\theta_y = \sin^{-1}(\lambda/\Lambda_y)$ are governed by the ratios of the wavelength of light to the period of the harmonic function in each direction. These geometrical relations follow from matching the wavefronts of the wave to the periodic pattern of the harmonic function in the $z = 0$ plane, as illustrated in Fig. 4.1-1.

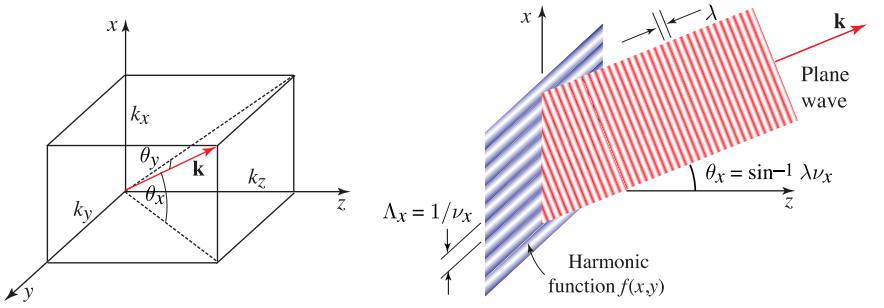


Figure 4.1-1 A harmonic function of spatial frequencies ν_x and ν_y at the plane $z = 0$ is consistent with a plane wave traveling at angles $\theta_x = \sin^{-1}\lambda\nu_x$ and $\theta_y = \sin^{-1}\lambda\nu_y$.

If $k_x \ll k$ and $k_y \ll k$, so that the wavevector $\mathbf{k}$ is paraxial, the angles θ_x and θ_y are small ($\sin\theta_x \approx \theta_x$ and $\sin\theta_y \approx \theta_y$) and

$$\theta_x \approx \lambda\nu_x, \quad \theta_y \approx \lambda\nu_y. \tag{4.1-2}$$

Spatial Frequencies and Angles (Paraxial Approximation)

The angles of inclination of the wavevector are then directly proportional to the spatial frequencies of the corresponding harmonic function. Apparently, there is a one-to-one

correspondence between the plane wave $U(x, y, z)$ and the harmonic function $f(x, y)$. Given one, the other can be readily determined, provided the wavelength λ is known: the harmonic function $f(x, y)$ is obtained by sampling at the $z = 0$ plane, $f(x, y) = U(x, y, 0)$. Given the harmonic function $f(x, y)$, on the other hand, the wave $U(x, y, z)$ is constructed by using the relation $U(x, y, z) = f(x, y) \exp(-jk_z z)$ with

$$k_z = \pm\sqrt{k^2 - k_x^2 - k_y^2}, \qquad k = 2\pi/\lambda. \tag{4.1-3}$$

A condition for the validity of this correspondence is that $k_x^2 + k_y^2 < k^2$, so that k_z is real. This condition implies that $\lambda\nu_x < 1$ and $\lambda\nu_y < 1$, so that the angles θ_x and θ_y defined by (4.1-1) exist. The $+$ and $-$ signs in (4.1-3) represent waves traveling in the forward and backward directions, respectively. We shall be concerned with forward waves only.

Spatial Spectral Analysis

When a plane wave of unity amplitude traveling in the z direction is transmitted through a thin optical element with complex amplitude transmittance $f(x, y) = \exp[-j2\pi(\nu_x x + \nu_y y)]$ the wave is modulated by the harmonic function, so that $U(x, y, 0) = f(x, y)$. The incident wave is then converted into a plane wave with a wavevector at angles $\theta_x = \sin^{-1} \lambda\nu_x$ and $\theta_y = \sin^{-1} \lambda\nu_y$ (see Fig. 4.1-2). The element thus acts much as a prism, bending the wave upward in this illustration. If the complex amplitude transmittance is $f(x, y) = \exp[+j2\pi(\nu_x x + \nu_y y)]$, the wave is converted into a plane wave whose wavevector makes angles $-\theta_x$ and $-\theta_y$ with the z axis, so the wave is bent downward instead.

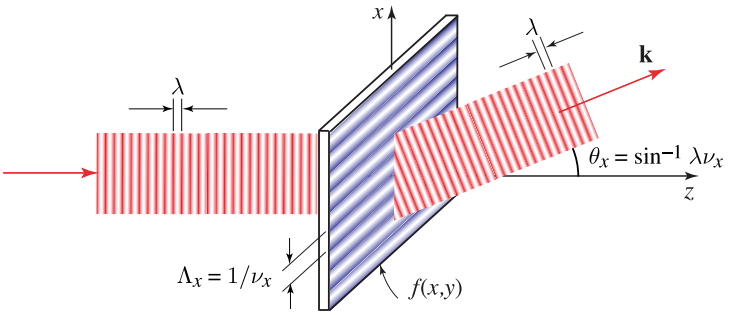


Figure 4.1-2 A thin element whose complex amplitude transmittance is a harmonic function of spatial frequency ν_x (period $\Lambda_x = 1/\nu_x$) bends a plane wave of wavelength λ by an angle $\theta_x = \sin^{-1} \lambda\nu_x = \sin^{-1}(\lambda/\Lambda_x)$. The dark blue and white stripes are used to indicate that the element is a *phase* grating (changing only the phase of the wave).

The wave-deflection property of an optical element with harmonic-function transmittance may be understood as an interference phenomenon. In a direction making an angle θ_x, two points on the element separated by a the period $\Lambda = 1/\nu_x$, have a relative pathlength difference of $\Lambda \sin\theta_x = (1/\nu_x)\lambda\nu_x = \lambda$, i.e., equal to a wavelength. Hence, all segments separated by a period interfere constructively in this direction.

If the transmittance of the optical element $f(x, y)$ is the sum of several harmonic functions of different spatial frequencies, the transmitted optical wave is also the sum of an equal number of plane waves dispersed into different directions; each spatial frequency is mapped into a corresponding direction, in accordance with (4.1-1). The amplitude of each wave is proportional to the amplitude of the corresponding harmonic component of $f(x, y)$.

Examples.

- A complex amplitude transmittance of the form $f(x,y) = \cos(2\pi\nu_x x) = \frac{1}{2}\{\exp(-j2\pi\nu_x x) + \exp(+j2\pi\nu_x x)\}$ bends an incident plane wave into components traveling at angles $\pm\sin^{-1}(\lambda\nu_x)$, namely in both the upward and downward directions.
- An element with a transmittance that varies as $1 + \cos(2\pi\nu_y y)$ behaves as a diffraction grating (see Exercise 2.4-5); the incident wave is bent into components that travel to the right and left, while a portion travels straight through.
- An element with transmittance $f(x,y) = \mathcal{U}[\cos(2\pi\nu_x x)]$, where $\mathcal{U}(x)$ is the unit step function [$\mathcal{U}(x) = 1$ if $x > 0$, and $\mathcal{U}(x) = 0$ if $x < 0$], represents a periodic set of slits with $f(x,y) = 1$ set in an opaque screen [$f(x,y) = 0$]. This periodic function may be analyzed via a Fourier series as a sum of harmonic functions of spatial frequencies $0, \pm\nu_x, \pm 2\nu_x, \ldots$, corresponding to waves traveling at angles $0, \pm\sin^{-1}\lambda\nu_x, \pm\sin^{-1} 2\lambda\nu_x, \ldots$, with amplitudes proportional to the coefficients of the Fourier series (in the case at hand, these vanish for even harmonics). At these angles, the waves transmitted through the slits interfere constructively.

More generally, if $f(x,y)$ is a superposition integral of harmonic functions,

$$f(x,y) = \iint_{-\infty}^{\infty} F(\nu_x,\nu_y)\exp\left[-j2\pi(\nu_x x + \nu_y y)\right] d\nu_x\, d\nu_y, \tag{4.1-4}$$

with frequencies (ν_x,ν_y) and amplitudes $F(\nu_x,\nu_y)$, the transmitted wave $U(x,y,z)$ is the superposition of plane waves,

$$U(x,y,z) = \iint_{-\infty}^{\infty} F(\nu_x,\nu_y)\exp\left[-j(2\pi\nu_x x + 2\pi\nu_y y)\right]\exp(-jk_z z)\, d\nu_x\, d\nu_y, \tag{4.1-5}$$

with complex envelopes $F(\nu_x,\nu_y)$ where $k_z = \sqrt{k^2 - k_x^2 - k_y^2} = 2\pi\sqrt{\lambda^{-2} - \nu_x^2 - \nu_y^2}$. Note that $F(\nu_x,\nu_y)$ is the Fourier transform of $f(x,y)$ [see (A.3-2) in Appendix A].

Since an arbitrary function may be Fourier analyzed as a superposition integral of the form (4.1-4), the light transmitted through a thin optical element of arbitrary transmittance may be written as a superposition of plane waves (see Fig. 4.1-3), provided that $\nu_x^2 + \nu_y^2 < \lambda^{-2}$.

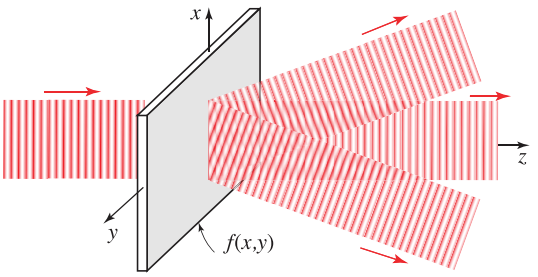


Figure 4.1-3 A thin optical element of amplitude transmittance $f(x,y)$ decomposes an incident plane wave into many plane waves. The plane wave traveling at the angles $\theta_x = \sin^{-1}\lambda\nu_x$ and $\theta_y = \sin^{-1}\lambda\nu_y$ has a complex envelope $F(\nu_x,\nu_y)$, the Fourier transform of $f(x,y)$.

This process of "spatial spectral analysis" is akin to the angular dispersion of different temporal-frequency components (wavelengths) provided by a prism. Free-space

propagation serves as a natural "spatial prism," sensitive to the spatial rather than the temporal frequencies of the optical wave.

Amplitude Modulation

Consider a transparency with complex amplitude transmittance $f_0(x, y)$. If the Fourier transform $F_0(\nu_x, \nu_y)$ extends over widths $\pm\Delta\nu_x$ and $\pm\Delta\nu_y$ in the x and y directions, the transparency will deflect an incident plane wave by angles θ_x and θ_y in the range $\pm \sin^{-1}(\lambda\Delta\nu_x)$ and $\pm \sin^{-1}(\lambda\Delta\nu_y)$, respectively.

Consider a second transparency of complex amplitude transmittance $f(x, y) = f_0(x, y) \exp[-j2\pi(\nu_{x0}x + \nu_{y0}y)]$, where $f_0(x, y)$ is slowly varying compared to $\exp[-j2\pi(\nu_{x0}x + \nu_{y0}y)]$ so that $\Delta\nu_x \ll \nu_{x0}$ and $\Delta\nu_y \ll \nu_{y0}$. We may regard $f(x, y)$ as an amplitude-modulated function with a carrier frequency ν_{x0} and ν_{y0} and modulation function $f_0(x, y)$. The Fourier transform of $f(x, y)$ is $F_0(\nu_x - \nu_{x0}, \nu_y - \nu_{y0})$, in accordance with the frequency-shifting property of the Fourier transform (see Appendix A). The transparency will deflect a plane wave to directions centered about the angles $\theta_{x0} = \sin^{-1} \lambda\nu_{x0}$ and $\theta_{y0} = \sin^{-1} \lambda\nu_{y0}$ (Fig. 4.1-4). This can also be readily seen by regarding $f(x, y)$ as a transparency of transmittance $f_0(x, y)$ in contact with a grating or prism of transmittance $\exp[-j2\pi(\nu_{x0}x + \nu_{y0}y)]$ that provides the angular deflection θ_{x0} and θ_{y0}.

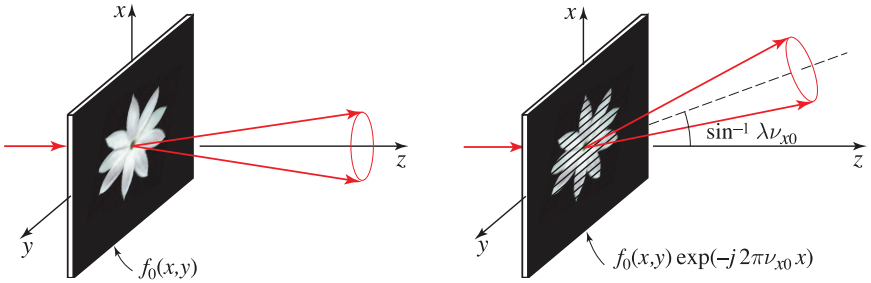


Figure 4.1-4 Deflection of light by the transparencies $f_0(x, y)$ and $f_0(x, y) \exp(-j2\pi\nu_{x0}x)$. The "carrier" harmonic function $\exp(-j2\pi\nu_{x0}x)$ acts as a prism that deflects the wave by an angle $\theta_{x0} = \sin^{-1} \lambda\nu_{x0}$.

This idea may be used to record two images $f_1(x, y)$ and $f_2(x, y)$ on the same transparency using the *spatial-frequency multiplexing* scheme $f(x, y) = f_1(x, y) \exp[-j2\pi(\nu_{x1}x + \nu_{y1}y)] + f_2(x, y) \exp[-j2\pi(\nu_{x2}x + \nu_{y2}y)]$. The two images may be easily separated by illuminating the transparency with a plane wave, whereupon the two images are deflected at different angles and are thus separated. This principle will prove useful in holography (Sec. 4.5), where it is often desired to separate two images recorded on the same transparency.

Frequency Modulation

The foregoing examples relate to the transmittance of plane waves through transparencies endowed with one or more 2D harmonic functions that extend over the entire region of the transparency. We now examine the transmission of a plane wave through a transparency comprising a "collage" of several regions, the transmittance of each of which is a harmonic function of some spatial frequency, as illustrated in Fig. 4.1-5. If the dimensions of each region are much greater than the period, each region acts as a grating or prism that deflects the wave in a particular direction, so that different portions of the incident wavefront are deflected into different directions. This principle may be used to create maps of optical interconnections, as described in Sec. 24.1A.

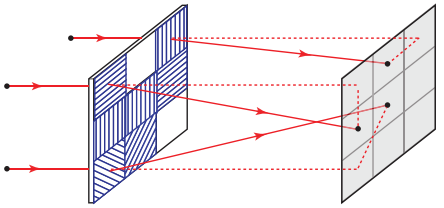


Figure 4.1-5 Deflection of light by a transparency made of several harmonic functions (phase gratings) of different spatial frequencies.

A transparency may also have a harmonic transmittance with a spatial frequency that varies continuously and slowly with position (in comparison with λ), much as the frequency of a frequency-modulated (FM) signal varies slowly with time. Consider, for example, the phase function $f(x,y) = \exp[-j2\pi\varphi(x,y)]$, where $\varphi(x,y)$ is a continuous slowly varying function of x and y. In the neighborhood of a point (x_0, y_0), we may use the Taylor-series expansion $\varphi(x,y) \approx \varphi(x_0,y_0)+(x-x_0)\nu_x+(y-y_0)\nu_y$, where the derivatives $\nu_x = \partial\varphi/\partial x$ and $\nu_y = \partial\varphi/\partial y$ are evaluated at the position (x_0, y_0). The local variation of $f(x,y)$ with x and y is therefore proportional to the quantity $\exp[-j2\pi(\nu_x x+\nu_y y)]$, which is a harmonic function with spatial frequencies $\nu_x = \partial\varphi/\partial x$ and $\nu_y = \partial\varphi/\partial y$. Since these derivatives vary with x and y, so do the spatial frequencies. The transparency $f(x,y) = \exp[-j2\pi\varphi(x,y)]$ therefore deflects the portion of the wave at the position (x,y) by the position-dependent angles $\theta_x = \sin^{-1}(\lambda\partial\varphi/\partial x)$ and $\theta_y = \sin^{-1}(\lambda\partial\varphi/\partial y)$.

EXAMPLE 4.1-1. ***Scanning.*** A thin transparency with complex amplitude transmittance $f(x,y) = \exp(j\pi x^2/\lambda f)$ introduces a phase shift $2\pi\varphi(x,y)$ where $\varphi(x,y) = -x^2/2\lambda f$, so that the wave is deflected at the position (x,y) by the angles $\theta_x = \sin^{-1}(\lambda\partial\varphi/\partial x) = \sin^{-1}(-x/f)$ and $\theta_y = 0$. If $|x/f| \ll 1$, $\theta_x \approx -x/f$ and the deflection angle θ_x is directly proportional to the transverse distance x. This transparency may be used to deflect a narrow beam of light. Moreover, if the transparency is moved at a uniform speed, the beam is deflected by a linearly increasing angle as illustrated in Fig. 4.1-6.

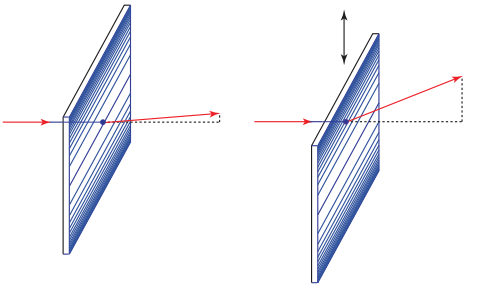


Figure 4.1-6 Making use of a frequency-modulated transparency to scan an optical beam.

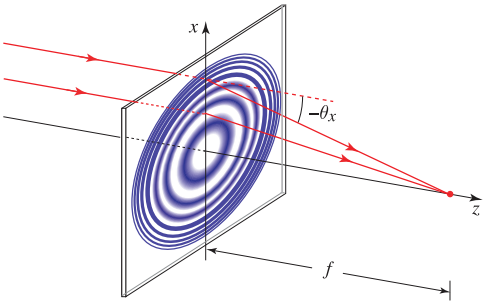


Figure 4.1-7 A transparency with transmittance $f(x,y) = \exp[j\pi(x^2+y^2)/\lambda f]$ acts as a spherical lens with focal length f.

EXAMPLE 4.1-2. ***Imaging.*** If the transparency illustrated in Example 4.1-1 is illuminated by a plane wave, each strip of the wave at a given value of x is deflected by a different angle and as a result the wavefront is altered. The local wavevector at position x bends by an angle $-x/f$ so that all wavevectors meet at a single line on the optical axis a distance f from the transparency. The transparency then acts as a cylindrical lens with a focal length f. Similarly, a transparency with transmittance $f(x,y) = \exp[j\pi(x^2+y^2)/\lambda f]$ acts as a spherical lens with focal length f, as illustrated in Fig. 4.1-7. Indeed, this is the expression for the amplitude transmittance of a thin lens provided in (2.4-9).

EXERCISE 4.1-1

Binary-Plate Cylindrical Lens. Use harmonic analysis near the position x to show that a transparency with complex amplitude transmittance equal to the binary function

$$f(x, y) = \mathcal{U}\left[\cos\left(\pi \frac{x^2}{\lambda f}\right)\right], \tag{4.1-6}$$

where $\mathcal{U}(x)$ is the unit step function [$\mathcal{U}(x) = 1$ if $x \geq 0$, and $\mathcal{U}(x) = 0$ if $x < 0$], acts as a cylindrical lens with multiple focal lengths equal to $\infty, \pm f, \pm f/3, \pm f/5, \ldots$.

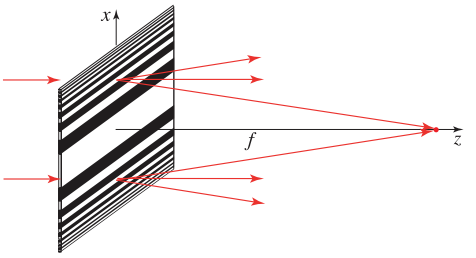


Figure 4.1-8 Binary plate as a cylindrical lens with multiple foci.

Fresnel Zone Plate

A two-dimensional generalization of the binary plate in Exercise 4.1-1 is a circularly symmetric transparency of complex amplitude transmittance

$$f(x, y) = \mathcal{U}\left[\cos\left(\pi \frac{x^2 + y^2}{\lambda f}\right)\right], \tag{4.1-7}$$

known as the **Fresnel zone plate**. It is a set of ring apertures of increasing radii, decreasing widths, and equal areas (see Fig. 4.1-9). The structure serves as a spherical lens with multiple focal lengths. A ray incident at each point is split into multiple rays, and the transmitted rays meet at multiple foci with focal lengths $\infty, \pm f, \pm f/3, \pm f/5, \ldots$, together with a component transmitted without deflection.

The operation of the Fresnel zone plate may also be described in terms of interference (see Sec. 2.5B). The center of the mth ring has a radius ρ_m at the mth peak of the cosine function, i.e., $\pi\rho_m^2/\lambda f = m2\pi$. At a focal point $z = f$, the distance R_m to the mth ring is given by $R_m^2 = f^2 + \rho_m^2$, so that $R_m = \sqrt{f^2 + 2m\lambda f}$. If f is sufficiently large so that the angles subtended by the rings are small, then $R_m \approx f + m\lambda$. Thus, the waves transmitted through consecutive rings have pathlengths differing by a wavelength, so that they interfere constructively at the focal point. A similar argument applies for the other foci.

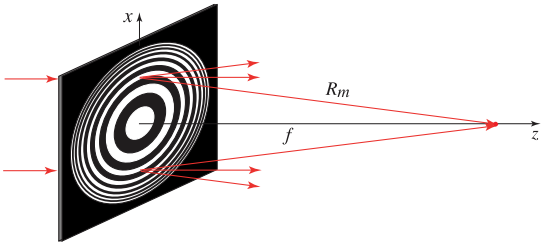


Figure 4.1-9 The Fresnel zone plate.

B. Transfer Function of Free Space

We now examine the propagation of a monochromatic optical wave of wavelength λ and complex amplitude $U(x, y, z)$ in the free space between the planes $z = 0$ and $z = d$, called the input and output planes, respectively (see Fig. 4.1-10). Given the complex amplitude of the wave at the input plane, $f(x, y) = U(x, y, 0)$, we shall determine the complex amplitude at the output plane, $g(x, y) = U(x, y, d)$.

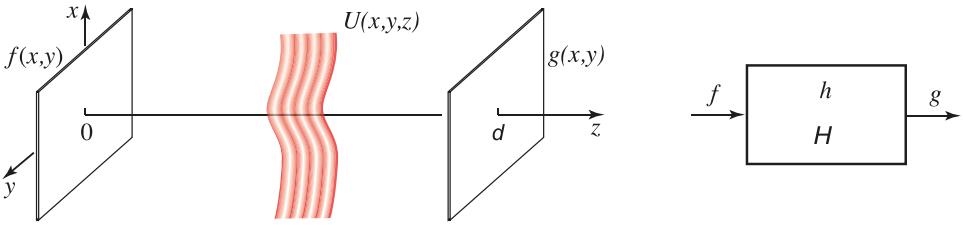


Figure 4.1-10 Propagation of light between two planes is regarded as a linear system whose input and output are the complex amplitudes of the wave in the two planes.

We regard $f(x, y)$ and $g(x, y)$ as the input and output of a linear system. The system is linear since the Helmholtz equation, which $U(x, y, z)$ must satisfy, is linear. The system is shift-invariant because of the invariance of free space to displacement of the coordinate system. A linear shift-invariant system is characterized by its impulse response function $h(x, y)$ or by its transfer function $H(\nu_x, \nu_y)$, as explained in Appendix B, Sec. B.2. We now proceed to determine expressions for these functions.

The transfer function $H(\nu_x, \nu_y)$ is the factor by which an input spatial harmonic function of frequencies ν_x and ν_y is multiplied to yield the output harmonic function. We therefore consider a harmonic input function $f(x, y) = A \exp[-j2\pi(\nu_x x + \nu_y y)]$. As explained earlier, this corresponds to a plane wave $U(x, y, z) = A \exp[-j(k_x x + k_y y + k_z z)]$ where $k_x = 2\pi\nu_x$, $k_y = 2\pi\nu_y$, and

$$k_z = \sqrt{k^2 - k_x^2 - k_y^2} = 2\pi\sqrt{\lambda^{-2} - \nu_x^2 - \nu_y^2}\,. \tag{4.1-8}$$

The output $g(x, y) = A \exp[-j(k_x x + k_y y + k_z d)]$, so that we can write $H(\nu_x, \nu_y) = g(x, y)/f(x, y) = \exp(-jk_z d)$, from which

$$\boxed{H(\nu_x, \nu_y) = \exp\left(-j2\pi d\sqrt{\lambda^{-2} - \nu_x^2 - \nu_y^2}\right)\,.} \tag{4.1-9}$$

Transfer Function of Free Space

The transfer function $H(\nu_x, \nu_y)$ is therefore a circularly symmetric complex function of the spatial frequencies ν_x and ν_y. Its magnitude and phase are sketched in Fig. 4.1-11.

For spatial frequencies for which $\nu_x^2 + \nu_y^2 \le \lambda^{-2}$ (i.e., frequencies lying within a circle of radius $1/\lambda$) the magnitude $|H(\nu_x, \nu_y)| = 1$ and the phase $\arg\{H(\nu_x, \nu_y)\}$ is a function of ν_x and ν_y. A harmonic function with such frequencies therefore undergoes a spatial phase shift as it propagates, but its magnitude is not altered.

At higher spatial frequencies, $\nu_x^2 + \nu_y^2 > \lambda^{-2}$, the quantity under the square root in (4.1-9) is negative so that the exponent is real and the transfer function $\exp[-2\pi d(\nu_x^2 + \nu_y^2 - \lambda^{-2})^{1/2}]$ represents an attenuation factor; the wave is then called an **evanescent**

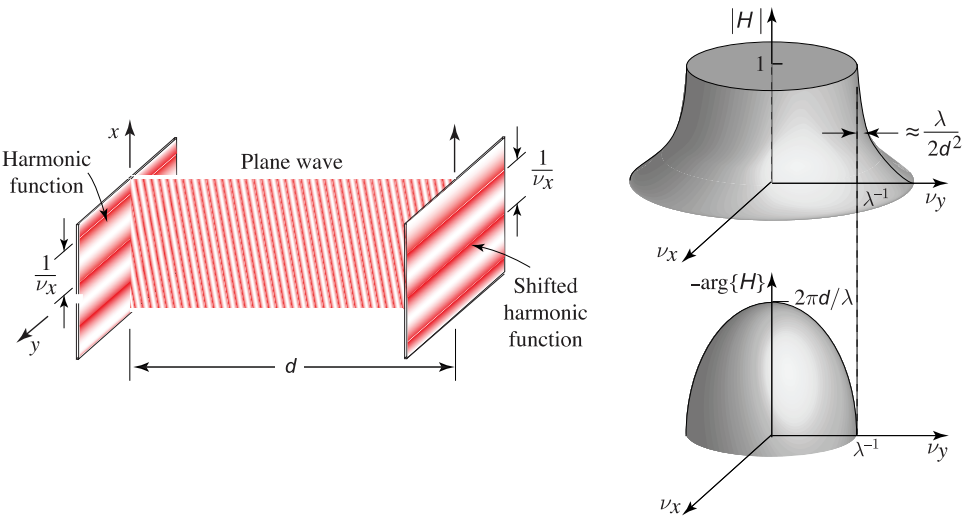


Figure 4.1-11 Magnitude and phase of the transfer function $H(\nu_x, \nu_y)$ for free-space propagation between two planes separated by a distance d.

wave.† When $\nu_\rho = (\nu_x^2 + \nu_y^2)^{1/2}$ exceeds λ^{-1} slightly, i.e., $\nu_\rho \approx \lambda^{-1}$, the attenuation factor is $\exp[-2\pi d(\nu_\rho^2 - \lambda^{-2})^{1/2}] = \exp[-2\pi d(\nu_\rho - \lambda^{-1})^{1/2}(\nu_\rho + \lambda^{-1})^{1/2}] \approx \exp[-2\pi d(\nu_\rho - \lambda^{-1})^{1/2}(2\lambda^{-1})^{1/2}]$, which equals $\exp(-2\pi)$ when $(\nu_\rho - \lambda^{-1}) \approx \lambda/2d^2$, or $(\nu_\rho - 1/\lambda)/(1/\lambda) \approx \frac{1}{2}(\lambda/d)^2$. For $d \gg \lambda$ the attenuation factor decreases sharply when the spatial frequency slightly exceeds λ^{-1}, as illustrated in Fig. 4.1-11. We may therefore regard λ^{-1} as the cutoff spatial frequency (the spatial bandwidth) of the system. Thus,

> *The spatial bandwidth of light propagation in free space is approximately* λ^{-1} *cycles/mm.*

Features contained in spatial frequencies greater than λ^{-1} (corresponding to details of size finer than λ) cannot be transmitted by an optical wave of wavelength λ over distances much greater than λ.

Fresnel Approximation

The expression for the transfer function in (4.1-9) may be simplified if the input function $f(x, y)$ contains only spatial frequencies that are much smaller than the cutoff frequency λ^{-1}, so that $\nu_x^2 + \nu_y^2 \ll \lambda^{-2}$. The plane-wave components of the propagating light then make small angles $\theta_x \approx \lambda\nu_x$ and $\theta_y \approx \lambda\nu_y$ corresponding to paraxial rays.

Denoting $\theta^2 = \theta_x^2 + \theta_y^2 \approx \lambda^2(\nu_x^2 + \nu_y^2)$, where θ is the angle with the optical axis, the phase factor in (4.1-9) is

$$2\pi d\sqrt{\lambda^{-2} - \nu_x^2 - \nu_y^2} = 2\pi\frac{d}{\lambda}\sqrt{1 - \theta^2}$$

† Evanescent waves are neither forward nor backward propagating; rather, they propagate in the transverse plane where they are generated. We select the negative sign before the square root in (4.1-3) because such waves must attenuate, rather than grow, in the positive z direction absent a gain mechanism enabling them to grow.

$$= 2\pi \frac{d}{\lambda}\left(1 - \frac{\theta^2}{2} + \frac{\theta^4}{8} - \cdots\right). \tag{4.1-10}$$

Neglecting the third and higher terms of this expansion, (4.1-9) may be approximated by

$$H(\nu_x, \nu_y) \approx H_0 \exp\left[j\pi\lambda d\left(\nu_x^2 + \nu_y^2\right)\right], \tag{4.1-11}$$

Transfer Function of Free Space (Fresnel Approximation)

where $H_0 = \exp(-jkd)$. In this approximation, the phase is a quadratic function of ν_x and ν_y, as illustrated in Fig. 4.1-12. This approximation is known as the **Fresnel approximation**.

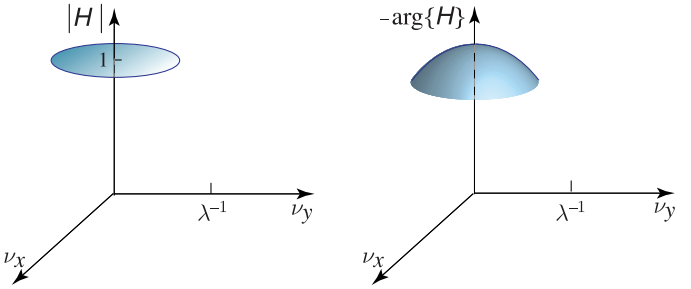
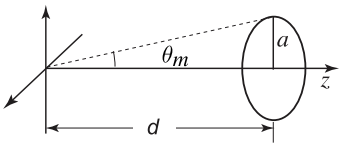


Figure 4.1-12 The transfer function of free-space propagation for low spatial frequencies (much less than $1/\lambda$ cycles/mm) has a constant magnitude and a quadratic phase.

The condition of validity of the Fresnel approximation is that the third term in (4.1-10) is much smaller than π for all θ. This is equivalent to

$$\frac{\theta^4 d}{4\lambda} \ll 1. \tag{4.1-12}$$

If a is the largest radial distance at the output plane, the largest angle $\theta_m \approx a/d$, and (4.1-12) may be written in the form [see (2.2-18)]



$$N_F \frac{\theta_m^2}{4} \ll 1, \tag{4.1-13}$$

Fresnel Approximation Condition of Validity

$$N_F = \frac{a^2}{\lambda d}, \tag{4.1-14}$$

Fresnel Number

where N_F is the Fresnel number. For example, if $a = 1$ cm, $d = 100$ cm, and $\lambda = 0.5$ μm, then $\theta_m = 10^{-2}$ radian, $N_F = 200$, and $N_F\theta^2/4 = 5 \times 10^{-3}$. In this case the Fresnel approximation is applicable.

Input–Output Relation

Given the input function $f(x, y)$, the output function $g(x, y)$ may be determined as follows: (1) we determine the Fourier transform

$$F(\nu_x, \nu_y) = \iint_{-\infty}^{\infty} f(x, y) \exp\left[j2\pi(\nu_x x + \nu_y y)\right] \, dx \, dy, \tag{4.1-15}$$

which represents the complex envelopes of the plane-wave components at the input plane; (2) the product $H(\nu_x, \nu_y)F(\nu_x, \nu_y)$ gives the complex envelopes of the plane-wave components at the output plane; and (3) the complex amplitude at the output plane is the sum of the contributions of these plane waves,

$$g(x, y) = \iint_{-\infty}^{\infty} H(\nu_x, \nu_y)F(\nu_x, \nu_y) \exp[-j2\pi(\nu_x x + \nu_y y)] \, d\nu_x d\nu_y. \tag{4.1-16}$$

Using the Fresnel approximation for $H(\nu_x, \nu_y)$, which is given by (4.1-11), we have

$$g(x, y) = H_0 \iint_{-\infty}^{\infty} F(\nu_x, \nu_y) \exp\left[j\pi\lambda d\left(\nu_x^2 + \nu_y^2\right)\right] \exp\left[-j2\pi(\nu_x x + \nu_y y)\right] \, d\nu_x \, d\nu_y. \tag{4.1-17}$$

Equations (4.1-17) and (4.1-15) serve to relate the output function $g(x, y)$ to the input function $f(x, y)$.

C. Impulse Response Function of Free Space

The impulse response function $h(x, y)$ of the system of free-space propagation is the response $g(x, y)$ when the input $f(x, y)$ is a point at the origin $(0, 0)$. It is the inverse Fourier transform of the transfer function $H(\nu_x, \nu_y)$. Using the results of Sec. A.3 and Table A.1-1 of Appendix A, together with $k = 2\pi/\lambda$, the inverse Fourier transform of (4.1-11) turns out to be

$$h(x, y) \approx h_0 \exp\left[-jk\frac{x^2 + y^2}{2d}\right], \tag{4.1-18}$$

Impulse Response Function
Free Space (Fresnel Approximation)

where $h_0 = (j/\lambda d)\exp(-jkd)$. This function is proportional to the complex amplitude at the $z = d$ plane of a paraboloidal wave centered about the origin $(0, 0)$ [see (2.2-17)]. Thus, each point at the input plane generates a paraboloidal wave; all such waves are superimposed at the output plane.

Free-Space Propagation as a Convolution

An alternative procedure for relating complex amplitudes $f(x, y)$ and $g(x, y)$ is to regard $f(x, y)$ as a superposition of different points (delta functions), each producing a paraboloidal wave. The wave originating at the point (x', y') has an amplitude $f(x', y')$

and is centered about (x', y') so that it generates a wave with amplitude $f(x', y')h(x - x', y - y')$ at the point (x, y) at the output plane. The sum of these contributions is the two-dimensional convolution

$$g(x, y) = \iint_{-\infty}^{\infty} f(x', y')h(x - x', y - y')\, dx'\, dy', \tag{4.1-19}$$

which, in the Fresnel approximation, becomes

$$g(x, y) = h_0 \iint_{-\infty}^{\infty} f(x', y') \exp\left[-j\pi \frac{(x - x')^2 + (y - y')^2}{\lambda d}\right] dx'\, dy', \tag{4.1-20}$$

where $h_0 = (j/\lambda d)\exp(-jkd)$.

In summary: within the Fresnel approximation, there are two approaches to determining the complex amplitude $g(x, y)$ at the output plane, given the complex amplitude $f(x, y)$ at the input plane: (1) Equation (4.1-20) is based on a space-domain approach in which the input wave is expanded in terms of paraboloidal elementary waves; and (2) Equation (4.1-17) is a frequency-domain approach in which the input wave is expanded as a sum of plane waves.

EXERCISE 4.1-2

Gaussian Beams Revisited. If the function $f(x, y) = A\exp[-(x^2 + y^2)/W_0^2]$ represents the complex amplitude of an optical wave $U(x, y, z)$ in the plane $z = 0$, show that $U(x, y, z)$ is the Gaussian beam displayed in (3.1-7). Use both space- and frequency-domain methods.

D. Huygens–Fresnel Principle

The **Huygens–Fresnel principle** states that each point on a wavefront generates a spherical wave (Fig. 4.1-13). The envelope of these secondary waves constitutes a new wavefront. Their superposition constitutes the wave in another plane. The system's impulse response function for propagation between the planes $z = 0$ and $z = d$ is

$$h(x, y) \propto \frac{1}{r}\exp(-jkr), \qquad r = \sqrt{x^2 + y^2 + d^2}\,. \tag{4.1-21}$$

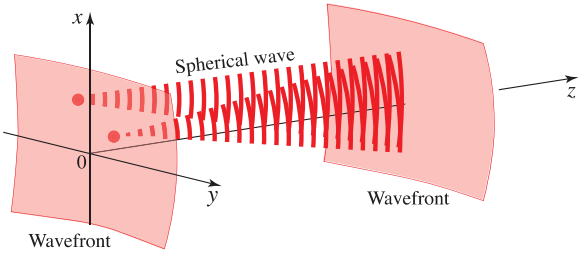


Figure 4.1-13 The Huygens–Fresnel principle. Each point on a wavefront generates a spherical wave.

In the paraxial approximation, the spherical wave given by (4.1-21) is approximated by the paraboloidal wave in (4.1-18) (see Sec. 2.2B). Our derivation of the impulse response function is therefore consistent with the Huygens–Fresnel principle.

4.2 OPTICAL FOURIER TRANSFORM

As has been shown in Sec. 4.1, the propagation of light in free space is described conveniently by Fourier analysis. If the complex amplitude of a monochromatic wave of wavelength λ in the $z = 0$ plane is a function $f(x, y)$ composed of harmonic components of different spatial frequencies, each harmonic component corresponds to a plane wave: the plane wave traveling at angles $\theta_x = \sin^{-1} \lambda\nu_x$, $\theta_y = \sin^{-1} \lambda\nu_y$ corresponds to the components with spatial frequencies ν_x and ν_y and has an amplitude $F(\nu_x, \nu_y)$, the Fourier transform of $f(x, y)$. This suggests that light can be used to compute the Fourier transform of a two-dimensional function $f(x, y)$, simply by making a transparency with amplitude transmittance $f(x, y)$ through which a uniform plane wave of unity magnitude is transmitted.

Because each of the plane waves has an infinite extent and therefore overlaps with the other plane waves, however, it is necessary to find a method of separating these waves. It will be shown that at a sufficiently large distance, only a single plane wave contributes to the total amplitude at each point at the output plane, so that the Fourier components are eventually separated naturally. A more practical approach is to use a lens to focus each of the plane waves into a single point, as described subsequently.

A. Fourier Transform in the Far Field

We now proceed to show that if the propagation distance d is sufficiently long, the only plane wave that contributes to the complex amplitude at a point (x, y) at the output plane is the wave with direction making angles $\theta_x \approx x/d$ and $\theta_y \approx y/d$ with the optical axis (see Fig. 4.2-1). This is the wave with wavevector components $k_x \approx (x/d)k$ and $k_y \approx (y/d)k$ and amplitude $F(\nu_x, \nu_y)$ with $\nu_x = x/\lambda d$ and $\nu_y = x/\lambda d$. The complex amplitudes $g(x, y)$ and $f(x, y)$ of the wave at the $z = d$ and $z = 0$ planes are related by

$$g(x, y) \approx h_0 \, F\left(\frac{x}{\lambda d}, \frac{y}{\lambda d}\right). \tag{4.2-1}$$

Free-Space Propagation as Fourier Transform (Fraunhofer Approximation)

where $F(\nu_x, \nu_y)$ is the Fourier transform of $f(x, y)$ and $h_0 = (j/\lambda d) \exp(-jkd)$. Contributions of all other waves cancel out as a result of destructive interference. This approximation is known as the **Fraunhofer approximation**.

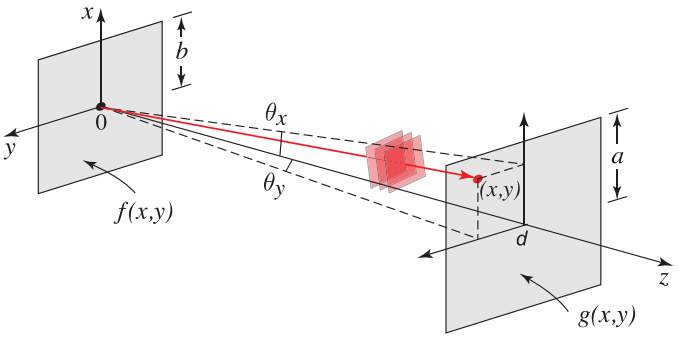


Figure 4.2-1 When the distance d is sufficiently long, the complex amplitude at point (x, y) in the $z = d$ plane is proportional to the complex amplitude of the plane-wave component with angles $\theta_x \approx x/d \approx \lambda\nu_x$ and $\theta_y \approx y/d \approx \lambda\nu_y$, i.e., to the Fourier transform $F(\nu_x, \nu_y)$ of $f(x, y)$, with $\nu_x = x/\lambda d$ and $\nu_y = y/\lambda d$.

As noted in the following proofs, the conditions of validity of Fraunhofer approximation are:

$$\boxed{N_{\mathrm{F}} \ll 1 \quad \text{and} \quad N'_{\mathrm{F}} \ll 1.} \tag{4.2-2}$$

Fraunhofer Approximation Condition of Validity
$N_{\mathrm{F}} = a^2/\lambda d$, $N'_{\mathrm{F}} = b^2/\lambda d$

The Fraunhofer approximation is therefore valid whenever the Fresnel numbers N_{F} and N'_{F} are small. The Fraunhofer approximation is more difficult to satisfy than the Fresnel approximation, which requires that $N_{\mathrm{F}}\theta_m^2/4 \ll 1$ [see (4.1-13)]. Since $\theta_m \ll 1$ in the paraxial approximation, it is possible to satisfy the Fresnel condition $N_{\mathrm{F}}\theta_m^2/4 \ll 1$ for Fresnel numbers N_{F} not necessarily $\ll 1$.

□ **Proofs of the Fourier Transform Property of Free-Space Propagation in the Fraunhofer Approximation.** We begin with the relation between $g(x, y)$ and $f(x, y)$ in (4.1-20). The phase in the argument of the exponent is $(\pi/\lambda d)[(x-x')^2+(y-y')^2] = (\pi/\lambda d)[(x^2+y^2)+(x'^2+y'^2)-2(xx'+yy')]$. If $f(x, y)$ is confined to a small area of radius b, and if the distance d is sufficiently large so that the Fresnel number $N'_{\mathrm{F}} = b^2/\lambda d$ is small, then the phase factor $(\pi/\lambda d)(x'^2 + y'^2) \leq \pi(b^2/\lambda d)$ is negligible and (4.1-20) may be approximated by

$$g(x, y) = h_0 \exp\left(-j\pi\frac{x^2+y^2}{\lambda d}\right) \iint_{-\infty}^{\infty} f(x', y') \exp\left(j2\pi\frac{xx'+yy'}{\lambda d}\right) dx'\, dy'. \tag{4.2-3}$$

The factors $x/\lambda d$ and $y/\lambda d$ may be regarded as the frequencies $\nu_x = x/\lambda d$ and $\nu_y = y/\lambda d$, so that

$$g(x, y) = h_0 \exp\left(-j\pi\frac{x^2+y^2}{\lambda d}\right) F\left(\frac{x}{\lambda d}, \frac{y}{\lambda d}\right), \tag{4.2-4}$$

where $F(\nu_x, \nu_y)$ is the Fourier transform of $f(x, y)$. The phase factor given by $\exp[-j\pi(x^2+y^2)/\lambda d]$ in (4.2-4) may also be neglected and (4.2-1) obtained if we also limit our interest to points at the output plane within a circle of radius a centered about the z-axis so that $\pi(x^2 + y^2)/\lambda d \leq \pi a^2/\lambda d \ll \pi$. This is applicable when the Fresnel number $N_{\mathrm{F}} = a^2/\lambda d \ll 1$.

Another proof is based on (4.1-17), which expresses the complex amplitude $g(x, y)$ as an integral of plane waves of different frequencies. If d is sufficiently large so that the phase in the integrand is much greater than 2π, it can be shown using the method of stationary phase[†] that only one value of ν_x contributes to the integral. This is the value for which the derivative of the phase $\pi\lambda d\nu_x^2 - 2\pi\nu_x x$ with respect to ν_x vanishes; i.e., $\nu_x = x/\lambda d$. Similarly, the only value of ν_y that contributes to the integral is $\nu_y = y/\lambda d$. This proves the assertion that only one plane wave contributes to the far field at a given point. ■

EXERCISE 4.2-1

Conditions of Validity of the Fresnel and Fraunhofer Approximations: A Comparison. Demonstrate that the Fraunhofer approximation is more restrictive than the Fresnel approximation by taking $\lambda = 0.5$ μm, and assuming that the object points lie within a circular aperture of radius $b = 1$ cm and the observation points lie within a circular aperture of radius $a = 2$ cm. Determine the range of distances d between the object plane and the observation plane for which each of these approximations is applicable.

[†] See, e.g., M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th expanded and corrected ed. 2002, Appendix III.

Summary

In the Fraunhofer approximation, the complex amplitude $g(x, y)$ of a wave of wavelength λ in the $z = d$ plane is proportional to the Fourier transform $F(\nu_x, \nu_y)$ of the complex amplitude $f(x, y)$ in the $z = 0$ plane, evaluated at the spatial frequencies $\nu_x = x/\lambda d$ and $\nu_y = y/\lambda d$. The approximation is valid if $f(x, y)$ at the input plane is confined to a circle of radius b satisfying $b^2/\lambda d \ll 1$, and at points at the output plane within a circle of radius a satisfying $a^2/\lambda d \ll 1$.

B. Fourier Transform Using a Lens

The plane-wave components that constitute a wave may also be separated by use of a lens. A thin spherical lens transforms a plane wave into a paraboloidal wave focused to a point in the lens focal plane (see Sec. 2.4 and Exercise 2.4-3). If the plane wave arrives at small angles θ_x and θ_y, the paraboloidal wave is centered about the point $(\theta_x f, \theta_y f)$, where f is the focal length (see Fig. 4.2-2). The lens therefore maps each direction (θ_x, θ_y) into a single point $(\theta_x f, \theta_y f)$ in the focal plane and thus separates the contributions of the different plane waves.

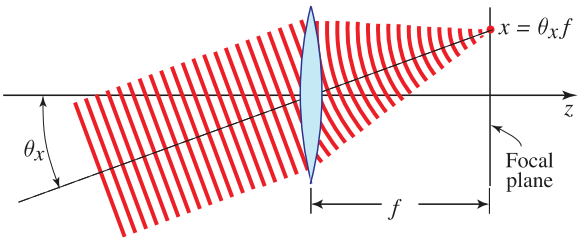


Figure 4.2-2 Focusing of a plane wave into a point. A direction (θ_x, θ_y) is mapped into a point $(x, y) = (\theta_x f, \theta_y f)$. (see Exercise 2.4-3.)

In reference to the optical system shown in Fig. 4.2-3, let $f(x, y)$ be the complex amplitude of the optical wave in the $z = 0$ plane. Light is decomposed into plane waves, with the wave traveling at small angles $\theta_x = \lambda\nu_x$ and $\theta_y = \lambda\nu_y$ having a complex amplitude proportional to the Fourier transform $F(\nu_x, \nu_y)$. This wave is focused by the lens into a point (x, y) in the focal plane where $x = \theta_x f = \lambda f \nu_x$ and $y = \theta_y f = \lambda f \nu_y$. The complex amplitude at point (x, y) at the output plane is therefore proportional to the Fourier transform of $f(x, y)$ evaluated at $\nu_x = x/\lambda f$ and $\nu_y = y/\lambda f$, so that

$$g(x, y) \propto F\left(\frac{x}{\lambda f}, \frac{y}{\lambda f}\right). \tag{4.2-5}$$

To determine the proportionality factor in (4.2-5), we analyze the input function $f(x, y)$ into its Fourier components and trace the plane wave corresponding to each component through the optical system. We then superpose the contributions of these waves at the output plane to obtain $g(x, y)$. Assuming that these waves are paraxial and using the Fresnel approximation, we obtain:

$$g(x, y) = h_l \exp\left[j\pi \frac{(x^2 + y^2)(d - f)}{\lambda f^2}\right] F\left(\frac{x}{\lambda f}, \frac{y}{\lambda f}\right), \tag{4.2-6}$$

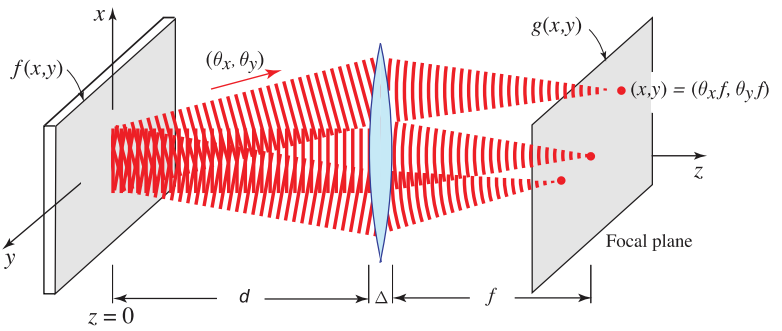


Figure 4.2-3 Focusing of the plane waves associated with the harmonic Fourier components of the input function $f(x, y)$ into points in the focal plane. The amplitude of the plane wave with direction $(\theta_x, \theta_y) = (\lambda\nu_x, \lambda\nu_y)$ is proportional to the Fourier transform $F(\nu_x, \nu_y)$ and is focused at the point $(x, y) = (\theta_x f, \theta_y f) = (\lambda f\nu_x, \lambda f\nu_y)$.

where $h_l = H_0 h_0 = (j/\lambda f)\exp[-jk(d + f)]$. Thus, the coefficient of proportionality in (4.2-5) contains a phase factor that is a quadratic function of x and y.

Since $|h_l| = 1/\lambda f$ it follows from (4.2-6) that the optical intensity at the output plane is

$$I(x, y) = \frac{1}{(\lambda f)^2}\left|F\left(\frac{x}{\lambda f}, \frac{y}{\lambda f}\right)\right|^2. \tag{4.2-7}$$

The intensity of light at the output plane (the back focal plane of the lens) is therefore proportional to the absolute-squared value of the Fourier transform of the complex amplitude of the wave at the input plane, regardless of the distance d.

The phase factor in (4.2-6) vanishes if $d = f$, so that

$$\boxed{g(x, y) = h_l\, F\left(\frac{x}{\lambda f}, \frac{y}{\lambda f}\right),} \tag{4.2-8}$$

Fourier-Transform Property of a Lens

where $h_l = (j/\lambda f)\exp(-j2kf)$. In this geometry, known as the **2-*f* system** (see Fig. 4.2-4), the complex amplitudes at the front and back focal planes of the lens are related by a Fourier transform, both magnitude and phase.

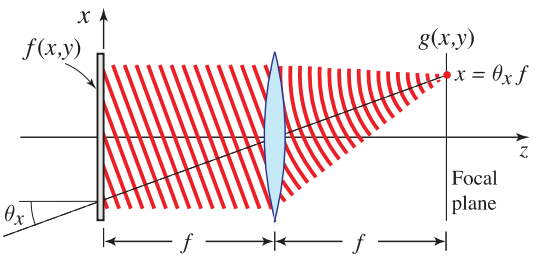


Figure 4.2-4 The 2-f system. The Fourier component of $f(x, y)$ with spatial frequencies ν_x and ν_y generates a plane wave at angles $\theta_x = \lambda\nu_x$ and $\theta_y = \lambda\nu_y$ and is focused by the lens to the point $(x, y) = (f\theta_x, f\theta_y) = (\lambda f\nu_x, \lambda f\nu_y)$ so that $g(x, y)$ is proportional to the Fourier transform $F(x/\lambda f, y/\lambda f)$.

Summary

The complex amplitude of light at a point (x, y) in the back focal plane of a lens of focal length f is proportional to the Fourier transform of the complex amplitude in the front focal plane evaluated at the frequencies $\nu_x = x/\lambda f, \nu_y/\lambda f$. This relation is valid in the Fresnel approximation. Without the lens, the Fourier transformation is obtained only in the Fraunhofer approximation, which is more restrictive.

□ *__Proof of the Fourier Transform Property of the Lens in the Fresnel Approximation.__ The proof takes the following four steps.

1. The plane wave with angles $\theta_x = \lambda\nu_x$ and $\theta_y = \lambda\nu_y$ has a complex amplitude $U(x, y, 0) = F(\nu_x, \nu_y) \exp[-j2\pi(\nu_x x + \nu_y y)]$ in the $z = 0$ plane and $U(x, y, d) = H(\nu_x, \nu_y)F(\nu_x, \nu_y) \exp[-j2\pi(\nu_x x + \nu_y y)]$ in the $z = d$ plane, immediately before crossing the lens, where $H(\nu_x, \nu_y) = H_0 \exp[j\pi\lambda d(\nu_x^2 + \nu_y^2)]$ is the transfer function of a distance d of free space and $H_0 = \exp(-jkd)$.
2. Upon crossing the lens, the complex amplitude is multiplied by the lens phase factor $\exp[j\pi(x^2 + y^2)/\lambda f]$ [the phase factor $\exp(-jk\Delta)$, where Δ is the width of the lens, has been ignored]. Thus,

$$U(x, y, d + \Delta) = H_0 \exp\left[j\pi(x^2 + y^2)/\lambda f\right] \times \exp\left[j\pi\lambda d\left(\nu_x^2 + \nu_y^2\right)\right] F(\nu_x, \nu_y) \exp\left[-j2\pi(\nu_x x + \nu_y y)\right]. \tag{4.2-9}$$

   This expression is simplified by writing $-2\nu_x x + x^2/\lambda f = (x^2 - 2\nu_x \lambda f x)/\lambda f = [(x - x_0)^2 - x_0^2]/\lambda f$, with $x_0 = \lambda\nu_x f$; a similar relation for y is written with $y_0 = \lambda\nu_y f$, so that

$$U(x, y, d + \Delta) = A(\nu_x, \nu_y) \exp\left[j\pi \frac{(x - x_0)^2 + (y - y_0)^2}{\lambda f}\right], \tag{4.2-10}$$

   where

$$A(\nu_x, \nu_y) = H_0 \exp\left[j\pi\lambda(d - f)\left(\nu_x^2 + \nu_y^2\right)\right] F(\nu_x, \nu_y). \tag{4.2-11}$$

   Equation (4.2-10) is recognized as the complex amplitude of a paraboloidal wave converging toward the point (x_0, y_0) in the lens focal plane, $z = d + \Delta + f$.
3. We now examine the propagation in the free space between the lens and the output plane to determine $U(x, y, d + \Delta + f)$. We apply (4.1-20) to (4.2-10), use the relation $\int \exp[j2\pi(x - x_0)x'/\lambda f]\, dx' = \lambda f \delta(x - x_0)$, and obtain

$$U(x, y, d + \Delta + f) = h_0(\lambda f)^2 A(\nu_x, \nu_y)\, \delta(x - x_0)\, \delta(y - y_0), \tag{4.2-12}$$

   where $h_0 = (j/\lambda f) \exp(-jkf)$. Indeed, the plane wave is focused into a single point at $x_0 = \lambda\nu_x f$ and $y_0 = \lambda\nu_y f$.
4. The last step is to integrate over all the plane waves (all ν_x and ν_y). By virtue of the sifting property of the delta function, $\delta(x - x_0) = \delta(x - \lambda f\nu_x) = (1/\lambda f)\delta(\nu_x - x/\lambda f)$, this integral gives $g(x, y) = h_0 A(x/\lambda f, y/\lambda f)$. Substituting from (4.2-11) we finally obtain (4.2-6). ■

EXERCISE 4.2-2

The Inverse Fourier Transform. In the single-lens optical system depicted in Fig. 4.2-4, the field distribution in the front focal plane ($z = 2f$) is a scaled version of the Fourier transform of the field distribution in the back focal plane ($z = 0$). Verify that if the coordinate system in the front focal plane is inverted, i.e., $(x, y) \to (-x, -y)$, then the resultant field distribution yields the inverse Fourier transform.

4.3 DIFFRACTION OF LIGHT

When an optical wave is transmitted through an aperture in an opaque screen and travels some distance in free space, its intensity distribution is called the diffraction pattern. If light were treated as rays, the diffraction pattern would be a shadow of the aperture. Because of the wave nature of light, however, the diffraction pattern may deviate slightly or substantially from the aperture shadow, depending on the distance between the aperture and observation plane, the wavelength, and the dimensions of the aperture. An example is illustrated in Fig. 4.3-1. It is difficult to determine exactly the manner in which the screen modifies the incident wave, but the propagation in free space beyond the aperture is always governed by the laws described earlier in this chapter.

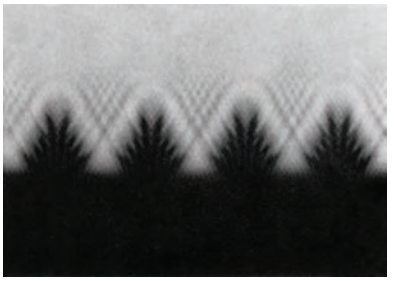


Figure 4.3-1 Diffraction pattern of the teeth of a saw. (Adapted from M. Cagnet, M. Françon, and J. C. Thrierr, *Atlas of Optical Phenomena*, Springer-Verlag, 1962.)

The simplest theory of diffraction is based on the *assumption* that the incident wave is transmitted without change at points within the aperture, but is reduced to zero at points on the back side of the opaque part of the screen. If $U(x, y)$ and $f(x, y)$ are the complex amplitudes of the wave immediately to the left and right of the screen (Fig. 4.3-2), respectively, then in accordance with this assumption,

$$f(x, y) = U(x, y)\, p(x, y), \tag{4.3-1}$$

where

$$p(x, y) = \begin{cases} 1 & \text{inside the aperture} \\ 0, & \text{outside the aperture} \end{cases} \tag{4.3-2}$$

is called the **aperture function**.

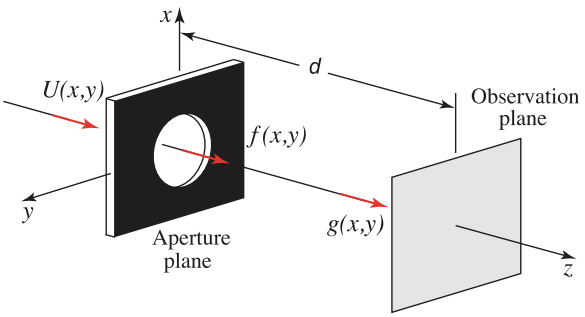


Figure 4.3-2 A wave $U(x, y)$ is transmitted through an aperture of amplitude transmittance $p(x, y)$, generating a wave of complex amplitude $f(x, y) = U(x, y)p(x, y)$. After propagation a distance d in free space, the complex amplitude is $g(x, y)$ and the diffraction pattern is the intensity $I(x, y) = |g(x, y)|^2$.

Given $f(x, y)$, the complex amplitude $g(x, y)$ at an observation plane a distance d from the screen may be determined using the methods described in Secs. 4.1 and 4.2. The diffraction pattern $I(x, y) = |g(x, y)|^2$ is known as **Fraunhofer diffraction** or

Fresnel diffraction, depending on whether free-space propagation is described using the Fraunhofer approximation or the Fresnel approximation, respectively.

Although this approach gives reasonably accurate results in most cases, it is not exact. The validity and self-consistency of the assumption that the complex amplitude $f(x, y)$ vanishes at points outside the aperture on the back of the screen are questionable since the transmitted wave propagates in all directions and therefore reaches those points as well. A theory of diffraction based on the exact solution of the Helmholtz equation under the boundary conditions imposed by the aperture is mathematically difficult. Only a few geometrical structures have yielded exact solutions. However, different theories of diffraction have been developed using a variety of assumptions, leading to results with varying accuracies. Rigorous diffraction theory is beyond the scope of this book.

A. Fraunhofer Diffraction

Fraunhofer diffraction is the theory of transmission of light through apertures, assuming that the incident wave is multiplied by the aperture function and that the Fraunhofer approximation determines the propagation of light in the free space beyond the aperture. The Fraunhofer approximation is valid if the propagation distance d between the aperture and observation planes is sufficiently large so that the Fresnel number $N_{\mathrm{F}}' = b^2/\lambda d \ll 1$, where b is the largest radial distance within the aperture.

Assuming that the incident wave is a plane wave of intensity I_i traveling in the z direction so that $U(x, y) = \sqrt{I_i}$, then $f(x, y) = \sqrt{I_i}\, p(x, y)$. In the Fraunhofer approximation [see (4.2-1)],

$$g(x, y) \approx \sqrt{I_i}\, h_0\, P\left(\frac{x}{\lambda d}, \frac{y}{\lambda d}\right), \tag{4.3-3}$$

where

$$P(\nu_x, \nu_y) = \iint_{-\infty}^{\infty} p(x, y) \exp\left[j2\pi\left(\nu_x x + \nu_y y\right)\right] dx\, dy \tag{4.3-4}$$

is the Fourier transform of $p(x, y)$ and $h_0 = (j/\lambda d)\exp(-jkd)$. The diffraction pattern is therefore

$$\boxed{I(x, y) = \frac{I_i}{(\lambda d)^2}\left|P\left(\frac{x}{\lambda d}, \frac{y}{\lambda d}\right)\right|^2.} \tag{4.3-5}$$

In summary: the Fraunhofer diffraction pattern at the point (x, y) is proportional to the squared magnitude of the Fourier transform of the aperture function $p(x, y)$ evaluated at the spatial frequencies $\nu_x = x/\lambda d$ and $\nu_y = y/\lambda d$.

EXERCISE 4.3-1

Fraunhofer Diffraction from a Rectangular Aperture. Verify that the Fraunhofer diffraction pattern from a rectangular aperture, of height and width D_x and D_y respectively, observed at a distance d is

$$I(x, y) = I_o\, \mathrm{sinc}^2\left(\frac{D_x x}{\lambda d}\right) \mathrm{sinc}^2\left(\frac{D_y y}{\lambda d}\right), \tag{4.3-6}$$

where $I_o = (D_x D_y/\lambda d)^2\, I_i$ is the peak intensity and $\mathrm{sinc}(x) \equiv \sin(\pi x)/(\pi x)$. Verify that the first

zeros of this pattern occur at $x = \pm\lambda d/D_x$ and $y = \pm\lambda d/D_y$, so that the angular divergence of the diffracted light is given by

$$\theta_x = \frac{\lambda}{D_x}, \qquad \theta_y = \frac{\lambda}{D_y}. \tag{4.3-7}$$

If $Dy < D_x$, the diffraction pattern is wider in the y direction than in the x direction, as illustrated in Fig. 4.3-3.

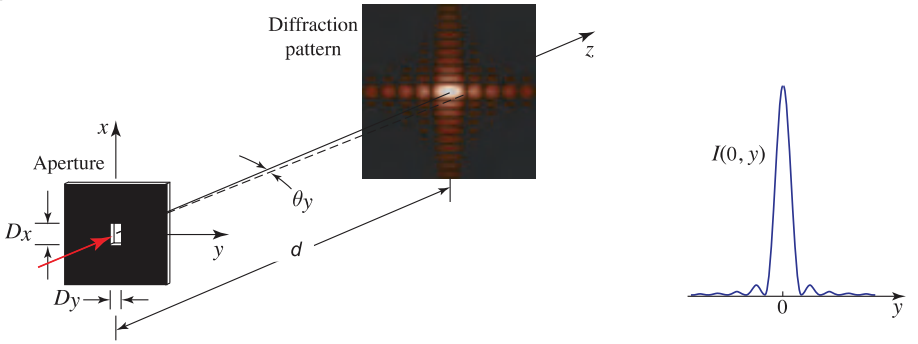


Figure 4.3-3 Fraunhofer diffraction from a rectangular aperture. The central lobe of the pattern has half-angular widths $\theta_x = \lambda/D_x$ and $\theta_y = \lambda/D_y$.

EXERCISE 4.3-2

Fraunhofer Diffraction from a Circular Aperture. Verify that the Fraunhofer diffraction pattern from a circular aperture of diameter D (Fig. 4.3-4) is

$$I(x, y) = I_o \left[\frac{2J_1(\pi D\rho/\lambda d)}{\pi D\rho/\lambda d}\right]^2, \qquad \rho = \sqrt{x^2 + y^2}, \tag{4.3-8}$$

where $I_o = (\pi D^2/4\lambda d)^2 I_i$ is the peak intensity and $J_1(\cdot)$ is the Bessel function of order 1. The Fourier transform of circularly symmetric functions is discussed in Appendix A, Sec. A.3. The circularly symmetric pattern (4.3-8), known as the **Airy pattern**, consists of a central disk surrounded by rings. Verify that the radius of the central disk, known as the **Airy disk**, is $\rho_s = 1.22\lambda d/D$ and subtends an angle

$$\theta = 1.22\frac{\lambda}{D}. \tag{4.3-9}$$

Airy Disk Half Angle

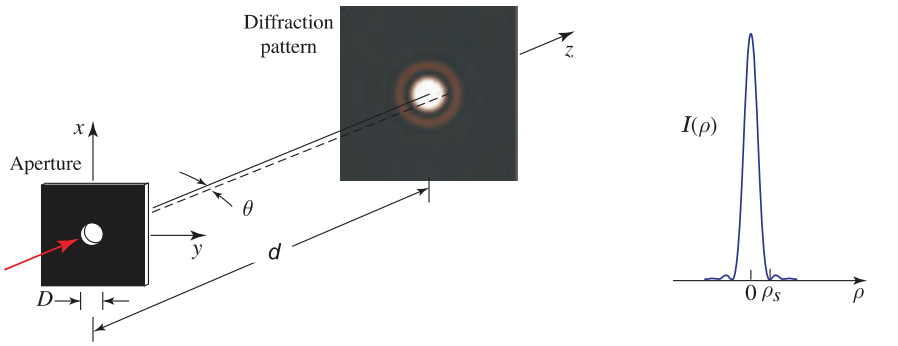


Figure 4.3-4 The Fraunhofer diffraction pattern from a circular aperture produces the Airy pattern with the radius of the central disk subtending an angle $\theta = 1.22\lambda/D$.

The Fraunhofer approximation is valid for distances d that are usually extremely large. It is satisfied, for example, in applications of long-distance free-space optical communications such as laser radar (lidar) and satellite communications. However, as shown in Sec. 4.2B, if a lens of focal length f is used to focus the diffracted light, the intensity pattern in the focal plane is proportional to the squared magnitude of the Fourier transform of $p(x, y)$ evaluated at $\nu_x = x/\lambda f$ and $\nu_y = y/\lambda_f$. The observed pattern is therefore identical to that obtained from (4.3-5), with the distance d replaced by the focal length f.

EXERCISE 4.3-3

Spot Size of a Focused Optical Beam. A beam of light is focused using a lens of focal length f with a circular aperture of diameter D (Fig. 4.3-5). If the beam is approximated by a plane wave at points within the aperture, verify that the pattern of the focused spot is

$$I(x, y) = I_o \left[\frac{2J_1(\pi D\rho/\lambda f)}{\pi D\rho/\lambda f} \right]^2, \qquad \rho = \sqrt{x^2 + y^2}, \tag{4.3-10}$$

where I_o is the peak intensity. Compare the radius of the focused spot,

$$\rho_s = 1.22\lambda \frac{f}{D}, \tag{4.3-11}$$

to the spot size obtained when a Gaussian beam of waist radius W_0 is focused by an ideal lens of infinite aperture [see (3.2-15)].

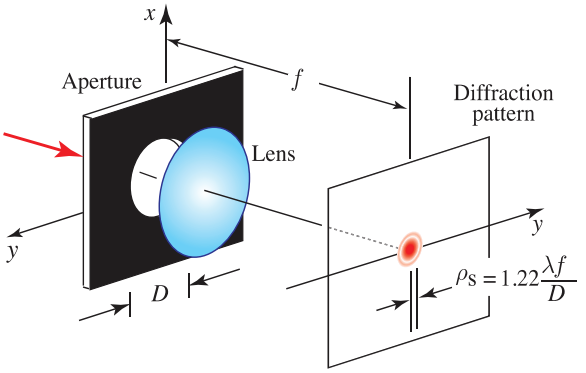


Figure 4.3-5 Focusing of a plane wave transmitted through a circular aperture of diameter D.

*B. Fresnel Diffraction

The theory of Fresnel diffraction is based on the assumption that the incident wave is multiplied by the aperture function $p(x, y)$ and propagates in free space in accordance with the Fresnel approximation. If the incident wave is a plane wave traveling in the z-direction with intensity I_i, the complex amplitude immediately after the aperture is $f(x, y) = \sqrt{I_i}\, p(x, y)$. Using (4.1-20), the diffraction pattern $I(x, y) = |g(x, y)|^2$ at a distance d is

$$I(x, y) = \frac{I_i}{(\lambda d)^2} \left| \int\int_{-\infty}^{\infty} p(x', y') \exp\left[-j\pi \frac{(x - x')^2 + (y - y')^2}{\lambda d} \right] dx'\, dy' \right|^2. \tag{4.3-12}$$

It is convenient to normalize all distances using $\sqrt{\lambda d}$ as a unit of distance, so that $X = x/\sqrt{\lambda d}$ and $X' = x'/\sqrt{\lambda d}$ are the normalized distances (and similarly for y and y'). Equation (4.3-12) then gives

$$I(X,Y) = I_i \left| \iint_{-\infty}^{\infty} p(X',Y') \exp\left\{-j\pi\left[(X-X')^2 + (Y-Y')^2\right]\right\} dX'\,dY' \right|^2. \tag{4.3-13}$$

The integral in (4.3-13) is the convolution of $p(X,Y)$ and $\exp[-j\pi(X^2+Y^2)]$. The real and imaginary parts of $\exp(-j\pi X^2)$, $\cos \pi X^2$ and $\sin \pi X^2$, respectively, are plotted in Fig. 4.3-6. They oscillate at an increasing frequency and their first lobes lie in the intervals $|X| < 1/\sqrt{2}$ and $|X| < 1$, respectively. The total area under the function $\exp(-j\pi X^2)$ is 1, with the main contribution to the area coming from the first few lobes, since subsequent lobes cancel out. If a is the radius of the aperture, the radius of the normalized function $p(X,Y)$ is $a/\sqrt{\lambda d}$. The result of the convolution, which depends on the relative size of the two functions, is therefore governed by the Fresnel number $N_F = a^2/\lambda d$.

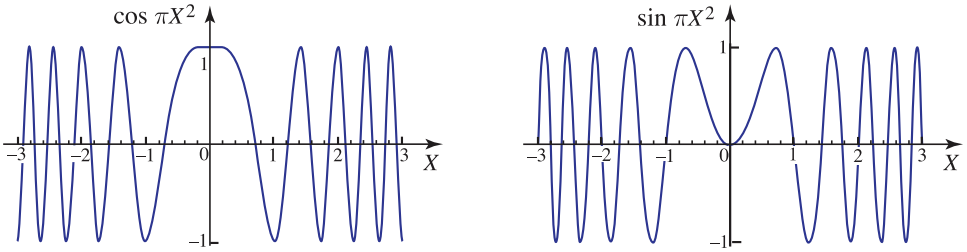


Figure 4.3-6 The real and imaginary parts of $\exp(-j\pi X^2)$.

If the Fresnel number is large, the normalized width of the aperture $a/\sqrt{\lambda d}$ is much greater than the width of the main lobe, and the convolution yields approximately the wider function $p(X,Y)$. Under this condition the Fresnel diffraction pattern is a shadow of the aperture, as would be expected from ray optics. Note that ray optics is applicable in the limit $\lambda \to 0$, which corresponds to the limit $N_F \to \infty$. In the opposite limit, when N_F is small, the Fraunhofer approximation becomes applicable and the Fraunhofer diffraction pattern is obtained.

EXAMPLE 4.3-1. ***Fresnel Diffraction from a Slit.*** Assume that the aperture is a slit of width $D = 2a$, so that $p(x,y) = 1$ when $|x| \le a$, and 0 elsewhere. The normalized coordinate $X = x/\sqrt{\lambda d}$ and

$$P(X,Y) = \begin{cases} 1, & |X| \le \dfrac{a}{\sqrt{\lambda d}} = \sqrt{N_F} \\ 0, & \text{elsewhere,} \end{cases} \tag{4.3-14}$$

where $N_F = a^2/\lambda d$ is the Fresnel number. Substituting into (4.3-13), we obtain $I(X,Y) = I_i|g(X)|^2$, where

$$g(X) = \int_{-\sqrt{N_F}}^{\sqrt{N_F}} \exp\left[-j\pi(X-X')^2\right] dX' = \int_{X-\sqrt{N_F}}^{X+\sqrt{N_F}} \exp(-j\pi X'^2)\, dX'. \tag{4.3-15}$$

This integral is usually written in terms of the Fresnel integrals

$$C(x) = \int_0^x \cos\frac{\pi\alpha^2}{2}\,d\alpha, \qquad S(x) = \int_0^x \sin\frac{\pi\alpha^2}{2}\,d\alpha, \tag{4.3-16}$$

which are available in the standard computer mathematical libraries.

The complex function $g(X)$ may also be evaluated using Fourier-transform techniques. Since $g(x)$ is the convolution of a rectangular function of width $\sqrt{N_F}$ and $\exp(-j\pi X^2)$, its Fourier transform $G(\nu_x) \propto \text{sinc}(\sqrt{N_F}\,\nu_x)\exp(j\pi\nu_x^2)$ (see Table A.1-1 in Appendix A). Thus, $g(X)$ may be computed by determining the inverse Fourier transform of $G(\nu_x)$. If $N_F \gg 1$, the width of $\text{sinc}(\sqrt{N_F}\,\nu_x)$ is much narrower than the width of the first lobe of $\exp(j\pi\nu_x^2)$ (see Fig. 4.3-6) so that $G(\nu_x) \approx \text{sinc}(\sqrt{N_F}\,\nu_x)$ and $g(X)$ is the rectangular function representing the aperture shadow.

The diffraction pattern from a slit is plotted in Fig. 4.3-7 for different Fresnel numbers corresponding to different distances d from the aperture. At very small distances (very large N_F), the diffraction pattern is a perfect shadow of the slit. As the distance increases (N_F decreases), the wave nature of light is exhibited in the form of small oscillations around the edges of the aperture (see also the diffraction pattern in Fig. 4.3-1). For very small N_F, the Fraunhofer pattern described by (4.3-6) is obtained. This is a sinc function with the first zero subtending an angle $\lambda/D = \lambda/2a$.

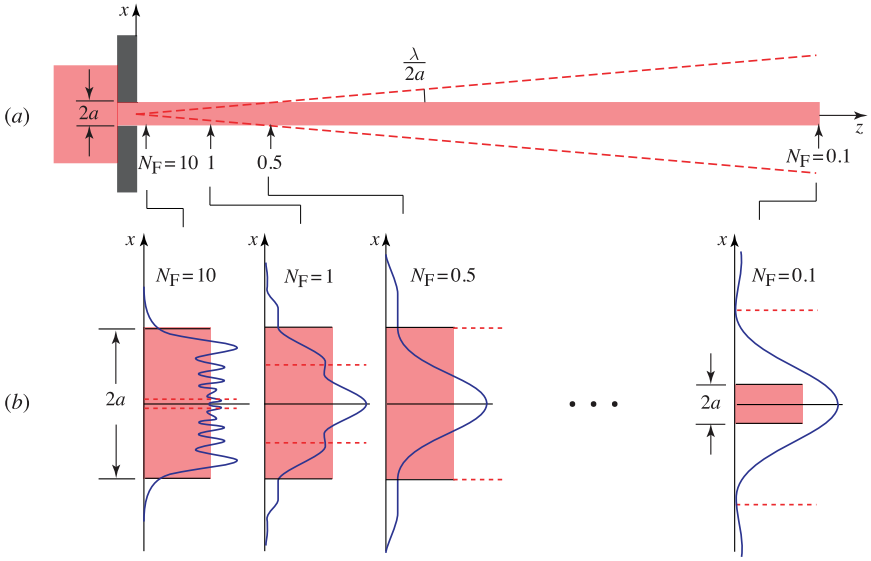


Figure 4.3-7 Fresnel diffraction from a slit of width $D = 2a$. (a) Shaded area is the geometrical shadow of the aperture. The dashed line is the width of the Fraunhofer diffracted beam. (b) Diffraction pattern at four axial positions marked by the arrows in (a) and corresponding to the Fresnel numbers $N_F = 10$, 1, 0.5, and 0.1. The shaded area represents the geometrical shadow of the slit. The dashed lines at $|x| = (\lambda/D)d$ represent the width of the Fraunhofer pattern in the far field. Where the dashed lines coincide with the edges of the geometrical shadow, the Fresnel number $N_F = a^2/\lambda d = 0.5$.

EXAMPLE 4.3-2. ***Fresnel Diffraction from a Gaussian Aperture.*** If the aperture function $p(x, y)$ is the Gaussian function $p(x, y) = \exp[-(x^2 + y^2)/W_0^2]$, the Fresnel diffraction equation (4.3-12) may be evaluated exactly by finding the convolution of $\exp[-(x^2 + y^2)/W_0^2]$ with $h_0 \exp[-j\pi(x^2 + y^2)/\lambda d]$ using, for example, Fourier transform techniques (see Appendix A). The resultant diffraction pattern is

$$I(x, y) = I_i \left[\frac{W_0}{W(d)}\right]^2 \exp\left[-2\frac{x^2 + y^2}{W^2(d)}\right], \tag{4.3-17}$$

where $W^2(d) = W_0^2 + \theta_0^2 d^2$ and $\theta_0 = \lambda/\pi W_0$.

The diffraction pattern is a Gaussian function of $1/e^2$ half-width $W(d)$. For small d, $W(d) \approx W_0$; but as d increases, $W(d)$ increases and approaches $W(d) \approx \theta_0 d$ when d is sufficiently large for the Fraunhofer approximation to be applicable, so that the angle subtended by the Fraunhofer diffraction

pattern is θ_0. These results are illustrated in Fig. 4.3-8, which is analogous to the illustration in Fig. 4.3-7 for diffraction from a slit. The wave diffracted from a Gaussian aperture is the Gaussian beam described in detail in Chapter 3.

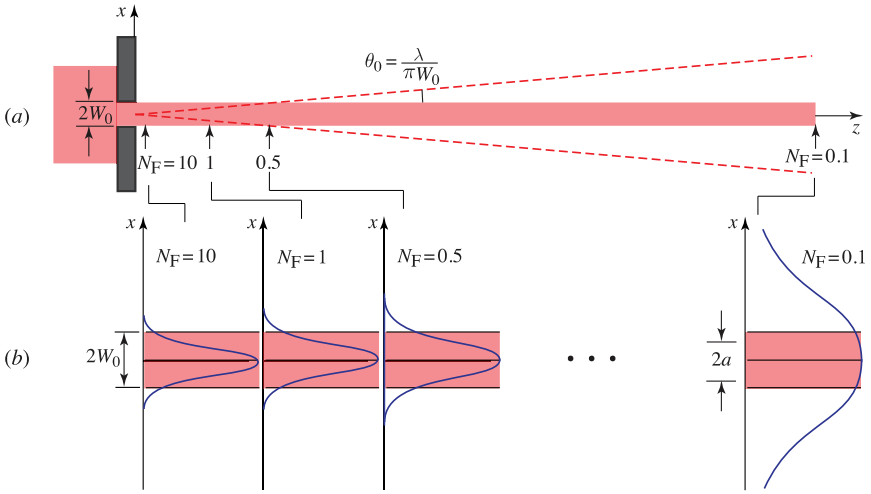


Figure 4.3-8 Fresnel diffraction pattern for a Gaussian aperture of radius W_0 at distances d such that the parameter $(\pi/2)W_0^2/\lambda d$, which is analogous to the Fresnel number N_F in Fig. 4.3-7, is 10, 1, 0.5, and 0.1. These values correspond to $W(d)/W_0 = 1.001$, 1.118, 1.414, and 5.099, respectively. The diffraction pattern is Gaussian at all distances.

Summary

In the order of increasing distance from the aperture, the diffraction pattern is:

1. A shadow of the aperture.
2. A Fresnel diffraction pattern, which is the convolution of the normalized aperture function with $\exp[-j\pi(X^2 + Y^2)]$.
3. A Fraunhofer diffraction pattern, which is the absolute-squared value of the Fourier transform of the aperture function. The far field has an angular divergence proportional to λ/D, where D is the diameter of the aperture.

Fresnel Diffraction from a Periodic Aperture: The Talbot Effect

Fresnel diffraction from a one-dimensional periodic aperture is best described in the Fourier domain by expanding the aperture function $p(x)$ in a Fourier series. If Λ is its period, then the Fourier expansion has frequencies $\nu_x = m/\Lambda$, where $m = 0, \pm 1, \pm 2, \ldots$. The transfer function of free space (4.1-11), at a distance z, is then

$$H_0 \exp\left(j\pi\lambda z\nu_x^2\right) = H_0 \exp\left(j\pi\lambda z m^2/\Lambda^2\right) = H_0 \exp\left(j2\pi m^2 z/z_T\right), \qquad (4.3\text{-}18)$$

where $z_T = 2\Lambda^2/\lambda$. At $z = z_T$, or multiples thereof, the transfer function is simply a constant H_0, independent of the harmonic order m. At these specific distances, then, each of the harmonic functions comprising the aperture function $p(x)$ is multiplied by the same factor so that the function $p(x)$ is reproduced. This process of self imaging is known as the **Talbot effect**, and the distance z_T is called the *Talbot*

distance. At distances z unequal to multiples of z_{T}, the field is given by $U(x,z) = H_0 \sum_{m=-\infty}^{\infty} c_m \exp(jmx/\Lambda) \exp\left(j2\pi m^2 z/z_{\mathrm{T}}\right)$, where the c_m are coefficients of the Fourier series expansion of $p(x)$. The corresponding intensity $I(x,z) = |U(x,z)|^2$ for an opaque screen with parallel slits exhibits a carpet-like pattern, as illustrated in Fig. 4.3-9.

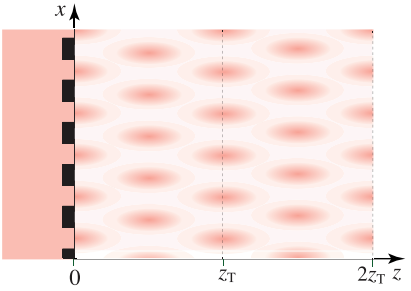


Figure 4.3-9 Talbot effect. Fresnel diffraction pattern from a periodic aperture that takes the form of parallel slits in an otherwise opaque screen. The pattern is reproduced at distances that are multiples of the Talbot distance z_{T}. The result has the appearance of a carpet with periodic patterns in x and z.

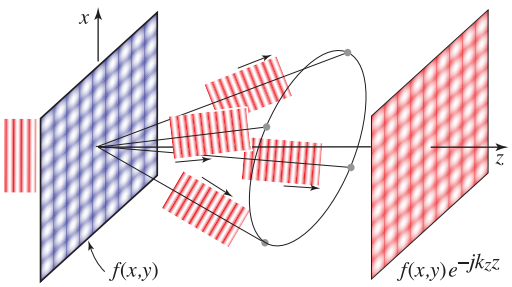
The Talbot effect is also observed for two-dimensional periodic apertures provided that the period is the same in the x and y directions.

*C. Nondiffracting Waves

In accordance with (4.1-8) and (4.1-9), the transfer function of free-space is $\exp(-jk_z z)$, where $k_z = \sqrt{k_o^2 - k_x^2 - k_y^2}$ is a circularly symmetric complex function of $k_T^2 = k_x^2 + k_y^2$. Any two input harmonic functions with spatial frequencies for which k_T^2 is the same thus have the same value of the transfer function. It follows that a function $f(x,y)$ that is a superposition of harmonic functions, all with the same value of k_T and therefore the same k_z, creates a stationary wave $U(x,y,z) = f(x,y)\exp(-jk_z z)$ that maintains its transverse distribution, and is therefore nondiffracting, as it travels through free space, regardless of the distance z. The wavefronts of such a wave are planes orthogonal to the z axis, and the propagation constant is k_z. Nondiffracting optical beams were considered in Sec. 3.5.

EXAMPLE 4.3-3. ***Two Plane Waves.*** The function $f(x,y) = \cos(\alpha x) = \frac{1}{2}[\exp(-j\alpha x) + \exp(+j\alpha x)]$ comprises two harmonic components with spatial angular frequencies $k_x = \pm\alpha$. On propagation through free space, each of these components is modified by the same factor $\exp(-jk_z z)$, where $k_z = \sqrt{k_o^2 - \alpha^2}$. The result is a stationary wave that takes the form $U(x,y,z) = \cos(\alpha x)\exp(-jk_z z)$, which has a sinusoidal transverse distribution representing the interference between two oblique plane waves at angles $\pm\sin^{-1}(\alpha/k_o)$, as provided by (2.5-7).

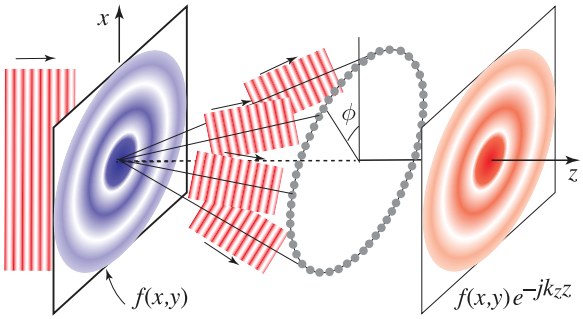
EXAMPLE 4.3-4. ***Four Plane Waves.*** A plane wave traveling in the z direction that is modulated by the function $f(x,y) = \cos(\alpha_x x)\cos(\alpha_y y)$ results in four waves, at angles $\pm\sin^{-1}(\alpha_x/k_o)$ and $\pm\sin^{-1}(\alpha_y/k_o)$ with respect to the x and y axes, respectively. Since the quantity $k_T^2 = \alpha_x^2 + \alpha_y^2$ is the same for all four waves, so too is $k_z = \sqrt{k_o^2 - k_T^2}$. The outcome at z is thus the stationary wave $U(x,y,z) = \cos(\alpha_x x)\cos(\alpha_y y)\exp(-jk_z z)$, as shown.



EXAMPLE 4.3-5. ***Infinite Number of Plane Waves.*** Consider now a function $f(x,y)$ composed of several harmonic functions of angular frequencies $k_x = k_T\cos\phi$ and $k_y = k_T\sin\phi$, with

fixed k_T but different ϕ. The quantities $k_x^2 + k_y^2 = k_T^2$ and $k_z = \sqrt{k_o^2 - k_T^2}$ are thus the same for each of these functions. This superposition of waves therefore corresponds to a stationary wave $U(x, y, z) = f(x, y) \exp(-jk_z z)$, no matter how many values of ϕ are included. A limiting case is the superposition in which a continuum of harmonic functions extends over all angles ϕ, which yields $f(x, y) = \int_{-\pi}^{\pi} \exp(-j\, k_T \cos\phi\, x - j\, k_T \sin\phi\, y)\, d\phi$.

The result is a continuum of plane waves whose directions form a cone of half-angle $\sin^{-1}(k_T/k_o)$. This superposition wave is nothing other than the Bessel beam $U(x, y, z) = 2\pi J_0\left(k_T\sqrt{x^2 + y^2}\right) \exp(-jk_z z)$ described in (3.5-4) and illustrated at right. The connection is explicitly forged via the identity $\int_{-\pi}^{\pi} \exp(-ju \sin\phi)\, d\phi = 2\pi J_0(u)$, where $J_0(u)$ is the Bessel function of the first kind and zeroth order.



The plane-wave superposition associated with the Bessel beam may be implemented by making use of an axicon (see Example 2.4-1 and Sec. 3.5A).

4.4 IMAGE FORMATION

An ideal image formation system is an optical system that replicates the distribution of light in one plane, the object plane, into another, the image plane. Since the optical transmission process is never perfect, the image is never an exact replica of the object. Aside from image magnification, there is also blur resulting from imperfect focusing and from the diffraction of optical waves. This section is devoted to the description of image formation systems and their fidelity. Methods of linear systems, such as the impulse response function and the transfer function (Appendix B), are used to characterize image formation. A simple ray-optics approach is presented first, then a treatment based on wave optics is subsequently developed.

A. Ray Optics of a Single-Lens Imaging System

Consider an imaging system using a lens of focal length f at distances d_1 and d_2 from the object and image planes, respectively, as shown in Fig. 4.4-1. When $1/d_1 + 1/d_2 = 1/f$, the system is focused so that paraxial rays emitted from each point in the object plane reach a single corresponding point in the image plane. Within the ray theory of light, the imaging is "ideal," with each point of the object producing a single point of the image. The impulse response function of the system is an impulse function.

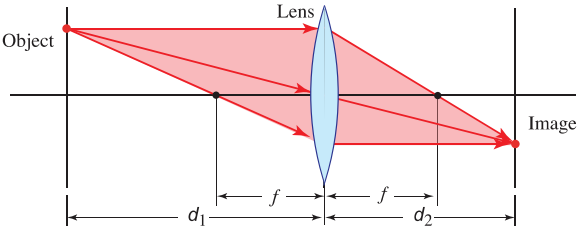


Figure 4.4-1 Rays in a focused imaging system.

Suppose now that the system is not in focus, as illustrated in Fig. 4.4-2, and assume that the focusing error is

$$\epsilon = \frac{1}{d_1} + \frac{1}{d_2} - \frac{1}{f}. \tag{4.4-1}$$

Focusing Error

A point in the object plane generates a patch of light in the image plane that is a shadow of the lens aperture. The distribution of this patch is the system's impulse response function. For simplicity, we shall consider an object point lying on the optical axis and determine the distribution of light $h(x, y)$ it generates in the image plane.

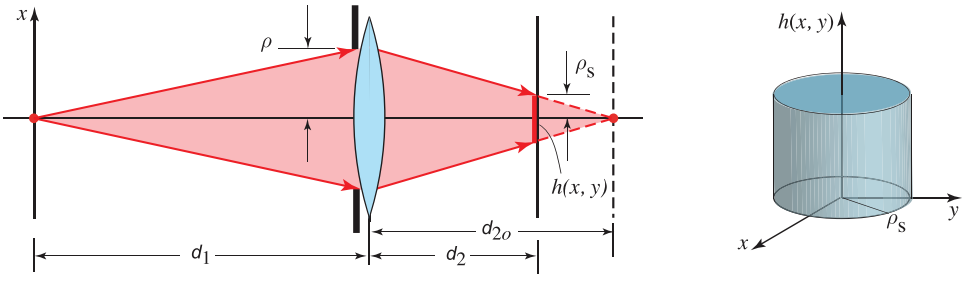


Figure 4.4-2 (a) Rays in a defocused imaging system. (b) The impulse response function of an imaging system with a circular aperture of diameter D is a circle of radius $\rho_s = \epsilon d_2 D/2$, where ϵ is the focusing error.

Assume that the plane of the focused image lies at a distance d_{2o} satisfying the imaging equation $1/d_{2o} + 1/d_1 = 1/f$. The shadow of a point on the edge of the aperture at a radial distance ρ is a point in the image plane with radial distance ρ_s where the ratio $\rho_s/\rho = (d_{2o} - d_2)/d_{2o} = 1 - d_2/d_{2o} = 1 - d_2(1/f - 1/d_1) = 1 - d_2(1/d_2 - \epsilon) = \epsilon d_2$. If $p(x, y)$ is the aperture function, also called the **pupil function** [$p(x, y) = 1$ for points inside the aperture, and 0 elsewhere], then $h(x, y)$ is a scaled version of $p(x, y)$ magnified by a factor $\rho_s/\rho = \epsilon d_2$, so that

$$h(x, y) \propto p\left(\frac{x}{\epsilon d_2}, \frac{y}{\epsilon d_2}\right). \tag{4.4-2}$$

Impulse Response Function (Ray-Optics)

As an example, a circular aperture of diameter D corresponds to an impulse response function confined to a circle of radius

$$\rho_s = \tfrac{1}{2}\epsilon d_2 D, \tag{4.4-3}$$

Blur Spot Radius

as illustrated in Fig. 4.4-2. The radius ρ_s of this "blur spot" is an inverse measure of resolving power and image quality. A small value of ρ_s means that the system is capable of resolving fine details. Since ρ_s is proportional to the aperture diameter D, the image

quality may be improved by use of a small aperture. A small aperture corresponds to a reduced sensitivity of the system to focusing errors, so that it corresponds to an increased "depth of focus."

B. Wave Optics of a 4-*f* Imaging System

Consider now the two-lens imaging system illustrated in Fig. 4.4-3. This system, called the **4-*f* system**, serves as a focused imaging system with unity magnification, as can be easily verified by ray tracing.

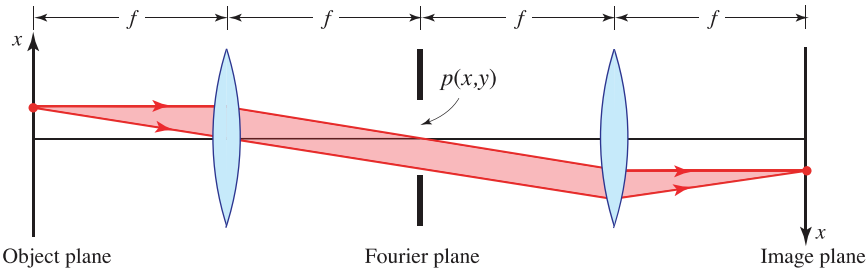


Figure 4.4-3 The 4-f imaging system. If an inverted coordinate system is used in the image plane, the magnification is unity.

The analysis of wave propagation through this system becomes simple if we recognize it as a cascade of two Fourier-transforming subsystems. The first subsystem (between the object plane and the Fourier plane) performs a Fourier transform, and the second (between the Fourier plane and the image plane) performs an inverse Fourier transform since the coordinate system in the image plane is inverted (see Exercise 4.2-2). As a result, in the absence of an aperture the image is a perfect replica of the object.

Let $f(x, y)$ be the complex amplitude transmittance of a transparency placed in the object plane and illuminated by a plane wave $\exp(-jkz)$ traveling in the z direction, as illustrated in Fig. 4.4-4, and let $g(x, y)$ be the complex amplitude in the image plane. The first lens system analyzes $f(x, y)$ into its spatial Fourier transform and separates its Fourier components so that each point in the Fourier plane corresponds to a single spatial frequency. These components are then recombined by the second lens system and the object distribution is perfectly reconstructed.

The 4-f imaging system can be used as a spatial filter in which the image $g(x, y)$ is a filtered version of the object $f(x, y)$. Since the Fourier components of $f(x, y)$ are available in the Fourier plane, a mask may be used to adjust them selectively, blocking some components and transmitting others, as illustrated in Fig. 4.4-5. The Fourier component of $f(x, y)$ at the spatial frequency (ν_x, ν_y) is located in the Fourier plane at the point $x = \lambda f \nu_x, y = \lambda f \nu_y$. To implement a filter of transfer function $H(\nu_x, \nu_y)$, the complex amplitude transmittance $p(x, y)$ of the mask must be proportional to $H(x/\lambda f, y/\lambda f)$. Thus, the transfer function of the filter realized by a mask of transmittance $p(x, y)$ is

$$H(\nu_x, \nu_y) = p(\lambda f \nu_x, \lambda f \nu_y), \tag{4.4-4}$$

Transfer Function 4-f System

where we have ignored the phase factor $j \exp(-j2kf)$ associated with each Fourier transform operation [the argument of h_l in (4.2-8)]. The Fourier transforms $G(\nu_x, \nu_y)$ and $F(\nu_x, \nu_y)$ of $g(x, y)$ and $f(x, y)$ are related by $G(\nu_x, \nu_y) = H(\nu_x, \nu_y)F(\nu_x, \nu_y)$.

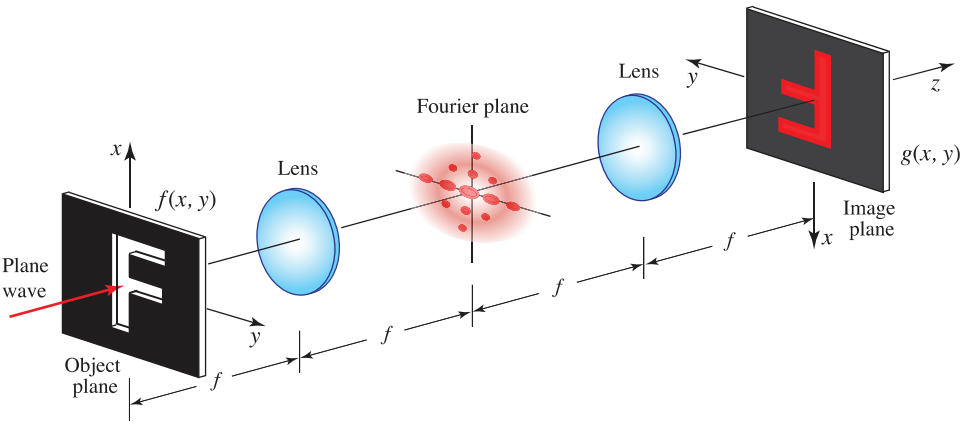


Figure 4.4-4 The 4-f imaging system performs a Fourier transform followed by an inverse Fourier transform, so that the image is a perfect replica of the object.

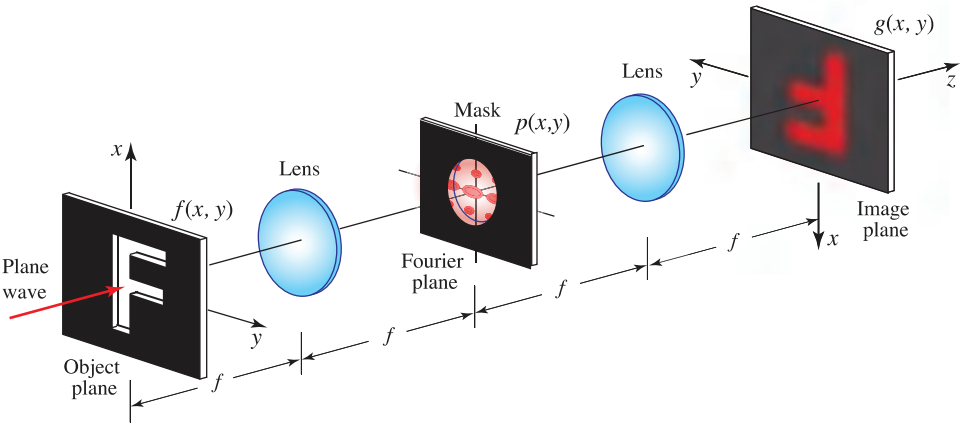


Figure 4.4-5 Spatial filtering. The transparencies in the object and Fourier planes have complex amplitude transmittances $f(x, y)$ and $p(x, y)$. A plane wave traveling in the z direction is modulated by the object transparency, Fourier transformed by the first lens, multiplied by the transmittance of the mask in the Fourier plane, and inverse Fourier transformed by the second lens. As a result, the complex amplitude in the image plane $g(x, y)$ is a filtered version of $f(x, y)$. The system has a transfer function $H(\nu_x, \nu_y) = p(\lambda f \nu_x, \lambda f \nu_y)$.

This is a rather simple result. *The transfer function has the same shape as the pupil function.* The corresponding impulse response function $h(x, y)$ is the inverse Fourier transform of $H(\nu_x, \nu_y)$,

$$h(x, y) = \frac{1}{(\lambda f)^2} P\left(\frac{x}{\lambda f}, \frac{y}{\lambda f}\right), \tag{4.4-5}$$

Impulse Response Function
4-f System

where $P(\nu_x, \nu_y)$ is the Fourier transform of $p(x, y)$.

Examples of Spatial Filters

- The ideal circularly symmetric *low-pass filter* has a transfer function $H(\nu_x, \nu_y) = 1$, $\nu_x^2 + \nu_y^2 < \nu_s^2$ and $H(\nu_x, \nu_y) = 0$, otherwise. It passes spatial frequencies that are smaller than the cutoff frequency ν_s and blocks higher frequencies. This filter is implemented by a mask in the form of a circular aperture of diameter D, with $D/2 = \nu_s \lambda f$. For example, if $D = 2$ cm, $\lambda = 1$ μm, and $f = 100$ cm, the cutoff frequency (spatial bandwidth) $\nu_s = D/2\lambda f = 10$ lines/mm. This filter eliminates spatial frequencies that are greater than 10 lines/mm, so that the smallest size of discernible detail in the filtered image is approximately 0.1 mm.
- The *high-pass filter* is the complement of the low-pass filter. It blocks low frequencies and transmits high frequencies. The mask is a clear transparency with an opaque central circle. The filter output is high at regions of large rate of change and small at regions of smooth or slow variation of the object. The filter is therefore useful for edge enhancement in image-processing applications.
- The *vertical-pass filter* blocks horizontal frequencies and transmits vertical frequencies. Only variations in the x direction are transmitted. If the mask is a vertical slit of width D, the highest transmitted frequency is $\nu_y = (D/2)/\lambda f$.

Examples of these filters and their effects on images are illustrated in Fig. 4.4-6.

C. Wave Optics of a Single-Lens Imaging System

We now consider image formation in the single-lens imaging system illustrated in Fig. 4.4-7, using a wave-optics approach. We first determine the impulse response function, and then derive the transfer function. These functions are determined by the defocusing error ϵ, given by (4.4-1), and by the pupil function $p(x, y)$ (the transmittance of the aperture in the lens plane). The pupil function in this single-lens imaging system plays the same role of the mask function in the 4-f imaging system described in the previous section.

Impulse Response Function

To determine the impulse response function we consider an object composed of a single point (an impulse) on the optical axis at the point $(0, 0)$, and follow the emitted optical wave as it travels to the image plane. The resultant complex amplitude is the impulse response function $h(x, y)$.

An impulse in the object plane produces in the aperture plane a spherical wave approximated by [see (4.1-18)]

$$U(x, y) \approx h_1 \exp\left[-jk\frac{x^2 + y^2}{2d_1}\right], \tag{4.4-6}$$

where $h_1 = (j/\lambda d_1)\exp(-jkd_1)$. Upon crossing the aperture and the lens, $U(x, y)$ is multiplied by the pupil function $p(x, y)$ and the lens quadratic phase factor $\exp[jk(x^2 + y^2)/2f]$, becoming

$$U_1(x, y) = U(x, y)\exp\left(jk\frac{x^2 + y^2}{2f}\right)p(x, y). \tag{4.4-7}$$

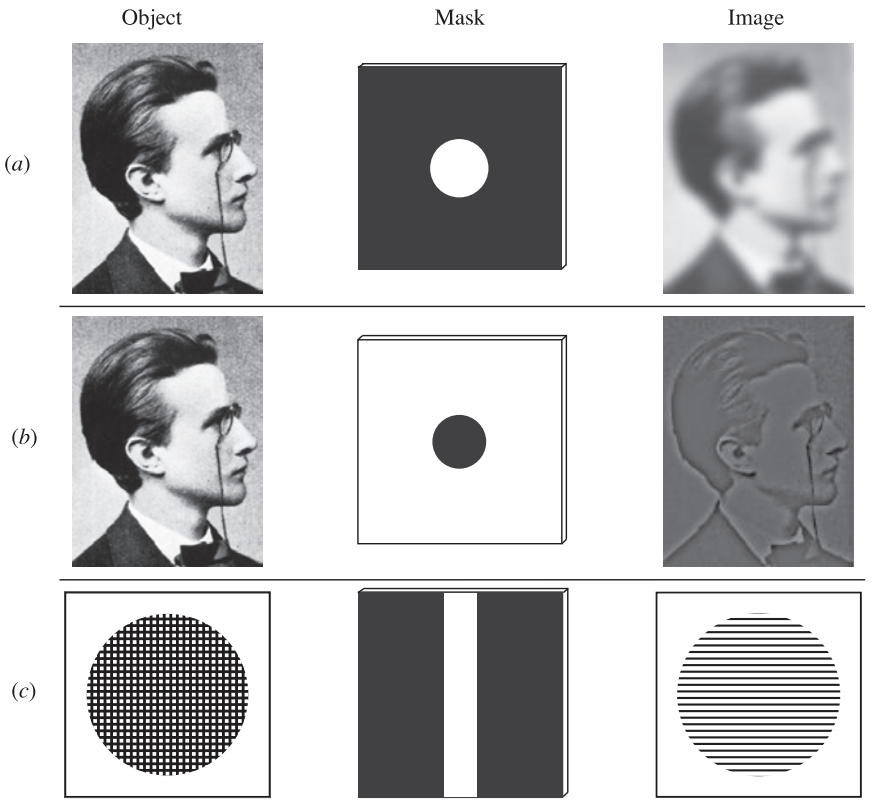


Figure 4.4-6 Examples of object, mask, and filtered image for three spatial filters: (*a*) low-pass filter; (*b*) high-pass filter; (*c*) vertical-pass filter. Black means the transmittance is zero and white means the transmittance is unity.

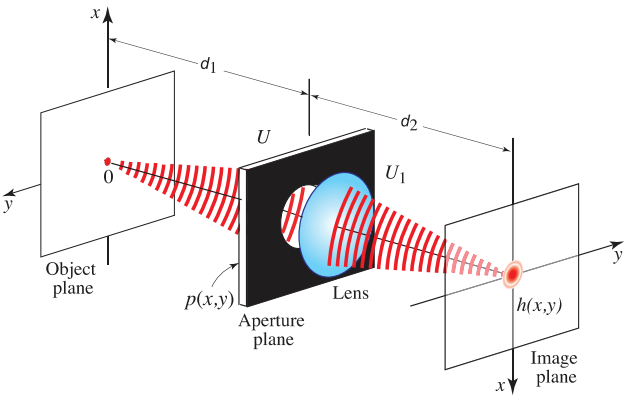


Figure 4.4-7 Single-lens imaging system.

The resultant field $U_1(x, y)$ then propagates in free space a distance d_2. In accordance

with (4.1-20) it produces the impulse response function

$$h(x,y) = h_2 \iint_{-\infty}^{\infty} U_1(x',y') \exp\left[-j\pi \frac{(x-x')^2 + (y-y')^2}{\lambda d_2}\right] dx'\,dy', \qquad (4.4\text{-}8)$$

where $h_2 = (j/\lambda d_2)\exp(-jkd_2)$. Substituting from (4.4-6) and (4.4-7) into (4.4-8) and casting the integrals as a Fourier transform, we obtain

$$h(x,y) = h_1 h_2 \exp\left(-j\pi \frac{x^2+y^2}{\lambda d_2}\right) P_1\left(\frac{x}{\lambda d_2}, \frac{y}{\lambda d_2}\right), \qquad (4.4\text{-}9)$$

where $P_1(\nu_x, \nu_y)$ is the Fourier transform of the function

$$\boxed{p_1(x,y) = p(x,y)\exp\left(-j\pi\epsilon \frac{x^2+y^2}{\lambda}\right),} \qquad (4.4\text{-}10)$$

Generalized Pupil Function

known as the **generalized pupil function**. The factor ϵ is the focusing error given by (4.4-1).

For a high-quality imaging system, the impulse response function is a narrow function, extending only over a small range of values of x and y. If the phase factor $\pi(x^2+y^2)/\lambda d_2$ in (4.4-9) is much smaller than 1 for all x and y within this range, it can be neglected, so that

$$\boxed{h(x,y) = h_0 P_1\left(\frac{x}{\lambda d_2}, \frac{y}{\lambda d_2}\right),} \qquad (4.4\text{-}11)$$

Impulse Response Function

where $h_0 = h_1 h_2$ is a constant of magnitude $(1/\lambda d_1)(1/\lambda d_2)$. It follows that the system's impulse response function is proportional to the Fourier transform of the generalized pupil function $p_1(x,y)$ evaluated at $\nu_x = x/\lambda d_2$ and $\nu_y = y/\lambda d_2$.

If the system is focused ($\epsilon = 0$), then $p_1(x,y) = p(x,y)$, and

$$h(x,y) \approx h_0 P\left(\frac{x}{\lambda d_2}, \frac{y}{\lambda d_2}\right), \qquad (4.4\text{-}12)$$

where $P(\nu_x, \nu_y)$ is the Fourier transform of $p(x,y)$. This result is similar to the corresponding result in (4.4-5) for the 4-f system.

EXAMPLE 4.4-1. ***Impulse Response Function of a Focused Imaging System with a Circular Aperture.*** If the aperture is a circle of diameter D so that $p(x,y) = 1$ if $\rho = \sqrt{x^2+y^2} \le D/2$, and zero otherwise, then the impulse response function is

$$h(x,y) = h(0,0)\frac{2J_1(\pi D\rho/\lambda d_2)}{\pi D\rho/\lambda d_2}, \qquad \rho = \sqrt{x^2+y^2}, \qquad (4.4\text{-}13)$$

and $|h(0,0)| = (\pi D^2/4\lambda^2 d_1 d_2)$. This is a circularly symmetric function whose cross section is shown in Fig. 4.4-8. It drops to zero at a radius

$$\rho_s = 1.22\lambda\frac{d_2}{D} \qquad (4.4\text{-}14)$$

and oscillates slightly before it vanishes. The radius ρ_s is therefore a measure of the size of the blur circle. If the system is focused at ∞, then $d_1 = \infty$ and $d_2 = f$, so that

$$\rho_s = 1.22\lambda F_{\#}\,, \tag{4.4-15}$$
Spot Radius

where $F_{\#} = f/D$ is the lens F-number. Thus, systems of smaller $F_{\#}$ (larger apertures) have better image quality. This assumes, of course, that the larger lens does not introduce geometrical aberrations.

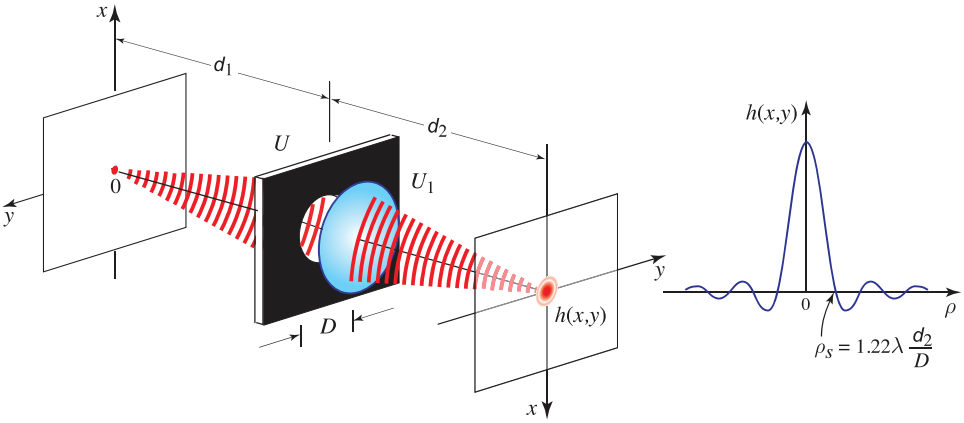


Figure 4.4-8 Impulse response function of an imaging system with a circular aperture.

Transfer Function

The transfer function of a linear system can only be defined when the system is shift invariant (see Appendix B). Evidently, the single-lens imaging system is not shift invariant since a shift Δ of a point in the object plane is accompanied by a *different* shift $M\Delta$ in the image plane, where $M = -d_2/d_1$ is the magnification.

The image is different from the object in two ways. First, the image is a magnified replica of the object, i.e., the point (x, y) of the object is located at a new point (Mx, My) in the image. Second, every point is smeared into a patch as a result of defocusing or diffraction. We can therefore think of image formation as a cascade of two systems — a system of ideal magnification followed by a system of blur, as depicted in Fig. 4.4-9. By its nature, the magnification system is shift-variant. For points near the optical axis, the blur system is approximately shift invariant and therefore can be described by a transfer function.

The transfer function $H(\nu_x, \nu_y)$ of the blur system is determined by obtaining the Fourier transform of the impulse response function $h(x, y)$ in (4.4-11). The result is

$$H(\nu_x, \nu_y) \approx p_1(\lambda d_2\nu_x, \lambda d_2\nu_y), \tag{4.4-16}$$
Transfer Function

where $p_1(x, y)$ is the generalized pupil function and we have ignored a constant phase factor $\exp(-jkd_1)\exp(-jkd_2)$. If the system is focused, then

$$H(\nu_x, \nu_y) \approx p(\lambda d_2\nu_x, \lambda d_2\nu_y), \tag{4.4-17}$$

where $p(x, y)$ is the pupil function. This result is identical to that obtained for the 4-f imaging system [see (4.4-4)]. If the aperture is a circle of diameter D, for example,

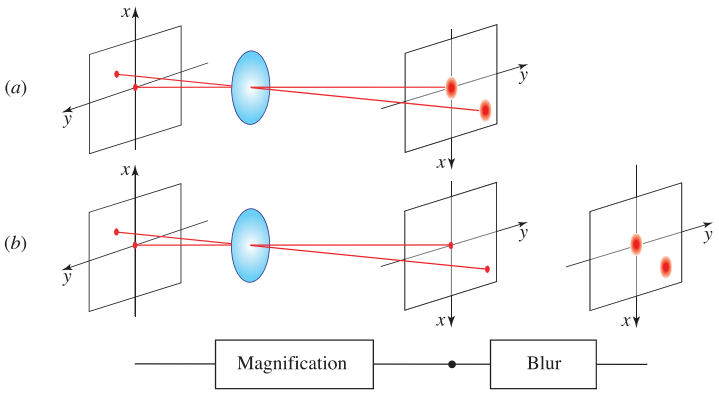


Figure 4.4-9 The imaging system in (a) is regarded in (b) as a combination of an ideal imaging system with only magnification, followed by shift-invariant blur in which each point is blurred into a patch with a distribution equal to the impulse response function.

then the transfer function is constant within a circle of radius ν_s, where

$$\nu_s = \frac{D}{2\lambda d_2}, \tag{4.4-18}$$

and vanishes elsewhere, as illustrated in Fig. 4.4-10.

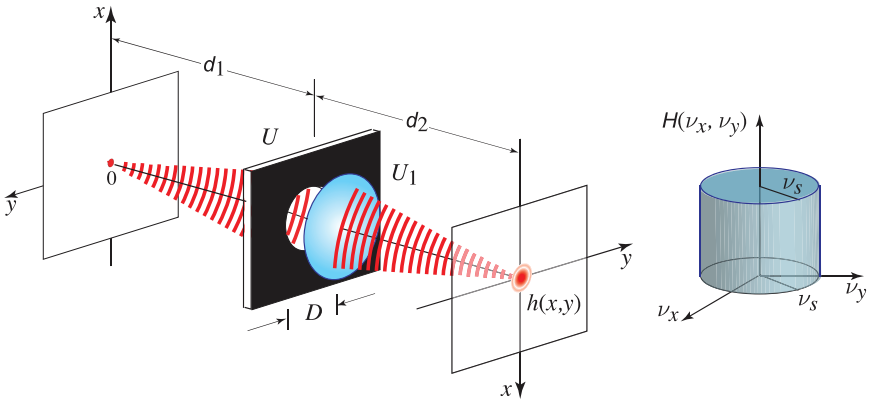


Figure 4.4-10 Transfer function of a focused imaging system with a circular aperture of diameter D. The system has a spatial bandwidth $\nu_s = D/2\lambda d_2$.

If the lens is focused at infinity, i.e., $d_2 = f$,

$$\boxed{\nu_s = \frac{1}{2\lambda F_\#},} \tag{4.4-19}$$

Spatial Bandwidth

where $F_\# = f/D$ is the lens F-number. For example, for an F-2 lens ($F_\# = f/D = 2$) and for $\lambda = 0.5\ \mu$m, $\nu_s = 500$ lines/mm. The frequency ν_s is the spatial bandwidth, i.e., the highest spatial frequency that the imaging system can transmit.

D. Near-Field Imaging

It was shown in Sec. 4.1B that the spatial bandwidth of light propagating in free space at a wavelength λ is λ^{-1} cycles/mm. Fourier components of the object distribution with spatial frequencies greater than λ^{-1} lead to evanescent waves that decay rapidly and diminish at distances from the object plane of the order of a wavelength, so that object features smaller than a wavelength cannot be transmitted. Moreover, it was shown in Sec. 4.4C that an imaging system using a lens with a specified $F_{\#}$ has an impulse response function whose radius is $1.22\lambda F_{\#}$, so that points separated by a distance smaller than $1.22\lambda F_{\#}$ cannot be discriminated [see Fig. 4.4-11(a)]. Another imaging modality that makes use of a laser beam focused by a lens to scan an object, as depicted in Fig. 4.4-11(b), behaves similarly. The resolution of this system is dictated by the size of the focused spot, which has a radius of $1.22\lambda F_{\#}$, as was shown in Example 4.4-1. In both of these cases, therefore, object details with dimensions much smaller than a wavelength are washed out in the scanned image. This fundamental limit on the resolution of image-formation systems is referred to as the **diffraction limit**.

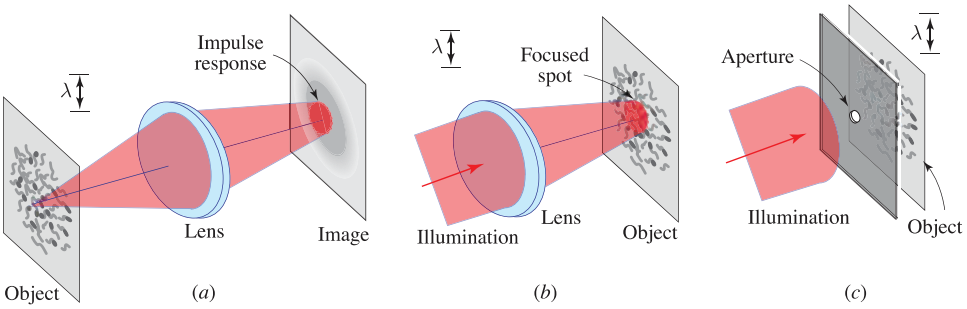


Figure 4.4-11 In a single-lens imaging system, the subwavelength spatial details of an object are washed out (a) in an image formed by a single lens, or (b) by making use of a focused laser-scanning system. (c) On the other hand, a scanning imaging system that makes use of illumination transmitted through a subwavelength aperture preserves the subwavelength details of the object, provided that the object plane is placed at a subwavelength distance from the aperture plane.

The diffraction limit may be circumvented, however. *Light can be localized to a spot with dimensions much smaller than a wavelength, within a single plane*. The difficulty is that the evanescent waves fully decay a short distance away from that plane, whereupon the spot diverges and acquires a size that exceeds the wavelength. At yet greater distances, the wave ultimately becomes a spherical wave. Hence, the diffraction limit can be circumvented if the object is illuminated in the very vicinity of the subwavelength spot, before the beam waist has an opportunity to grow. This may be implemented in a scanning configuration by passing the illumination beam through an aperture of diameter much smaller than a wavelength, as depicted in Fig. 4.4-11(c). The object is placed at a distance from the aperture that is usually less than half the diameter of the aperture so that the beam illuminates a subwavelength-size area of the object. Upon transmission through the object, the traveling components of the wave form a spherical wave whose amplitude is proportional to the object transmittance at the location of the spot illumination. The resolution of this imaging system is therefore of the order of the aperture size, which is much smaller than the wavelength. An image is constructed by raster-scanning the illuminated aperture across the object and recording the optical response via a conventional far-field imaging system. This technique is known as near-field optical imaging or **scanning near-field optical microscopy** (**SNOM**). Subwavelength imaging falls in the domain of **nanophotonics** since the

imaging takes place over a subwavelength (nanometer) spatial scale. Other approaches for achieving subwavelength imaging make use of negative-index and hyperbolic materials, as considered in Sec. 8.1.

SNOM is typically implemented by sending the illumination light through an optical fiber with an aluminum-coated tapered tip, as illustrated in Fig. 4.4-12. The light is guided through the fiber by total internal reflection. As the diameter of the fiber decreases, the light is guided by reflection from the metallic surface, which acts like a conical mirror. As the fiber diameter grows yet smaller in the region of the tip, the wave can no longer be guided (see Sec. 9.1) and becomes evanescent. The distribution of the illumination wave at, and beyond, the end of the tip is complex and is typically determined numerically. Aperture diameters and spatial resolutions of the order of tens of nanometers are usually achieved in SNOM with visible light. Since the tip of the fiber scans the object at a distance of only a few nanometers, an elaborate feedback system must be employed to maintain the distance for a specimen of arbitrary topography. Applications of SNOM include the non-destructive characterization of inorganic, organic, composite, and biological materials and nanostructures.

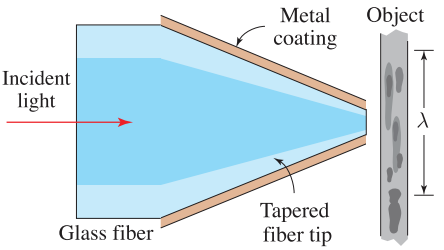


Figure 4.4-12 An optical fiber with a tapered metal-coated tip for near-field imaging.

4.5 HOLOGRAPHY

Holography involves the recording and reconstruction of optical waves. A hologram is a 2D transparency that contains a coded record of the amplitude and phase of an optical wave originating from a 3D object. Consider a monochromatic optical wave whose complex amplitude in some plane, say the $z = 0$ plane, is $U_o(x, y)$. If, somehow, a thin optical element (call it a transparency) with complex amplitude transmittance $\mathrm{t}(x, y)$ equal to $U_o(x, y)$ were able to be made, it would provide a complete record of the wave. The wave could then be reconstructed simply by illuminating the transparency with a uniform plane wave of unit amplitude traveling in the z direction. The transmitted wave would have a complex amplitude in the $z = 0$ plane $U(x, y) = 1 \cdot \mathrm{t}(x, y) = U_o(x, y)$. The original wave would then be reproduced at all points in the $z = 0$ plane, and therefore reconstructed everywhere in the space $z > 0$.

As an example, we know that a uniform plane wave traveling at an angle θ with respect to the z axis in the x–z plane has a complex amplitude $U_o(x, y) = \exp[-jkx \sin\theta]$. A record of this wave would be a transparency with complex amplitude transmittance $\mathrm{t}(x, y) = \exp[-jkx \sin\theta]$. Such a transparency acts as a prism that bends an incident plane wave $\exp(-jkz)$ by an angle θ (see Sec. 2.4B), thus reproducing the original wave.

The question is how to make a transparency $\mathrm{t}(x, y)$ from the original wave $U_o(x, y)$. One key impediment is that optical detectors, including the photographic emulsions used to make transparencies, are responsive to the optical intensity, $|U_o(x, y)|^2$, and are therefore insensitive to the phase $\arg\{U_o(x, y)\}$. Phase information is obviously important and cannot be disregarded, however. For example, if the phase of the oblique wave $U_o(x, y) = \exp[-jkx \sin\theta]$ were not recorded, neither would the direction of

travel of the wave. To record the phase of $U_o(x, y)$, a code must be found that transforms phase into intensity. The recorded information could then be optically decoded in order to reconstruct the wave.

The Holographic Code

The holographic code is based on mixing the original wave (hereafter called the **object wave**) U_o with a known **reference wave** U_r and recording their interference pattern in the $z = 0$ plane. The intensity of the sum of the two waves is photographically recorded and a transparency of complex amplitude transmittance t, proportional to the intensity, is made [Fig. 4.5-1(a)]. The transmittance is therefore given by

$$\begin{aligned} \mathrm{t} &\propto |U_o + U_r|^2 = |U_r|^2 + |U_o|^2 + U_r^* U_o + U_r U_o^*, \\ &= I_r + I_o + U_r^* U_o + U_r U_o^*, \\ &= I_r + I_o + 2\sqrt{I_r I_o}\, \cos\left[\arg\{U_r\} - \arg\{U_o\}\right], \end{aligned} \tag{4.5-1}$$

where I_r and I_o are, respectively, the intensities of the reference wave and the object wave at the $z = 0$ plane.

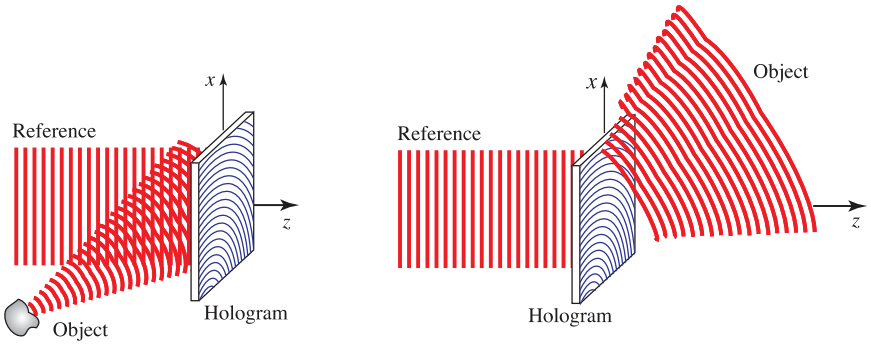


Figure 4.5-1 (a) A hologram is a transparency on which the interference pattern between the original wave (object wave) and a reference wave is recorded. (b) The original wave is reconstructed by illuminating the hologram with the reference wave.

The transparency, called a **hologram**, clearly carries coded information pertinent to the magnitude and phase of the wave U_o. In fact, as an interference pattern the transmittance t is highly sensitive to the difference between the phases of the two waves, as was shown in Sec. 2.5 (the temporal analog to holography is heterodyning, discussed in Sec. 2.6). As indicated above, ordinary photography is responsive only to the intensity of the incident wave and therefore records no phase information.

To decode the information in the hologram and reconstruct the object wave, the reference wave U_r is again used to illuminate the hologram [Fig. 4.5-1(b)]. The result is a wave with complex amplitude

$$U = \mathrm{t}U_r \propto U_r I_r + U_r I_o + I_r U_o + U_r^2 U_o^* \tag{4.5-2}$$

in the hologram plane $z = 0$. The third term on the right-hand side is the original wave multiplied by the intensity I_r of the reference wave. If I_r is uniform (independent of x and y), this term constitutes the desired reconstructed wave. But it must be separated from the other three terms. The fourth term is a conjugated version of the original wave modulated by U_r^2. The first two terms represent the reference wave, modulated by the sum of the intensities of the two waves.

If the reference wave is selected to be a uniform plane wave propagating along the z axis $\sqrt{I_r}\exp(-jkz)$, then in the $z = 0$ plane $U_r(x,y) = \sqrt{I_r}$ is a constant independent of x and y. Dividing (4.5-2) by $U_r = \sqrt{I_r}$ gives

$$U(x,y) \propto I_r + I_o(x,y) + \sqrt{I_r}\,U_o(x,y) + \sqrt{I_r}\,U_o^*(x,y). \tag{4.5-3}$$

Reconstructed Wave in Plane of Hologram

The significance of the various terms in (4.5-3), and the methods of extracting the original wave (the third term), are clarified by means of a number of examples.

EXAMPLE 4.5-1. ***Hologram of an Oblique Plane Wave.*** If the object wave is an oblique plane wave at angle θ [Fig. 4.5-2(a)], $U_o(x,y) = \sqrt{I_o}\exp(-jkx\sin\theta)$, then (4.5-3) gives $U(x,y) \propto I_r + I_o + \sqrt{I_r I_o}\exp(-jkx\sin\theta) + \sqrt{I_r I_o}\exp(-jkx\sin\theta)$. Since the first two terms are constant, they correspond to a wave propagating in the z direction (the continuance of the reference wave). The third term corresponds to the original object wave, whereas the fourth term represents the **conjugate wave**, a plane wave traveling at an angle $-\theta$. The object wave is therefore separable from the other waves. In fact, this hologram is nothing but a recording of the interference pattern formed from two oblique plane waves at an angle θ (Sec. 2.5A). It serves as a sinusoidal diffraction grating that splits an incident reference wave into three waves at angles 0, θ, and $-\theta$ [see Fig. 4.5-2(b) and Sec. 2.4B].

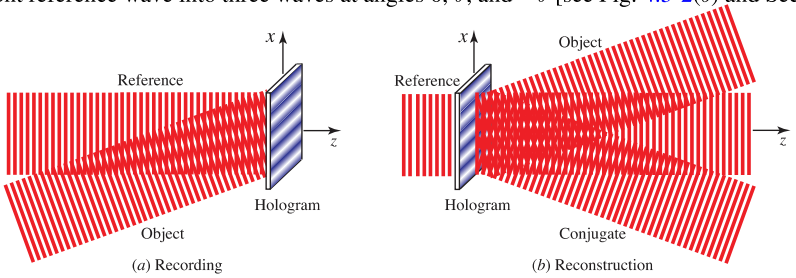


Figure 4.5-2 The hologram of an oblique plane wave is a sinusoidal diffraction grating.

EXAMPLE 4.5-2. ***Hologram of a Point Source.*** Here the object wave is a spherical wave originating at the point $\mathbf{r}_0 = (0,0,-d)$, as illustrated in Fig. 4.5-3, so that $U_o(x,y) \propto \exp(-jk|\mathbf{r}-\mathbf{r}_0|)/|\mathbf{r}-\mathbf{r}_0|$, where $\mathbf{r} = (x,y,0)$. The first term of (4.5-3) corresponds to a plane wave traveling in the z direction, whereas the third is proportional to the amplitude of the original spherical wave originating at $(0,0,-d)$. The fourth term is proportional to the amplitude of the conjugate wave $U_o^*(x,y) \propto \exp(jk|\mathbf{r}-\mathbf{r}_0|)/|\mathbf{r}-\mathbf{r}_0|$, which is a converging spherical wave centered at the point $(0,0,d)$. The second term is proportional to $1/|\mathbf{r}-\mathbf{r}_0|^2$ and its corresponding wave therefore travels in the z direction with very small angular spread since its intensity varies slowly in the transverse plane.

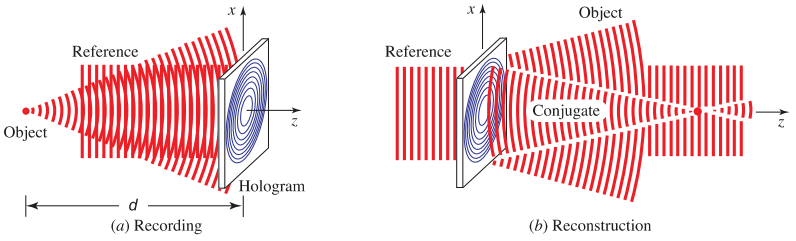


Figure 4.5-3 Hologram of a spherical wave originating from a point source. The conjugate wave forms a real image of the point.

Off-Axis Holography

One means of separating the four components of the reconstructed wave is to ensure that they vary at well-separated spatial frequencies, so that they have well-separated directions. This form of spatial-frequency multiplexing (see Sec. 4.1A) is assured if the object and reference waves are offset so that they arrive from well-separated directions.

Let us consider the case when the object wave has a complex amplitude $U_o(x, y) = f(x, y) \exp(-jkx \sin\theta)$. This is a wave of complex envelope $f(x, y)$ modulated by a phase factor equal to that introduced by a prism with deflection angle θ. It is assumed that $f(x, y)$ varies slowly so that its maximum spatial frequency ν_s corresponds to an angle $\theta_s = \sin^{-1} \lambda\nu_s$ much smaller than θ. The object wave therefore has directions centered about the angle θ, as illustrated in Fig. 4.5-4. Equation (4.5-3) gives

$$U(x, y) \propto I_r + |f(x, y)|^2 + \sqrt{I_r}\, f(x, y) \exp(-jkx \sin\theta) + \sqrt{I_r}\, f^*(x, y) \exp(+jkx \sin\theta). \tag{4.5-4}$$

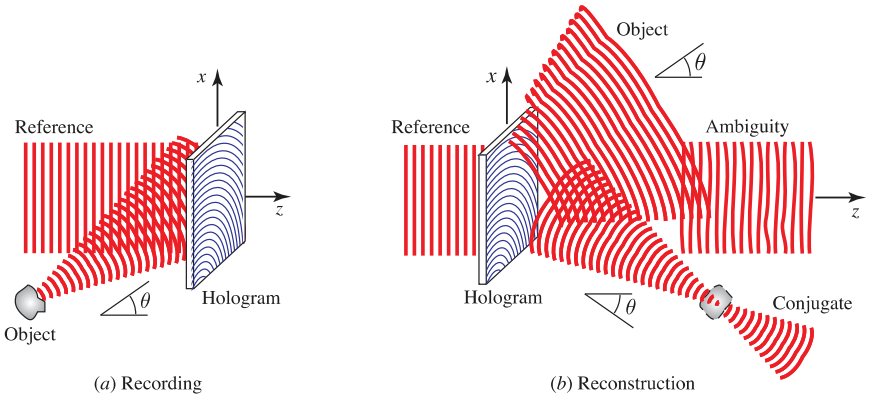


Figure 4.5-4 Hologram of an off-axis object wave. The object wave is separated from both the reference and conjugate waves.

The third term is evidently a replica of the object wave, which arrives from a direction at angle θ. The presence of the phase factor $\exp(+jkx \sin\theta)$ in the fourth term indicates that it is deflected in the $-\theta$ direction. The first term corresponds to a plane wave traveling in the z direction. The second term, usually known as the **ambiguity term**, corresponds to a nonuniform plane wave in directions within a cone of small angle $2\theta_s$ around the z direction. The offset of the directions of the object and reference waves results in a natural angular separation of the object and conjugate waves from each other and from the other two waves if $\theta > 3\theta_s$, thus allowing the original wave to be recovered unambiguously. An alternative method of reducing the effect of the ambiguity wave is to make the intensity of the reference wave much greater than that of the object wave. The ambiguity wave [second term of (4.5-3)] is then much smaller than the other terms since it involves only object waves; it is therefore relatively negligible.

Fourier-Transform Holography

The Fourier transform $F(\nu_x, \nu_y)$ of a function $f(x, y)$ may be computed optically by use of a lens (see Sec. 4.2). If $f(x, y)$ is the complex amplitude in one focal plane of the lens, then $F(x/\lambda f, y/\lambda f)$ is the complex amplitude in the other focal plane, where

f is the focal length of the lens and λ is the wavelength. Since the Fourier transform is usually a complex-valued function, it cannot be recorded directly.

The Fourier transform $F(x/\lambda f, y/\lambda f)$ may be recorded holographically by regarding it as an object wave, $U_o(x, y) = f(x/\lambda f, y/\lambda f)$, mixing it with a reference wave $U_f(x, y)$, and recording the superposition as a hologram [Fig. 4.5-5(a)]. Reconstruction is achieved by illumination of the hologram with the reference wave as usual. The reconstructed wave may be inverse Fourier transformed using a lens so that the original function $f(x, y)$ is recovered [Fig. 4.5-5(b)].

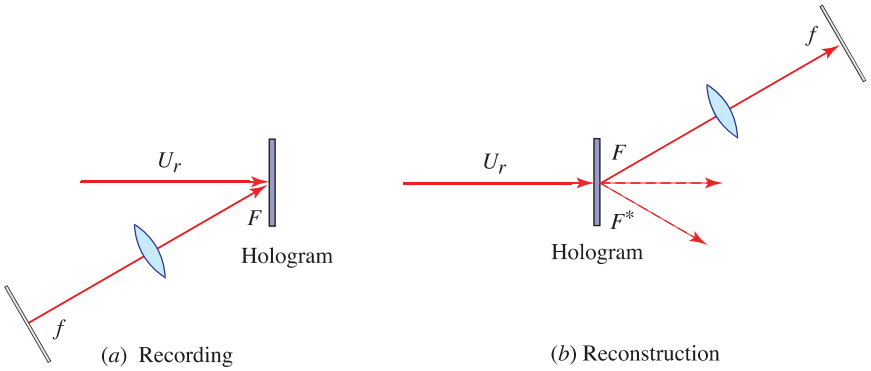


Figure 4.5-5 (a) Hologram of a wave whose complex amplitude represents the Fourier transform of a function $f(x, y)$. (b) Reconstruction of $f(x, y)$ by use of a Fourier-transform lens.

Holographic Spatial Filters

A spatial filter of transfer function $H(\nu_x, \nu_y)$ may be implemented by use of a 4-f optical system with a mask of complex amplitude transmittance $p(x, y) = H(x/\lambda f, y/\lambda f)$ placed in the Fourier plane (see Sec. 4.4B). Since the transfer function $H(\nu_x, \nu_y)$ is usually complex-valued, the mask transmittance $p(x, y)$ has a phase component and is difficult to fabricate using conventional printing techniques. If the filter impulse response function $h(x, y)$ is real-valued, however, a Fourier-transform hologram of $h(x, y)$ may be created by holographically recording the Fourier transform $U_o(x, y) = H(x/\lambda f, y/\lambda f)$, as depicted in Fig. 4.5-6(a).

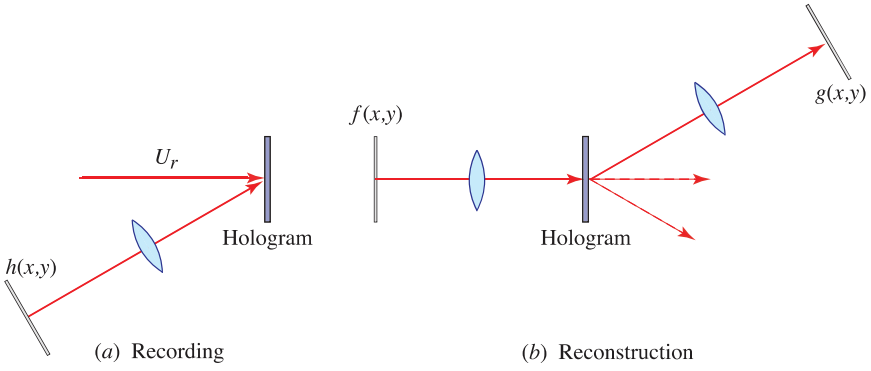


Figure 4.5-6 The VanderLugt holographic filter. (a) A hologram of the Fourier transform of $h(x, y)$ is recorded. (b) The Fourier transform of $f(x, y)$ is transmitted through the hologram and inverse Fourier transformed by a lens. The result is a function $g(x, y)$ proportional to the convolution of $f(x, y)$ and $h(x, y)$. The overall process provides a spatial filter with impulse response function $h(x, y)$.

Using the Fourier transform of the input $f(x, y)$ as a reference, $U_r(x, y) = F(x/\lambda f, y/\lambda f)$, the hologram constructs the wave

$$U_r(x, y)U_o(x, y) = F(x/\lambda f, y/\lambda f)\, H(x/\lambda f, y/\lambda f). \tag{4.5-5}$$

The inverse Fourier transform of the reconstructed object wave, obtained with a lens of focal length f as illustrated in Fig. 4.5-6(b), therefore yields a complex amplitude $g(x, y)$ with a Fourier transform $G(\nu_x, \nu_y) = H(\nu_x, \nu_y)F(\nu_x, \nu_y)$. Thus, $g(x, y)$ is the convolution of $f(x, y)$ with $h(x, y)$. The overall system, known as the **VanderLugt filter**, performs the operation of convolution, which is the basis of spatial filtering.

If the conjugate wave $U_r(x, y)U_o^*(x, y) = F(x/\lambda f, y/\lambda f)H^*(x/\lambda f, y/\lambda f)$ is, instead, inverse Fourier transformed, the correlation, instead of the convolution, of the functions $f(x, y)$ and $h(x, y)$ is obtained. The operation of correlation is useful in image-processing applications, including pattern recognition.

The Holographic Apparatus

An essential condition for the successful fabrication of a hologram is the availability of a monochromatic light source with minimal phase fluctuations. The presence of phase fluctuations results in the random shifting of the interference pattern and the washing out of the hologram. For this reason, a coherent light source (usually a laser) is a necessary part of the apparatus. The coherence requirements for the interference of light waves are discussed in Chapter 12.

Figure 4.5-7 illustrates a typical experimental configuration used to record a hologram and reconstruct the optical wave scattered from the surface of a physical object. Using a beamsplitter, laser light is split into two portions; one is used as the reference wave, whereas the other is scattered from the object to form the object wave. The optical path difference between the two waves should be as small as possible to ensure that the two beams maintain a nonrandom phase difference [the term $\arg\{U_r\} - \arg\{U_o\}$ in (4.5-1)].

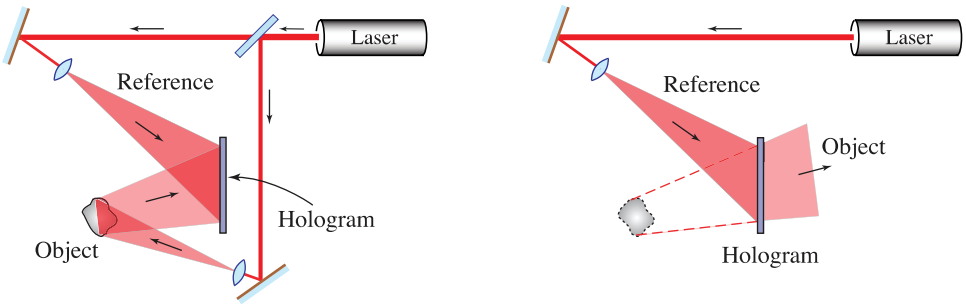


Figure 4.5-7 Holographic recording and reconstruction.

Since the interference pattern forming the hologram is composed of fine lines separated by distances of the order of $\lambda/\sin\theta$, where θ is the angular offset between the reference and object waves, the photographic film must be of high resolution and the system must not vibrate during the exposure. The larger θ, the smaller the distances between the hologram lines, and the more stringent these requirements are. The object wave is reconstructed when the recorded hologram is illuminated with the reference wave, so that a viewer see the object as if it were actually there, with its three-dimensional character preserved.

Volume Holography

It has been assumed so far that the hologram is a thin planar transparency on which the interference pattern of the object and reference waves is recorded. We now consider recording the hologram in a relatively thick medium and show that this offers an advantage. Consider the simple case when the object and reference waves are plane waves with wavevectors $\mathbf{k}_r$ and $\mathbf{k}_o$. The recording medium extends between the planes $z = 0$ and $z = \Delta$, as illustrated in Fig. 4.5-8. The interference pattern is now a function of x, y, and z:

$$\begin{aligned} I(x,y,z) &= \left| \sqrt{I_r}\, \exp(-j\mathbf{k}_r \cdot \mathbf{r}) + \sqrt{I_o}\, \exp(-j\mathbf{k}_o \cdot \mathbf{r}) \right|^2 \\ &= I_r + I_o + 2\sqrt{I_r I_o}\, \cos(\mathbf{k}_o \cdot \mathbf{r} - \mathbf{k}_r \cdot \mathbf{r}) \\ &= I_r + I_o + 2\sqrt{I_r I_o}\, \cos(\mathbf{k}_g \cdot \mathbf{r}), \end{aligned} \tag{4.5-6}$$

where $\mathbf{k}_g = \mathbf{k}_o - \mathbf{k}_r$. This is a sinusoidal pattern of period $\Lambda = 2\pi/|\mathbf{k}_g|$ and with the surfaces of constant intensity normal to the vector $\mathbf{k}_g$.

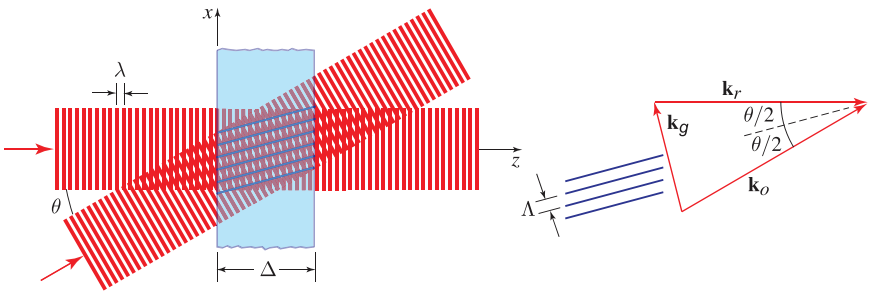


Figure 4.5-8 Interference pattern when the reference and object waves are plane waves. Since $|\mathbf{k}_r| = |\mathbf{k}_o| = 2\pi/\lambda$ and $|\mathbf{k}_g| = 2\pi/\Lambda$, from the geometry of the vector diagram $2\pi/\Lambda = 2(2\pi/\lambda)\sin(\theta/2)$, so that $\Lambda = \lambda/2\sin(\theta/2)$.

For example, if the reference wave points in the z direction and the object wave makes an angle θ with the z axis, $|\mathbf{k}_g| = 2k\sin(\theta/2)$ and the period is

$$\Lambda = \frac{\lambda}{2\sin(\theta/2)} \tag{4.5-7}$$

as illustrated in Fig. 4.5-8.

If recorded in emulsion, this pattern serves as a thick diffraction grating, a **volume hologram**. The vector $\mathbf{k}_g$ is called the **grating vector**. When illuminated with the reference wave as illustrated in Fig. 4.5-9, the parallel planes of the grating reflect the wave only when the Bragg condition $\sin\phi = \lambda/2\Lambda$ is satisfied, where ϕ is the angle between the planes of the grating and the incident reference wave (Exercise 2.5-3). In our case $\phi = \theta/2$, so that $\sin(\theta/2) = \lambda/2\Lambda$. In view of (4.5-7), the Bragg condition is indeed satisfied, so that the reference wave is indeed reflected. As evident from the geometry, the reflected wave is an extension of the object wave, so that the reconstruction process is successful.

Suppose now that the hologram is illuminated with a reference wave of different wavelength λ'. Evidently, the Bragg condition, $\sin(\theta/2) = \lambda'/2\Lambda$, will not be satisfied and the wave will not be reflected. It follows that the object wave is reconstructed only

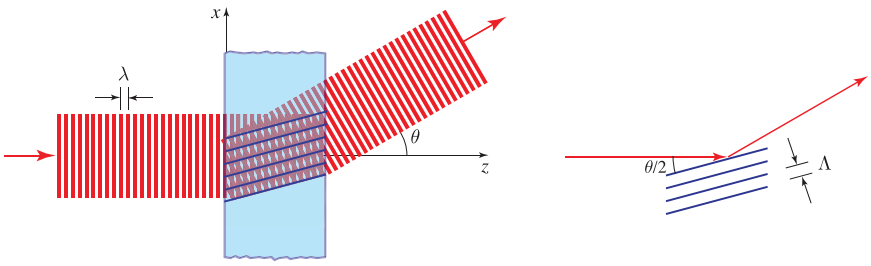
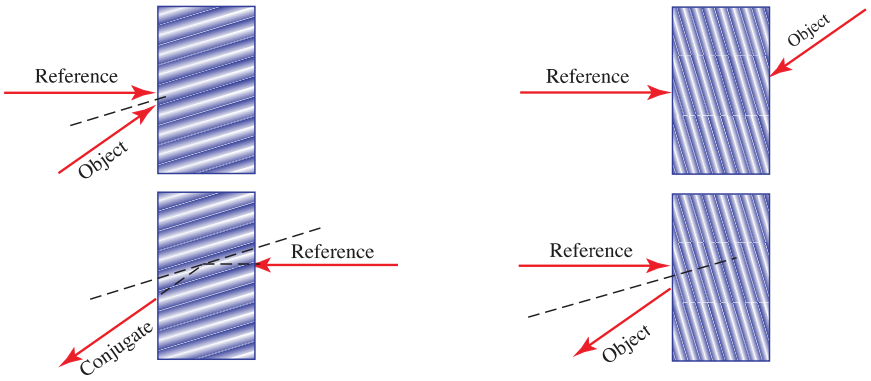


Figure 4.5-9 The reference wave is Bragg reflected from the thick hologram and the object wave is reconstructed.

if the wavelength of the reconstruction source matches that of the recording source. If light with a broad spectrum (white light) is used as a reconstruction source, only the "correct" wavelength would be reflected and the reconstruction process would be successful.

Although the recording process must be done with monochromatic light, the reconstruction can be achieved with white light. This provides a clear advantage in many applications of holography. Other geometries for recording a reconstruction of a volume hologram are illustrated in Fig. 4.5-10.



(*a*) Transmission hologram (*b*) Reflection hologram

Figure 4.5-10 Two geometries for recording and reconstruction of a volume hologram. (*a*) This hologram is recorded with the reference and object waves arriving from the same side, and is reconstructed by use of a reversed reference wave; the reconstructed wave is a conjugate wave traveling in a direction opposite to the original object wave. (*b*) A reflection hologram is recorded with the reference and object waves arriving from opposite sides; the object wave is reconstructed by reflection from the grating.

Another type of hologram that may be viewed with white light is the **rainbow hologram**. This hologram is recorded through a narrow slit so that the reconstructed image, of course, also appears as if seen through a slit. However, if the wavelength of reconstruction differs from the recording wavelength, the reconstructed wave will appear to be coming from a displaced slit since a magnification effect will be introduced. If white light is used for reconstruction, the reconstructed wave appears as the object seen through many displaced slits, each with a different wavelength (color). The result is a rainbow of images seen through parallel slits. Each slit displays the object with parallax effect in the direction of the slit, but not in the orthogonal direction. Rainbow holograms enjoy wide commercial use as displays.

Computer-Generated Holography

A **computer-generated hologram** is a hologram of an object that does not physically exist. The hologram is generated by computing, and then digitally recording, the interference pattern of a reference wave with a mathematically defined wave that represents light scattered from a particular virtual object. The hologram may take the form of a mask, film, or spatial light modulator; when illuminated by the reference wave it generates the desired object wave. Computer-generated holography is principally geared toward 3D visualization, including applications in CAD (computer-aided design), gaming, and video displays.

An important application is the generation of **holographic optical elements** (HOEs). One example is the hologram of a point source, described in Example 4.5-2, which functions as a lens. A HOE that converts a planar wave into another optical beam with a mathematically defined complex amplitude, such as a Hermite–Gaussian, Laguerre–Gaussian, Bessel, or Airy beam (see Secs. 3.4 and 3.5), may be created by computing and digitally recording the interference pattern of the desired beam with a planar wave, as described in Example 4.5-3.

EXAMPLE 4.5-3. ***Holographic Optical Element for Generating a Spiral-Phased Wave.*** The task at hand is to create a HOE that converts a reference planar wave into a spiral-phased object wave with complex amplitude $U_o = \exp(-jl\phi)$ at the $z = 0$ plane. Here $\phi = \arctan(y/x)$ is the azimuthal angle [see Fig. 3.4-1(b)] and $l = 1, 2, \ldots$ is the topological charge of the associated optical vortex, as discussed in Sec. 3.4. Choosing a reference planar wave that propagates in the x–z plane, at an angle θ with respect to the z axis (see Fig. 2.5-4), gives rise to $U_r = \exp(-jk_o \sin\theta\, x)$ at the $z = 0$ plane. Making use of (4.5-1) thus yields an interference pattern given by

$$I(x, y) = 2 + 2\cos(2\pi x/\Lambda - l\phi), \quad \phi = \arctan(y/x), \tag{4.5-8}$$

where $\Lambda = \lambda_o/\sin\theta$.

The resultant holograms take the form of vertical sinusoidal fringes of period Λ with dislocations near $x = 0$, as displayed in Fig. 4.5-11. Illuminating the recorded hologram with the reference wave will result in the generation of a spiral-phased wave with a helical wavefront and the associated value of l.

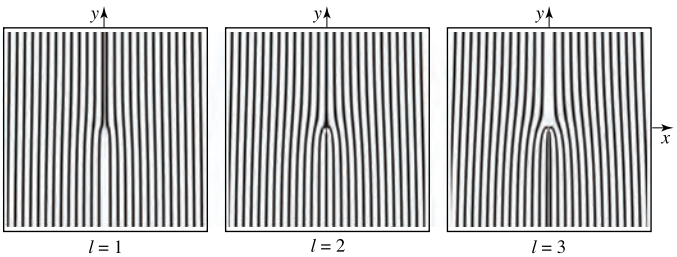


Figure 4.5-11 Computer-generated holographic optical elements for introducing a spiral phase into a planar wave, for three values of l.

READING LIST

Fourier Optics and Optical Signal Processing

J. W. Goodman, *Introduction to Fourier Optics*, Freeman, 4th ed. 2017.

K. Khare, *Fourier Optics and Computational Imaging*, Wiley, 2016.

R. K. Tyson, *Principles and Applications of Fourier Optics*, IOP Publishing, 2014.

D. Voelz, *Computational Fourier Optics: A MATLAB Tutorial*, SPIE Optical Engineering Press, 2011.

O. K. Ersoy, *Diffraction, Fourier Optics, and Imaging*, Wiley, 2006.

E. G. Steward, *Fourier Optics: An Introduction*, Halsted Press, 2nd ed. 1987; Dover, reissued 2004.

W. Lauterborn and T. Kurz, *Coherent Optics: Fundamentals and Applications*, Springer-Verlag, 2nd ed. 2003.

E. L. O'Neill, *Introduction to Statistical Optics*, Addison–Wesley, 1963; Dover, reissued 2003.

M. A. Abushagur and H. Caulfield, eds., *Selected Papers on Fourier Optics*, SPIE Optical Engineering Press (Milestone Series Volume 105), 1995.

P. W. Hooijmans, *Coherent Optical System Design*, Wiley, 1994.

A. VanderLugt, *Optical Signal Processing*, Wiley, 1992.

F. T. Yu and S. Yin, eds., *Selected Papers on Coherent Optical Processing*, SPIE Optical Engineering Press (Milestone Series Volume 52), 1992.

G. Reynolds, J. B. DeVelis, G. B. Parrent, and B. J. Thompson, *The New Physical Optics Notebook: Tutorials in Fourier Optics*, SPIE Optical Engineering Press, 1989.

G. Harburn, C. A. Taylor, and T. R. Welberry, *Atlas of Optical Transforms*, Cornell University Press, 1975.

M. Cagnet, M. Françon, and S. Mallick, *Atlas of Optical Phenomena*, Springer-Verlag, reprinted with supplement 1971.

M. Cagnet, M. Françon, and J. C. Thrierr, *Atlas of Optical Phenomena*, Springer-Verlag, 1962.

Diffraction

P. Ya. Ufimtsev, *Fundamentals of the Physical Theory of Diffraction*, Wiley, 2nd ed. 2014.

M. Nieto-Vesperinas, *Scattering and Diffraction in Physical Optics*, World Scientific, 2nd ed. 2006.

H. M. Nussenzveig, *Diffraction Effects in Semiclassical Scattering*, Cambridge University Press, 1992, paperback ed. 2006.

A. Sommerfeld, *Mathematical Theory of Diffraction*, Mathematische Annalen, 1896; Birkhäuser, 2004.

D. C. O'Shea, T. J. Suleski, A. D. Kathman, and D. W. Prather, *Diffractive Optics: Design, Fabrication, and Test*, SPIE Optical Engineering Press, 2003.

J. M. Cowley, *Diffraction Physics*, Elsevier, 3rd revised ed. 1995.

K. E. Oughstun, ed., *Selected Papers on Scalar Wave Diffraction*, SPIE Optical Engineering Press (Milestone Series Volume 51), 1992.

M. Françon, *Diffraction: Coherence in Optics*, Pergamon, 1966.

Imaging

G. de Villiers and E. R. Pike, *The Limits of Resolution*, CRC Press/Taylor & Francis, 2017.

D. F. Buscher, *Practical Optical Interferometry: Imaging at Visible and Infrared Wavelengths*, Cambridge University Press, 2015.

G. Saxby, *The Science of Imaging*, CRC Press/Taylor & Francis, 2nd ed. 2011.

B. E. A. Saleh, *Introduction to Subsurface Imaging*, Cambridge University Press, 2011.

V. N. Mahajan, *Optical Imaging and Aberrations, Part II. Wave Diffraction Optics*, SPIE Optical Engineering Press, 2nd ed. 2011.

R. L. Easton, Jr., *Fourier Methods in Imaging*, Wiley, 2010.

D. J. Brady, *Optical Imaging and Spectroscopy*, Wiley, 2009.

H. H. Barrett and K. J. Myers, *Foundations of Image Science*, Wiley, 2004.

S. Jutamulia, ed., *Selected Papers on Near-Field Optics*, SPIE Optical Engineering Press (Milestone Series Volume 172), 2002.

C. S. Williams and O. A. Becklund, *Introduction to the Optical Transfer Function*, SPIE Optical Engineering Press, 2002.

G. D. Boreman, *Modulation Transfer Function in Optical and Electro-Optical Systems*, SPIE Optical Engineering Press, 2001.

M. Françon, *Optical Image Formation and Processing*, Academic Press, 1979.

Holography

U. Schnars, C. Falldorf, J. Watson, and W. Jüptner, *Digital Holography and Wavefront Sensing: Principles, Techniques and Applications*, Springer-Verlag, 2nd ed. 2015.

T.-C. Poon and J.-P. Liu, *Introduction to Modern Digital Holography: With Matlab*, Cambridge University Press, 2014.

P.-A. Blanche, *Field Guide to Holography*, SPIE Optical Engineering Press, 2014.

V. Toal, *Introduction to Holography*, CRC Press/Taylor & Francis, 2012.

F. Unterseher, J. Hansen, and B. Schlesinger, *Holography Handbook: Making Holograms the Easy Way*, Ross, paperback 3rd ed. 2010.

L. Yaroslavsky, *Digital Holography and Digital Image Processing: Principles, Methods, Algorithms*, Kluwer, 2004, paperback ed. 2010.

G. Saxby, *Practical Holography*, Institute of Physics, 3rd ed. 2004.

P. Hariharan, *Basics of Holography*, Cambridge University Press, 2002.

H. I. Bjelkhagen and H. J. Caulfield, eds., *Selected Papers on Fundamental Techniques in Holography*, SPIE Optical Engineering Press (Milestone Series Volume 171), 2001.

J. E. Kasper and S. A. Feller, *Complete Book of Holograms: How They Work and How to Make Them*, Wiley, 1987; Dover, reissued 2001.

R. S. Sirohi and K. D. Hinsch, eds., *Selected Papers on Holographic Interferometry Principles and Techniques*, SPIE Optical Engineering Press (Milestone Series Volume 144), 1998.

M. Françon, *Holography*, Academic Press, 1974.

D. Gabor, Holography, 1948–1971 (Nobel Lecture in Physics, 1971), in S. Lundqvist, ed., *Nobel Lectures in Physics 1971–1980*, World Scientific, 1992.

PROBLEMS

4.1-3 **Correspondence Between Harmonic Functions and Plane Waves.** The complex amplitudes of a monochromatic wave of wavelength λ in the $z = 0$ and $z = d$ planes are $f(x, y)$ and $g(x, y)$, respectively. Assuming that $d = 10^4\lambda$, use harmonic analysis to determine $g(x, y)$ in the following cases:

(a) $f(x, y) = 1$;

(b) $f(x, y) = \exp[(-j\pi/\lambda)(x + y)]$;

(c) $f(x, y) = \cos(\pi x/2\lambda)$;

(d) $f(x, y) = \cos^2(\pi y/2\lambda)$;

(e) $f(x, y) = \sum_m \text{rect}[(x/10\lambda) - 2m]$, $m = 0, \pm 1, \pm 2, \ldots$, where $\text{rect}(x) = 1$ if $|x| \leq 1/2$ and 0, otherwise.

Describe the physical nature of the wave in each case.

4.1-4 **Conical Confinement Angle.** In Prob. 4.1-3, if $f(x, y)$ is a circularly symmetric function with a maximum spatial frequency of 200 lines/mm, determine the angle of the cone within which the wave directions are confined. Assume that $\lambda = 633$ nm.

4.1-5 **Logarithmic Interconnection Map.** A transparency of amplitude transmittance $\mathrm{t}(x, y) = \exp[-j2\pi\phi(x)]$ is illuminated with a uniform plane wave of wavelength $\lambda = 1$ μm. The transmitted light is focused by an adjacent lens of focal length $f = 100$ cm. What must $\phi(x)$ be so that the ray that hits the transparency at position x is deflected and focused to a position $x' = \ln(x)$ for all $x > 0$? (Note that x and x' are measured in millimeters.) If the lens is removed, how should $\phi(x)$ be modified so that the system performs the same function? This system may be used to perform a logarithmic coordinate transformation, as discussed in Chapter 24 [see Exercise 24.1-2].

4.2-3 **Proof of the Lens Fourier-Transform Property.**

(a) Show that the convolution of $f(x)$ and $\exp(-j\pi x^2/\lambda d)$ may be obtained via three steps: multiply $f(x)$ by $\exp(-j\pi x^2/\lambda d)$; evaluate the Fourier transform of the product at the frequency $\nu_x = x/\lambda d$; and multiply the result by $\exp(-j\pi x^2/\lambda d)$.

(b) The Fourier transform system in Fig. 4.2-4 is a cascade of three systems — propagation a distance f in free space, transmission through a lens of focal length f, and propagation a distance f in free space. Noting that propagation a distance d in free space is equivalent

to convolution with $\exp(-j\pi x^2/\lambda d)$ [see (4.1-20)], and using the result in (a), derive the lens' Fourier-transform equation (4.2-8). For simplicity ignore the y dependence.

4.2-4 **Fourier Transform of the Line Functions.** A transparency of amplitude transmittance $\mathrm{t}(x, y)$ is illuminated with a plane wave of wavelength $\lambda = 1$ μm and focused with a lens of focal length $f = 100$ cm. Sketch the intensity distribution in the plane of the transparency and in the lens focal plane in the following cases (all distances are measured in mm):

(a) $\mathrm{t}(x, y) = \delta(x - y)$;

(b) $\mathrm{t}(x, y) = \delta(x + a) + \delta(x - a), \quad a = 1$ mm;

(c) $\mathrm{t}(x, y) = \delta(x + a) + j\delta(x - a), \quad a = 1$ mm;

where $\delta(\cdot)$ is the delta function (see Appendix A, Sec. A.1).

4.2-5 **Design of an Optical Fourier-Transform System.** Consider a lens used to display the Fourier transform of a two-dimensional function with spatial frequencies between 20 and 200 lines/mm. If the wavelength of light is $\lambda = 488$ nm, what should be the focal length of the lens so that the highest and lowest spatial frequencies are separated by a distance of 9 cm in the Fourier plane?

*4.2-6 **Generation of the Airy Beam by Use of an Optical Fourier-Transform System.** As described in Sec. 3.5B, the Airy beam has an amplitude $A(x, 0) = \mathrm{Ai}(x/W_0)$ in the $z = 0$ plane, where $\mathrm{Ai}(x)$ is the Airy function and W_0 is a measure of the beam width. Given that the Fourier transform of the phase function $\exp(jx^3/3)$ is equal to $2\pi\mathrm{Ai}(2\pi\nu_x)$, design an optical Fourier-transform system that generates the Airy beam using a lens of focal length f and a mask whose amplitude transmittance is $\exp(jx^3/3)$. Determine an expression for W_0 of the beam generated in terms of f and the wavelength λ.

4.3-4 **Fraunhofer Diffraction from a Diffraction Grating.** Derive an expression for the Fraunhofer diffraction pattern for an aperture made of $M = 2L + 1$ parallel slits of infinitesimal widths separated by equal distances $a = 10\lambda$,

$$p(x, y) = \sum_{m=-L}^{L} \delta(x - ma).$$

Sketch the pattern as a function of the observation angle $\theta = x/d$, where d is the observation distance.

4.3-5 **Fraunhofer Diffraction with an Oblique Incident Wave.** The diffraction pattern from an aperture with aperture function $p(x, y)$ is proportional to $|P(x/\lambda d, y/\lambda d)|^2$, where $P(\nu_x, \nu_y)$ is the Fourier transform of $p(x, y)$ and d is the distance between the aperture and observation planes. What is the diffraction pattern when the direction of the incident wave makes a small angle $\theta_x \ll 1$, with the z-axis in the x–z plane?

*4.3-6 **Fresnel Diffraction from Two Pinholes.** Show that the Fresnel diffraction pattern from two pinholes separated by a distance $2a$, i.e., $p(x, y) = [\delta(x-a)+\delta(x+a)]\delta(y)$, at an observation distance d is the periodic pattern, $I(x, y) = (2/\lambda d)^2 \cos^2(2\pi ax/\lambda d)$.

*4.3-7 **Relation Between Fresnel and Fraunhofer Diffraction.** Show that the Fresnel diffraction pattern of the aperture function $p(x, y)$ is equal to the Fraunhofer diffraction pattern of the aperture function $p(x, y) \exp[-j\pi(x^2 + y^2)/\lambda d]$.

4.4-1 **Blurring a Sinusoidal Grating.** An object $f(x, y) = \cos^2(2\pi x/a)$ is imaged by a defocused single-lens imaging system whose impulse response function $h(x, y) = 1$ within a square of width D, and is 0 elsewhere. Derive an expression for the distribution of the image $g(x, 0)$ in the x direction. Derive an expression for the contrast of the image in terms of the ratio D/a. The contrast is defined as $(\max - \min)/(\max + \min)$, where max and min are the maximum and minimum values of $g(x, 0)$, respectively.

4.4-2 **Image of a Phase Object.** An imaging system has an impulse response function $h(x, y) = \mathrm{rect}(x)\,\delta(y)$. If the input wave is

$$f(x, y) = \begin{cases} \exp\left(j\dfrac{\pi}{2}\right) & \text{for } x > 0 \\ \exp\left(-j\dfrac{\pi}{2}\right) & \text{for } x \le 0, \end{cases}$$

determine and sketch the intensity $|g(x, y)|^2$ of the output wave $g(x, y)$. Verify that even though the intensity of the input wave $|f(x, y)|^2 = 1$, the intensity of the output wave is not uniform.

4.4-3 **Optical Spatial Filtering.** Consider the spatial filtering system shown in Fig. 4.4-5 with $f = 1000$ mm. The system is illuminated with a uniform plane wave of unit amplitude and wavelength $\lambda = 10^{-3}$ mm. The input transparency has amplitude transmittance $f(x, y)$ and the mask has amplitude transmittance $p(x, y)$. Write an expression relating the complex amplitude $g(x, y)$ of light in the image plane to $f(x, y)$ and $p(x, y)$. Assuming that all distances are measured in mm, sketch $g(x, 0)$ in the following cases:

(a) $f(x, y) = \delta(x - 5)$ and $p(x, y) = \text{rect}(x)$;

(b) $f(x, y) = \text{rect}(x)$ and $p(x, y) = \text{sinc}(x)$.

Determine $p(x, y)$ such that $g(x, y) = \nabla_T^2 f(x, y)$, where $\nabla_T^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the transverse Laplacian operator.

4.4-4 **Optical Correlation.** Show how a spatial filter may be used to perform the operation of optical correlation (see Appendix A) between two images described by the real-valued functions $f_1(x, y)$ and $f_2(x, y)$. Under what conditions would the complex amplitude transmittances of the masks and transparencies used be real-valued?

*4.4-5 **Impulse Response Function of a Severely Defocused System.** Using wave optics, show that the impulse response function of a severely defocused imaging system (where the defocusing error ϵ is very large) may be approximated by $h(x, y) = p(x/\epsilon d_2, y/\epsilon d_2)$, where $p(x, y)$ is the pupil function. *Hint:* Use the method of stationary phase described on page 125 (second proof) to evaluate the integral resulting from the use of (4.4-11) and (4.4-10). Note that this is the same result as that predicted by the ray theory of light [see (4.4-2)].

4.4-6 **Two-Point Resolution.**

(a) Consider the single-lens imaging system discussed in Sec. 4.4C. Assuming a square aperture of width D, unit magnification, and perfect focus, write an expression for the impulse response function $h(x, y)$.

(b) Determine the response of the system to an object consisting of two points separated by a distance b, i.e.,

$$f(x, y) = \delta(x)\,\delta(y) + \delta(x - b)\,\delta(y).$$

(c) If $\lambda d_2/D = 0.1$ mm, sketch the magnitude of the image $g(x, 0)$ as a function of x when the points are separated by a distance $b = 0.5$, 1, and 2 mm. What is the minimum separation between the two points such that the image remains discernible as two spots instead of a single spot, i.e., has two peaks?

4.4-7 **Ring Aperture.**

(a) A focused single-lens imaging system, with magnification $M = 1$ and focal length $f = 100$ cm has an aperture in the form of a ring

$$p(x, y) = \begin{cases} 1, & a \le \sqrt{x^2 + y^2} \le b, \\ 0, & \text{otherwise}, \end{cases}$$

where $a = 5$ mm and $b = 6$ mm. Determine the transfer function $H(\nu_x, \nu_y)$ of the system and sketch its cross section $H(\nu_x, 0)$. Assume that the wavelength $\lambda = 1$ μm.

(b) If the image plane is now moved closer to the lens so that its distance from the lens becomes $d_2 = 25$ cm, with the distance between the object plane and the lens d_1 as in (a), use the ray-optics approximation to determine the impulse response function of the imaging system $h(x, y)$ and sketch $h(x, 0)$.

4.5-1 **Holography with a Spherical Reference Wave.** The choice of a uniform plane wave as a reference wave is not essential to holography; other waves can be used. Assuming that the reference wave is a spherical wave centered about the point $(0, 0, -d)$, determine the hologram pattern and examine the reconstructed wave when:

(a) the object wave is a plane wave traveling at an angle θ_x;

(b) the object wave is a spherical wave centered at $(-x_0, 0, -d_1)$.

Approximate spherical waves by paraboloidal waves.

4.5-2 **Optical Correlation via Holography.** A transparency with an amplitude transmittance given by $f(x, y) = f_1(x - a, y) + f_2(x + a, y)$ is Fourier transformed by a lens and the intensity is recorded on a transparency (hologram). The hologram is subsequently illuminated with a reference wave and the reconstructed wave is Fourier transformed with a lens to generate the function $g(x, y)$. Derive an expression relating $g(x, y)$ to $f_1(x, y)$ and $f_2(x, y)$. Show how the correlation of $f_1(x, y)$ and $f_2(x, y)$ may be determined with this system.

CHAPTER

5

ELECTROMAGNETIC OPTICS

5.1 ELECTROMAGNETIC THEORY OF LIGHT 162

5.2 ELECTROMAGNETIC WAVES IN DIELECTRIC MEDIA 166

A. Linear, Nondispersive, Homogeneous, and Isotropic Media

B. Nonlinear, Dispersive, Inhomogeneous, or Anisotropic Media

5.3 MONOCHROMATIC ELECTROMAGNETIC WAVES 172

5.4 ELEMENTARY ELECTROMAGNETIC WAVES 175

A. Plane, Dipole, and Gaussian Electromagnetic Waves

B. Relation Between Electromagnetic Optics and Scalar Wave Optics

C. Vector Beams

5.5 ABSORPTION AND DISPERSION 181

A. Absorption

B. Dispersion

C. The Resonant Medium

5.6 SCATTERING OF ELECTROMAGNETIC WAVES 192

A. Born Approximation

B. Rayleigh Scattering

C. Mie Scattering

D. Attenuation in a Medium with Scatterers

5.7 PULSE PROPAGATION IN DISPERSIVE MEDIA 199

James Clerk Maxwell (1831–1879) advanced the theory that light is an electromagnetic wave phenomenon. He formulated a set of fundamental equations of enormous importance that bear his name.

Lord Rayleigh (John William Strutt) (1842–1919) contributed extensively to many areas of optics, including blackbody radiation, image formation, and scattering. He was awarded the Nobel Prize in 1904.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

It is apparent from the results presented in Chapters 2–4 that wave optics has a far greater reach than ray optics. Remarkably, both approaches provide similar results for many simple optical phenomena involving paraxial waves, such as the focusing of light by a lens and the behavior of light in graded-index media and periodic systems. But it is also clear that wave optics offers something that ray optics cannot: the ability to explain phenomena such as interference and diffraction, which involve phase, and therefore lie hopelessly beyond the reach of a simple construct like ray optics. In spite of its many successes, however, wave optics, like ray optics, is unable to quantitatively account for some simple observations in an optics experiment, such as the division of light at a beamsplitter. The fraction of light reflected (and transmitted) turns out to depend on the polarization of the incident light, which means that the light must be treated in the context of a vector, rather than a scalar, theory. That's where electromagnetic optics enters the picture. In common with radio waves and X-rays, as shown in Fig. 5.0-1, light is an electromagnetic phenomenon that is described by a *vector* wave theory. Electromagnetic radiation propagates in the form of two mutually coupled vector waves, an electric-field wave and a magnetic-field wave. From this perspective, the wave-optics approach set forth in Chapter 2, and developed in Chapters 3 and 4, is merely a *scalar* approximation to the more complete electromagnetic theory.

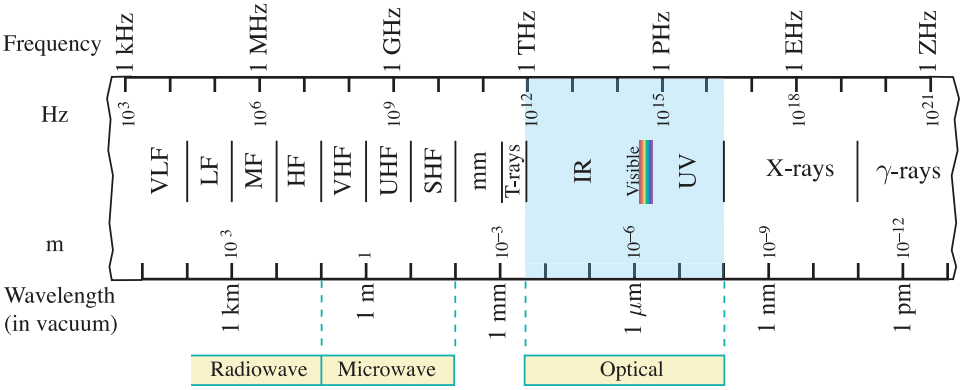


Figure 5.0-1 The electromagnetic spectrum from low frequencies (long wavelengths) to high frequencies (short wavelengths). The optical region, shown as shaded, is displayed in greater detail in Fig. 2.0-1.

Electromagnetic optics thus encompasses wave optics, which in turn reduces to ray optics in the limit of short wavelengths, as shown in Chapter 2. This hierarchy is displayed in Fig. 5.0-2.

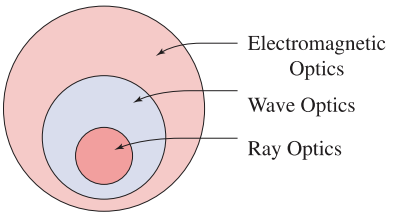


Figure 5.0-2 Electromagnetic optics is a *vector* theory comprising an electric field and a magnetic field that vary in time and space. Wave optics is an approximation to electromagnetic optics that relies on the wavefunction, a *scalar* function of time and space. Ray optics is the limit of wave optics when the wavelength is very short.

Optical frequencies occupy a band of the electromagnetic spectrum that extends from the infrared through the visible to the ultraviolet, as shown in Fig. 5.0-1. The

range of wavelengths that is generally considered to lie in the optical domain extends from 10 nm to 300 μm (as is shown in greater detail in Fig. 2.0-1). Because these wavelengths are substantially shorter than those of radiowaves, or even microwaves, the techniques involved in their generation, transmission, and detection have traditionally been rather distinct. In recent years, however, the march toward miniaturization has served to blur these differences: it is now commonplace to encounter wavelength- and subwavelength-size resonators, antennas, waveguides, lasers, and other structures.

This Chapter

This chapter offers a brief review of those aspects of electromagnetic theory that are of paramount importance in optics. The fundamental theoretical construct — Maxwell's equations — is set forth in Sec. 5.1. The behavior of optical electromagnetic waves in dielectric media is examined in Sec. 5.2. Together, these sections lay out the fundamentals of electromagnetic optics and provide the set of laws that govern the remaining sections of the chapter. These rules simplify considerably for the special case of monochromatic light, as discussed in Sec. 5.3. Elementary electromagnetic waves (plane waves, dipole waves, and Gaussian beams), introduced in Sec. 5.4, provide important examples that are often encountered in practice. Section 5.5 is devoted to a study of the propagation of light in dispersive media, which exhibit wavelength-dependent absorption and refraction, as do real media. The scattering of electromagnetic waves, considered in Sec. 5.6, plays an important role in optics and plasmonics, as discussed in Chapter 8. Finally, in Sec. 5.7, we consider pulse propagation in dispersive media, which provides a basic underpinning for Chapters 10, 23, and 25.

Chapter 6, which is based on the theory of electromagnetic optics presented in this chapter, deals explicitly with the polarization of light and the interaction of polarized light with dielectric and anisotropic media such as liquid crystals. The material set forth here also forms the basis for the expositions provided in Chapters 7–11, which deal, respectively, with the optics of layered and periodic media, metals and metamaterials, guided waves, fibers, and resonators. Chapters 12 and 22, devoted to statistical optics and nonlinear optics, respectively, are also based on electromagnetic optics.

5.1 ELECTROMAGNETIC THEORY OF LIGHT

An electromagnetic field is described by two related *vector* fields that are functions of position and time: the **electric field** $\mathcal{E}(\mathbf{r}, t)$ and the **magnetic field** $\mathcal{H}(\mathbf{r}, t)$. In general, therefore, six scalar functions of position and time are required to describe light in free space. Fortunately, these six functions are interrelated since they must satisfy the celebrated set of coupled partial differential equations known as Maxwell's equations.

Maxwell's Equations in Free Space

The electric- and magnetic-field vectors in free space satisfy **Maxwell's equations**:

$$\nabla \times \mathcal{H} = \epsilon_o \frac{\partial \mathcal{E}}{\partial t} \tag{5.1-1}$$

$$\nabla \times \mathcal{E} = -\mu_o \frac{\partial \mathcal{H}}{\partial t} \tag{5.1-2}$$

$$\nabla \cdot \mathcal{E} = 0 \tag{5.1-3}$$

$$\nabla \cdot \mathcal{H} = 0, \tag{5.1-4}$$

Maxwell's Equations (Free Space)

where the constants $\epsilon_o \approx (1/36\pi) \times 10^{-9}$ F/m and $\mu_o = 4\pi \times 10^{-7}$ H/m (MKS units) are, respectively, the **electric permittivity** and the **magnetic permeability** of free space. The vector operators $\nabla\cdot$ and $\nabla\times$ represent the divergence and curl, respectively.[†]

The Wave Equation

A necessary condition for $\mathcal{E}$ and $\mathcal{H}$ to satisfy Maxwell's equations is that each of their components satisfy the wave equation

$$\nabla^2 u - \frac{1}{c_o^2}\frac{\partial^2 u}{\partial t^2} = 0. \tag{5.1-5}$$

Wave Equation (Free Space)

Here

$$c_o = \frac{1}{\sqrt{\epsilon_o \mu_o}} \approx 3 \times 10^8 \text{ m/s} \tag{5.1-6}$$

Speed of Light (Free Space)

is the speed of light in vacuum, and the scalar function $u(\mathbf{r}, t)$ represents any of the three components $(\mathcal{E}_x, \mathcal{E}_y, \mathcal{E}_z)$ of $\mathcal{E}$ or the three components $(\mathcal{H}_x, \mathcal{H}_y, \mathcal{H}_z)$ of $\mathcal{H}$.

The wave equation may be derived from Maxwell's equations by applying the curl operation $\nabla\times$ to (5.1-2), making use of the vector identity $\nabla \times (\nabla \times \mathcal{E}) = \nabla(\nabla \cdot \mathcal{E}) - \nabla^2\mathcal{E}$, and then using (5.1-1) and (5.1-3) to show that each component of $\mathcal{E}$ satisfies the wave equation. A similar procedure is followed for $\mathcal{H}$. Since Maxwell's equations and the wave equation are linear, the principle of superposition applies: if two sets of electric and magnetic fields are solutions to these equations separately, their sum is also a solution.

The connection between electromagnetic optics and wave optics is now evident. The wave equation (2.1-2), which is the basis of wave optics, is embedded in the structure of electromagnetic theory; the speed of light is related to the electromagnetic constants ϵ_o and μ_o by (5.1-6); and the scalar wavefunction $u(\mathbf{r}, t)$ in Chapter 2 represents any of the six components of the electric- and magnetic-field vectors. Electromagnetic optics reduces to wave optics in problems for which the vector nature of the electromagnetic fields is not of essence. As we shall see in this and the following chapters, the vector character of light underlies polarization phenomena and governs the amount of light reflected or transmitted through boundaries between different media, and therefore determines the characteristics of light propagation in waveguides, layered media, and optical resonators.

Maxwell's Equations in a Medium

In a medium devoid of free electric charges and currents, two additional vector fields are required — the **electric flux density** (also called the **electric displacement**) $\mathcal{D}(\mathbf{r}, t)$ and the **magnetic flux density** $\mathcal{B}(\mathbf{r}, t)$. The four fields, $\mathcal{E}$, $\mathcal{H}$, $\mathcal{D}$, and $\mathcal{B}$, are related by Maxwell's equations in a source-free medium:

[†] In a Cartesian coordinate system $\nabla \cdot \mathcal{E} = \partial\mathcal{E}_x/\partial x + \partial\mathcal{E}_y/\partial y + \partial\mathcal{E}_z/\partial z$ whereas $\nabla \times \mathcal{E}$ is a vector with Cartesian components $(\partial\mathcal{E}_z/\partial y - \partial\mathcal{E}_y/\partial z)$, $(\partial\mathcal{E}_x/\partial z - \partial\mathcal{E}_z/\partial x)$, and $(\partial\mathcal{E}_y/\partial x - \partial\mathcal{E}_x/\partial y)$.

$$\nabla \times \mathcal{H} = \frac{\partial \mathcal{D}}{\partial t} \tag{5.1-7}$$

$$\nabla \times \mathcal{E} = -\frac{\partial \mathcal{B}}{\partial t} \tag{5.1-8}$$

$$\nabla \cdot \mathcal{D} = 0 \tag{5.1-9}$$

$$\nabla \cdot \mathcal{B} = 0. \tag{5.1-10}$$

Maxwell's Equations (Source-Free Medium)

Conductive media such as metals have free electric charges, requiring the addition of an associated current density $\mathcal{J}$ to the right-hand side of (5.1-7), as discussed in Sec. 8.2A. Maxwell's original formulation in 1865 comprised 20 simultaneous equations with 20 variables; these were condensed into their present form by Oliver Heaviside in 1885.

The relationship between the electric flux density $\mathcal{D}$ and the electric field $\mathcal{E}$ depends on the electric properties of the medium, which are characterized by the **polarization density** $\mathcal{P}$. In a dielectric medium, the polarization density is the macroscopic sum of the electric dipole moments induced by the electric field. Similarly, the relation between the magnetic flux density $\mathcal{B}$ and the magnetic field $\mathcal{H}$ depends on the magnetic properties of the medium, embodied in the **magnetization density** $\mathcal{M}$, which is defined analogously to the polarization density. The equations relating the flux densities and the fields are

$$\mathcal{D} = \epsilon_o \mathcal{E} + \mathcal{P} \tag{5.1-11}$$

$$\mathcal{B} = \mu_o \mathcal{H} + \mu_o \mathcal{M}. \tag{5.1-12}$$

The vector fields $\mathcal{P}$ and $\mathcal{M}$ are in turn related to the externally applied electric and magnetic fields $\mathcal{E}$ and $\mathcal{H}$ by relationships that depend on the electric and magnetic character of the medium, respectively, as will be described in Sec. 5.2. Equations relating $\mathcal{P}$ and $\mathcal{E}$, as well as $\mathcal{M}$ and $\mathcal{H}$, are established once the medium is specified. When these latter equations are substituted into Maxwell's equations in a source-free medium, the flux densities disappear.

In free space, $\mathcal{P} = \mathcal{M} = 0$, so that $\mathcal{D} = \epsilon_o \mathcal{E}$ and $\mathcal{B} = \mu_o \mathcal{H}$ whereupon (5.1-7)–(5.1-10) reduce to the free-space Maxwell's equations, (5.1-1)–(5.1-4).

Boundary Conditions

In a homogeneous medium, all components of the fields $\mathcal{E}$, $\mathcal{H}$, $\mathcal{D}$, and $\mathcal{B}$ are continuous functions of position. At the boundary between two dielectric media, in the absence of free electric charges and currents, the tangential components of the electric and magnetic fields $\mathcal{E}$ and $\mathcal{H}$, and the normal components of the electric and magnetic flux densities $\mathcal{D}$ and $\mathcal{B}$, must be continuous (Fig. 5.1-1).

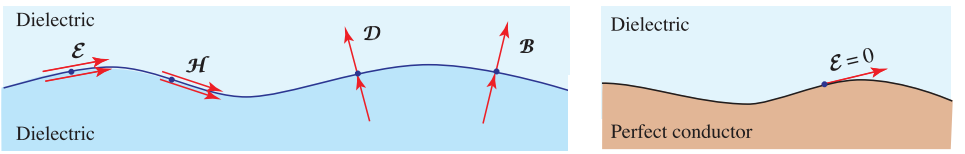


Figure 5.1-1 Boundary conditions at: (a) the interface between two dielectric media; (b) the interface between a perfect conductor and a dielectric material.

At the boundary between a dielectric medium and a perfectly conductive medium, the tangential components of the electric-field vector must vanish. Since a perfect

mirror is made of a perfectly conductive material (a metal), the component of the electric field parallel to the surface of the mirror must be zero. This requires that at normal incidence the electric fields of the reflected and incident waves must have equal magnitudes and a phase shift of π so that their sum adds up to zero.

These boundary conditions are an integral part of Maxwell's equations. They are used to determine the reflectance and transmittance of waves at various boundaries (see Sec. 6.2), and the propagation of waves in periodic dielectric structures (see Sec. 7.1) and waveguides (see Sec. 9.2).

Intensity, Power, and Energy

The flow of electromagnetic power is governed by the vector

$$\mathcal{S} = \mathcal{E} \times \mathcal{H}, \tag{5.1-13}$$

which is known as the **Poynting vector**. The direction of power flow is along the direction of the Poynting vector, i.e., orthogonal to both $\mathcal{E}$ and $\mathcal{H}$. The **optical intensity** $I(\mathbf{r}, t)$ (power flow across a unit area normal to the vector $\mathcal{S}$)[†] is the magnitude of the time-averaged Poynting vector $\langle \mathcal{S} \rangle$. The average is taken over times that are long in comparison with an optical cycle, but short compared to other times of interest. The wave-optics equivalent is given in (2.1-3).

Using the vector identity $\nabla \cdot (\mathcal{E} \times \mathcal{H}) = (\nabla \times \mathcal{E}) \cdot \mathcal{H} - (\nabla \times \mathcal{H}) \cdot \mathcal{E}$, together with Maxwell's equations (5.1-7)–(5.1-8) and (5.1-11)–(5.1-12), we obtain

$$\nabla \cdot \mathcal{S} = -\frac{\partial}{\partial t}\left(\frac{1}{2}\epsilon_o \mathcal{E}^2 + \frac{1}{2}\mu_o \mathcal{H}^2\right) - \mathcal{E} \cdot \frac{\partial \mathcal{P}}{\partial t} - \mu_o \mathcal{H} \cdot \frac{\partial \mathcal{M}}{\partial t}. \tag{5.1-14}$$

The first and second terms in parentheses in (5.1-14) represent the energy densities (per unit volume) stored in the electric and magnetic fields, respectively. The third and fourth terms represent the power densities associated with the material's electric and magnetic dipoles. Equation (5.1-14), known as the **Poynting theorem**, therefore represents conservation of energy: the power flow escaping from the surface of an incremental volume equals the time rate of change of the energy stored inside the volume.

Momentum

An electromagnetic wave carries linear momentum, which results in radiation pressure on objects from which the wave reflects or scatters. In free space, the linear momentum density (per unit volume) is a vector

$$\epsilon_o \mathcal{E} \times \mathcal{B} = \frac{1}{c^2}\mathcal{S} \tag{5.1-15}$$

Linear Momentum Density

proportional to the Poynting vector $\mathcal{S}$. The average momentum in a cylinder of length c and unit area is $(\langle \mathcal{S} \rangle / c^2) \cdot c = \langle \mathcal{S} \rangle / c$. This momentum crosses the unit area in a unit time, so that the average rate (per unit time) of momentum flow across a unit area oriented perpendicular to the direction of $\mathcal{S}$ is $\langle \mathcal{S} \rangle / c$.

An electromagnetic wave may also carry angular momentum and may therefore exert torque on an object. The average rate of angular momentum transported by an electromagnetic field is $\mathbf{r} \times \langle \mathcal{S} \rangle / c$. For example, the Laguerre–Gaussian beams introduced in Sec. 3.4 have helical wavefronts; the Poynting vector has an azimuthal component that leads to an orbital angular momentum.

[†] For a discussion of this interpretation, see M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th expanded and corrected ed. 2002, pp. 7–10.

5.2 ELECTROMAGNETIC WAVES IN DIELECTRIC MEDIA

The character of the medium is embodied in the relation between the polarization and magnetization densities, $\mathcal{P}$ and $\mathcal{M}$, on the one hand, and the electric and magnetic fields, $\mathcal{E}$ and $\mathcal{H}$, on the other; these are known as the **constitutive relation**. In most media, the constitutive relation separates into a pair of constitutive relations, one between $\mathcal{P}$ and $\mathcal{E}$, and another between $\mathcal{M}$ and $\mathcal{H}$. The former describes the dielectric properties of the medium, whereas the latter describes its magnetic properties. With the notable exceptions of magnetic materials, optically active materials, and metamaterials, the principal emphasis in this book is on the dielectric properties. We therefore direct our attention to the $\mathcal{P}$-$\mathcal{E}$ relations for various dielectric media; the $\mathcal{M}$-$\mathcal{H}$ relations for magnetic media obey similar relations under similar conditions.

It is useful to regard the $\mathcal{P}$-$\mathcal{E}$ constitutive relation as arising from a system in which $\mathcal{E}$ is the input and $\mathcal{P}$ is the output or response (Fig. 5.2-1). Note that $\mathcal{E} = \mathcal{E}(\mathbf{r}, t)$ and $\mathcal{P} = \mathcal{P}(\mathbf{r}, t)$ are functions of both position and time.

Figure 5.2-1 In response to an applied electric field $\mathcal{E}$, the dielectric medium creates a polarization density $\mathcal{P}$.

Definitions

- A dielectric medium is said to be *linear* if the vector field $\mathcal{P}(\mathbf{r}, t)$ is linearly related to the vector field $\mathcal{E}(\mathbf{r}, t)$. The principle of superposition then applies.
- The medium is said to be *nondispersive* if its response is instantaneous, i.e., if $\mathcal{P}$ at time t is determined by $\mathcal{E}$ at the same time t and not by prior values of $\mathcal{E}$. Nondispersiveness is clearly an idealization since all physical systems, no matter how rapidly they may respond, do have a response time that is finite.
- The medium is said to be *homogeneous* if the relation between $\mathcal{P}$ and $\mathcal{E}$ is independent of the position $\mathbf{r}$.
- The medium is said to be *isotropic* if the relation between the vectors $\mathcal{P}$ and $\mathcal{E}$ is independent of the direction of the vector $\mathcal{E}$, so that the medium exhibits the same behavior from all directions. The vectors $\mathcal{P}$ and $\mathcal{E}$ must then be parallel.
- The medium is said to be *spatially nondispersive* if the relation between $\mathcal{P}$ and $\mathcal{E}$ is local, i.e., if $\mathcal{P}$ at each position $\mathbf{r}$ is influenced only by $\mathcal{E}$ at the same position $\mathbf{r}$. The medium is assumed to be spatially nondispersive throughout this chapter (optically active media, considered in Sec. 6.4A, are spatially dispersive).

A. Linear, Nondispersive, Homogeneous, and Isotropic Media

Let us first consider the simplest case of *linear*, *nondispersive*, *homogeneous*, and *isotropic* dielectric media. The vectors $\mathcal{P}$ and $\mathcal{E}$ at every position and time are then parallel and proportional, so that

$$\mathcal{P} = \epsilon_o \chi \mathcal{E}, \tag{5.2-1}$$

where the scalar constant χ is called the **electric susceptibility** (Fig. 5.2-2).

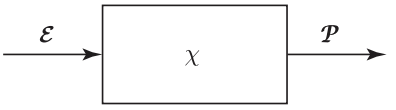


Figure 5.2-2 A linear, nondispersive, homogeneous, and isotropic medium is fully characterized by a single constant, the electric susceptibility χ.

Substituting (5.2-1) in (5.1-11) shows that $\mathcal{D}$ and $\mathcal{E}$ are also parallel and proportional,

$$\mathcal{D} = \epsilon \mathcal{E}, \tag{5.2-2}$$

where the scalar quantity

$$\epsilon = \epsilon_o(1 + \chi) \tag{5.2-3}$$

is defined as the **electric permittivity** of the medium. The **relative permittivity** $\epsilon/\epsilon_o = 1 + \chi$ is also called the **dielectric constant** of the medium.

Under similar conditions, the magnetic relation can be written in the form

$$\mathcal{B} = \mu \mathcal{H}, \tag{5.2-4}$$

where μ is the **magnetic permeability** of the medium.

With the relations (5.2-2) and (5.2-4), Maxwell's equations in (5.1-7)–(5.1-10) relate only the two vector fields $\mathcal{E}(\mathbf{r}, t)$ and $\mathcal{H}(\mathbf{r}, t)$, simplifying to

$$\nabla \times \mathcal{H} = \epsilon \frac{\partial \mathcal{E}}{\partial t} \tag{5.2-5}$$

$$\nabla \times \mathcal{E} = -\mu \frac{\partial \mathcal{H}}{\partial t} \tag{5.2-6}$$

$$\nabla \cdot \mathcal{E} = 0 \tag{5.2-7}$$

$$\nabla \cdot \mathcal{H} = 0. \tag{5.2-8}$$

Maxwell's Equations
(Linear, Nondispersive, Homogeneous, Isotropic, Source-Free Medium)

It is apparent that (5.2-5)–(5.2-8) are identical in form to the free-space Maxwell's equations in (5.1-1)–(5.1-4) except that ϵ replaces ϵ_o and μ replaces μ_o. Each component of $\mathcal{E}$ and $\mathcal{H}$ therefore satisfies the wave equation

$$\nabla^2 u - \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = 0, \tag{5.2-9}$$

Wave Equation
(in a Medium)

where the speed of light in the medium is denoted c:

$$c = \frac{1}{\sqrt{\epsilon\,\mu}}. \tag{5.2-10}$$

Speed of Light
(in a Medium)

The ratio of the speed of light in free space to that in the medium, c_o/c, is defined as the *refractive index* n:

$$n = \frac{c_o}{c} = \sqrt{\frac{\epsilon}{\epsilon_o}\frac{\mu}{\mu_o}}, \tag{5.2-11}$$

Refractive Index

where (5.1-6) provides

$$c_o = \frac{1}{\sqrt{\epsilon_o \mu_o}}. \tag{5.2-12}$$

For a nonmagnetic material, $\mu = \mu_o$ and

$$n = \sqrt{\frac{\epsilon}{\epsilon_o}} = \sqrt{1+\chi}, \tag{5.2-13}$$

Refractive Index (Nonmagnetic Media)

so that the refractive index is the square root of the relative permittivity. These relations provide another point of connection with scalar wave optics (Sec. 2.1), as discussed further in Sec. 5.4B.

Finally, the Poynting theorem (5.1-14) based on Maxwell's equations (5.2-5) and (5.2-6) takes the form of a continuity equation

$$\nabla \cdot \mathcal{S} = -\frac{\partial \mathcal{W}}{\partial t} \tag{5.2-14}$$

where

$$\mathcal{W} = \tfrac{1}{2}\epsilon\, \mathcal{E}^2 + \tfrac{1}{2}\mu \mathcal{H}^2 \tag{5.2-15}$$

is the energy density stored in the medium.

B. Nonlinear, Dispersive, Inhomogeneous, or Anisotropic Media

We now consider nonmagnetic dielectric media for which one or more of the properties of linearity, nondispersiveness, homogeneity, and isotropy are not satisfied.

Inhomogeneous Media

We first consider an inhomogeneous dielectric (such as a graded-index medium) that is linear, nondispersive, and isotropic. The simple proportionalities, $\mathcal{P} = \epsilon_o \chi \mathcal{E}$ and $\mathcal{D} = \epsilon \mathcal{E}$, remain intact, but the coefficients χ and ϵ become functions of position: $\chi = \chi(\mathbf{r})$ and $\epsilon = \epsilon(\mathbf{r})$ (Fig. 5.2-3). The refractive index therefore also becomes position dependent so that $n = n(\mathbf{r})$.

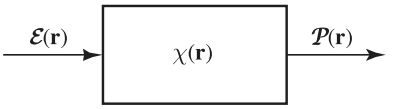


Figure 5.2-3 An inhomogeneous (but linear, nondispersive, and isotropic) medium is characterized by a position dependent susceptibility $\chi(\mathbf{r})$.

Beginning with Maxwell's equations, (5.1-7)–(5.1-10), and noting that $\epsilon = \epsilon(\mathbf{r})$ is position dependent, we apply the curl operation $\nabla\times$ to both sides of (5.1-8). We then use (5.1-7) to write

$$\boxed{\frac{\epsilon_o}{\epsilon}\,\nabla\times(\nabla\times\boldsymbol{\mathcal{E}}) = -\frac{1}{c_o^2}\frac{\partial^2\boldsymbol{\mathcal{E}}}{\partial t^2}.} \qquad \begin{matrix}(5.2\text{-}16)\\ \text{Wave Equation}\\ \text{(Inhomogeneous Medium)}\end{matrix}$$

The magnetic field satisfies a different equation:

$$\boxed{\nabla\times\left(\frac{\epsilon_o}{\epsilon}\,\nabla\times\boldsymbol{\mathcal{H}}\right) = -\frac{1}{c_o^2}\frac{\partial^2\boldsymbol{\mathcal{H}}}{\partial t^2}.} \qquad \begin{matrix}(5.2\text{-}17)\\ \text{Wave Equation}\\ \text{(Inhomogeneous Medium)}\end{matrix}$$

Equation (5.2-16) may also be written in the form

$$\nabla^2\boldsymbol{\mathcal{E}} + \nabla\left(\frac{1}{\epsilon}\,\nabla\epsilon\cdot\boldsymbol{\mathcal{E}}\right) - \mu_o\epsilon\,\frac{\partial^2\boldsymbol{\mathcal{E}}}{\partial t^2} = 0. \tag{5.2-18}$$

The validity of (5.2-18) can be demonstrated by employing the following procedure. Use the identity $\nabla\times(\nabla\times\boldsymbol{\mathcal{E}}) = \nabla(\nabla\cdot\boldsymbol{\mathcal{E}}) - \nabla^2\boldsymbol{\mathcal{E}}$, valid for a rectilinear coordinate system. Invoke (5.1-9), which yields $\nabla\cdot\epsilon\boldsymbol{\mathcal{E}} = 0$, together with the identity $\nabla\cdot\epsilon\boldsymbol{\mathcal{E}} = \epsilon\nabla\cdot\boldsymbol{\mathcal{E}} + \nabla\epsilon\cdot\boldsymbol{\mathcal{E}}$, which provides $\nabla\cdot\boldsymbol{\mathcal{E}} = -(1/\epsilon)\nabla\epsilon\cdot\boldsymbol{\mathcal{E}}$. Finally, substitute in (5.2-16) to obtain (5.2-18).

For media with gradually varying dielectric properties, i.e., when $\epsilon(\mathbf{r})$ varies sufficiently slowly so that it can be assumed constant within distances of the order of a wavelength, the second term on the left-hand side of (5.2-18) is negligible in comparison with the first, so that

$$\nabla^2\boldsymbol{\mathcal{E}} - \frac{1}{c^2(\mathbf{r})}\frac{\partial^2\boldsymbol{\mathcal{E}}}{\partial t^2} \approx 0, \tag{5.2-19}$$

where $c(\mathbf{r}) = 1/\sqrt{\mu_o\epsilon} = c_o/n(\mathbf{r})$ is spatially varying and $n(\mathbf{r}) = \sqrt{\epsilon(\mathbf{r})/\epsilon_o}$ is the refractive index at position $\mathbf{r}$. This relation was invoked without proof in Sec. 2.1, but it is clearly an approximate consequence of Maxwell's equations.

For a homogeneous dielectric medium of refractive index n perturbed by a slowly varying spatially dependent change Δn, it is often useful to write (5.2-19) in the form

$$\nabla^2\mathcal{E} - \frac{1}{c^2}\frac{\partial^2\mathcal{E}}{\partial t^2} \approx -\mathcal{S}, \qquad \mathcal{S} = -\mu_o\frac{\partial^2\Delta\mathcal{P}}{\partial t^2}, \qquad \Delta\mathcal{P} = 2\epsilon_o n\,\Delta n\,\mathcal{E}, \tag{5.2-20}$$

where $c = c_o/n$ is the speed of light in the homogenous medium. Thus, $\mathcal{E}$ satisfies the wave equation with a radiation source $\mathcal{S}$ created by a perturbation of the polarization density $\Delta\mathcal{P}$, which in turn is proportional to Δn and $\mathcal{E}$ itself. These equations may be verified by expanding the term $1/c^2(\mathbf{r})$ in (5.2-19) as $(n+\Delta n)^2/c_o^2 \approx (n^2+2n\Delta n)/c_o^2$ and bringing the perturbation term to the right-hand side of the equation. The term $\Delta\mathcal{P}$ is the perturbation in $\mathcal{P}$, as can be shown by noting that $\mathcal{P} = \epsilon_o\chi\mathcal{E} = \epsilon_o(\epsilon/\epsilon_o - 1)\mathcal{E} = \epsilon_o(n^2-1)\mathcal{E}$, so that $\Delta\mathcal{P} = \epsilon_o\Delta(n^2-1)\mathcal{E} = 2\epsilon_o n\Delta n\,\mathcal{E}$.

Anisotropic Media

The relation between the vectors $\mathcal{P}$ and $\mathcal{E}$ in an anisotropic dielectric medium depends on the direction of the vector $\mathcal{E}$; the requirement that the two vectors remain parallel is not maintained. If the medium is linear, nondispersive, and homogeneous, each component of $\mathcal{P}$ is a linear combination of the three components of $\mathcal{E}$:

$$\mathcal{P}_i = \sum_j \epsilon_o \chi_{ij} \mathcal{E}_j, \tag{5.2-21}$$

where the indices $i, j = 1, 2, 3$ denote the x, y, and z components, respectively.

The dielectric properties of the medium are then described by a 3×3 array of constants $\{\chi_{ij}\}$, which are elements of what is called the **electric susceptibility tensor** $\boldsymbol{\chi}$ (Fig. 5.2-4). A similar relation between $\mathcal{D}$ and $\mathcal{E}$ applies:

$$\mathcal{D}_i = \sum_j \epsilon_{ij} \mathcal{E}_j, \tag{5.2-22}$$

where $\{\epsilon_{ij}\}$ are the elements of the **electric permittivity tensor** $\boldsymbol{\epsilon}$.

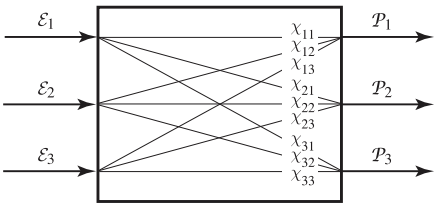


Figure 5.2-4 An anisotropic (but linear, homogeneous, and nondispersive) medium is characterized by nine constants, the components of the electric susceptibility tensor χ_{ij}. Each component of $\mathcal{P}$ is a weighted superposition of the three components of $\mathcal{E}$.

The optical properties of anisotropic media are examined in Chapter 6. The relation between $\mathcal{B}(t)$ and $\mathcal{H}(t)$ for anisotropic magnetic media takes a form similar to that of (5.2-22), under similar assumptions.

Dispersive Media

The relation between the vectors $\mathcal{P}$ and $\mathcal{E}$ in a dispersive dielectric medium is dynamic rather than instantaneous. The vector $\mathcal{E}(t)$ may be thought of as an input that induces the bound electrons in the atoms of the medium to oscillate, which then collectively give rise to the polarization-density vector $\mathcal{P}(t)$ as the output. The presence of a time delay between the output and the input indicates that the system possesses memory. Only when this time is short in comparison with other times of interest can the response be regarded as instantaneous, in which case the medium is approximately nondispersive.

For dispersive media that are linear, homogeneous, and isotropic, the dynamic relation between $\mathcal{P}(t)$ and $\mathcal{E}(t)$ may be described by a linear differential equation such as that associated with a driven harmonic oscillator: $a_1 \, d^2\mathcal{P}/dt^2 + a_2 \, d\mathcal{P}/dt + a_3 \, \mathcal{P} = \mathcal{E}$, where a_1, a_2, and a_3 are constants. A simple analysis along these lines (see Sec. 5.5C) provides a physical rationale for the presence of dispersion (and absorption).

More generally, the linear-systems approach provided in Appendix B may be used to investigate an arbitrary linear system, which is characterized by its response to an impulse (impulse response function). An electric-field impulse of magnitude $\delta(t)$ applied at time $t = 0$ induces a time-dispersed polarization density of magnitude $\epsilon_o \mathrm{x}(t)$, where $\mathrm{x}(t)$ is a scalar function of time with finite duration that begins at $t = 0$. Since the medium is linear, an arbitrary electric field $\mathcal{E}(t)$ then induces a polarization density

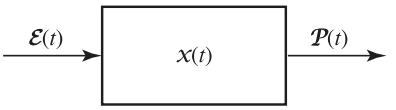


Figure 5.2-5 In a dispersive (but linear, homogeneous, and isotropic) medium the relation between $\mathcal{P}(t)$ and $\mathcal{E}(t)$ is governed by a dynamic linear system described by an impulse response function $\epsilon_o \chi(t)$ that corresponds to a frequency dependent susceptibility $\chi(\nu)$.

that is a superposition of the effects of $\mathcal{E}(t')$ for all $t' \leq t$, so that the polarization density can be expressed as a convolution, as defined in Appendix A:

$$\mathcal{P}(t) = \epsilon_o \int_{-\infty}^{\infty} \chi(t - t')\, \mathcal{E}(t')\, dt'. \tag{5.2-23}$$

This dielectric medium is completely described by its impulse response function $\epsilon_o \chi(t)$.

Alternatively, a dynamic linear system may be described by its transfer function, which governs the response to harmonic inputs. The transfer function is the Fourier transform of the impulse response function (see Appendix B). In the example at hand, the transfer function at frequency ν is $\epsilon_o \chi(\nu)$, where $\chi(\nu)$ is the Fourier transform of $\chi(t)$ so that it is a frequency-dependent susceptibility (Fig. 5.2-5). This concept is discussed further in Secs. 5.3 and 5.5.

For magnetic media under similar assumptions, the relation between $\mathcal{M}(t)$ and $\mathcal{H}(t)$ is analogous to (5.2-23).

Nonlinear Media

A nonlinear dielectric medium is defined as one in which the relation between $\mathcal{P}$ and $\mathcal{E}$ is nonlinear, in which case the wave equation as written in (5.2-9) is not applicable. Rather, Maxwell's equations can be used to derive a nonlinear wave equation that electromagnetic waves obey in a such a medium.

We first derive a general wave equation valid for homogeneous and isotropic nonmagnetic media. Operating on Maxwell's equation (5.1-8) with the curl operator $\nabla\times$, and using the relation $\mathcal{B} = \mu_o \mathcal{H}$ from (5.2-4) together with (5.1-7), we obtain $\nabla\times(\nabla\times\mathcal{E}) = -\mu_o \partial^2 \mathcal{D}/\partial t^2$. Making use of the vector identity $\nabla\times(\nabla\times\mathcal{E}) = \nabla(\nabla\cdot\mathcal{E}) - \nabla^2\mathcal{E}$ and the relation $\mathcal{D} = \epsilon_o \mathcal{E} + \mathcal{P}$ from (5.1-11) then yields

$$\nabla(\nabla\cdot\mathcal{E}) - \nabla^2\mathcal{E} = -\epsilon_o\mu_o \frac{\partial^2 \mathcal{E}}{\partial t^2} - \mu_o \frac{\partial^2 \mathcal{P}}{\partial t^2}. \tag{5.2-24}$$

For homogeneous and isotropic media $\mathcal{D} = \epsilon\mathcal{E}$; thus $\nabla\cdot\mathcal{D} = 0$ from (5.1-9) is equivalent to $\nabla\cdot\mathcal{E} = 0$. Substituting this, along with $\epsilon_o\mu_o = 1/c_o^2$ from (5.1-6), into (5.2-24) therefore provides

$$\boxed{\nabla^2\mathcal{E} - \frac{1}{c_o^2}\frac{\partial^2 \mathcal{E}}{\partial t^2} = \mu_o \frac{\partial^2 \mathcal{P}}{\partial t^2}.} \tag{5.2-25}$$

Wave Equation
(Homogeneous and Isotropic Medium)

Equation (5.2-25) is applicable for all homogeneous and isotropic dielectric media: nonlinear or linear, nondispersive or dispersive.

Now, if the medium is nonlinear, nondispersive, and nonmagnetic, the polarization density $\mathcal{P}$ can be written as a memoryless nonlinear function of $\mathcal{E}$, say $\mathcal{P} = \Psi(\mathcal{E})$, valid at every position and time. (The simplest example of such a function is $\mathcal{P} =$

$a_1\,\mathcal{E}+a_2\,\mathcal{E}^2$, where a_1 and a_2 are constants.) Under these conditions (5.2-25) becomes a nonlinear partial differential equation for the electric-field vector $\mathcal{E}(\mathbf{r},t)$:

$$\nabla^2\mathcal{E}-\frac{1}{c_o^2}\frac{\partial^2\mathcal{E}}{\partial t^2}=\mu_o\frac{\partial^2\Psi(\mathcal{E})}{\partial t^2}. \tag{5.2-26}$$

The principle of superposition is no longer applicable by virtue of the nonlinear nature of this wave equation. Nonlinear magnetic materials may be similarly described.

Most dielectric media are approximately linear unless the optical intensity is substantial, as in the case of focused laser beams. Nonlinear optics is discussed in Chapter 22.

5.3 MONOCHROMATIC ELECTROMAGNETIC WAVES

For the special case of *monochromatic* electromagnetic waves in an optical medium, all components of the electric and magnetic fields are harmonic functions of time with the same frequency ν and corresponding angular frequency $\omega=2\pi\nu$. Adopting the complex representation used in Sec. 2.2A, these six real field components may be expressed as

$$\begin{aligned}\mathcal{E}(\mathbf{r},t)&=\text{Re}\{\mathbf{E}(\mathbf{r})\,\exp(j\omega t)\}\\ \mathcal{H}(\mathbf{r},t)&=\text{Re}\{\mathbf{H}(\mathbf{r})\exp(j\omega t)\},\end{aligned} \tag{5.3-1}$$

where $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ represent electric- and magnetic-field complex-amplitude vectors, respectively. Analogous complex-amplitude vectors $\mathbf{P}$, $\mathbf{D}$, $\mathbf{M}$, and $\mathbf{B}$ are similarly associated with the real vectors $\mathcal{P}$, $\mathcal{D}$, $\mathcal{M}$, and $\mathcal{B}$, respectively.

Maxwell's Equations in a Medium

Inserting (5.3-1) into Maxwell's equations (5.1-7)–(5.1-10), and using the relation $(\partial/\partial t)\,e^{j\omega t}=j\omega\,e^{j\omega t}$ for monochromatic waves of angular frequency ω, yields a set of equations obeyed by the field complex-amplitude vectors:

$$\nabla\times\mathbf{H}=j\omega\mathbf{D} \tag{5.3-2}$$
$$\nabla\times\mathbf{E}=-j\omega\mathbf{B} \tag{5.3-3}$$
$$\nabla\cdot\mathbf{D}=0 \tag{5.3-4}$$
$$\nabla\cdot\mathbf{B}=0. \tag{5.3-5}$$

Maxwell's Equations (Source-Free Medium; Monochromatic Fields)

Similarly, (5.1-11) and (5.1-12) give rise to

$$\mathbf{D}=\epsilon_o\mathbf{E}+\mathbf{P} \tag{5.3-6}$$
$$\mathbf{B}=\mu_o\mathbf{H}+\mu_o\mathbf{M}. \tag{5.3-7}$$

Intensity and Power

As indicated in Sec. 5.1, the flow of electromagnetic power is governed by the time average of the Poynting vector $\mathcal{S}=\mathcal{E}\times\mathcal{H}$. Casting this expression in terms of complex

amplitudes yields

$$\mathcal{S} = \text{Re}\left\{\mathbf{E}e^{j\omega t}\right\} \times \text{Re}\left\{\mathbf{H}e^{j\omega t}\right\} = \tfrac{1}{2}\left(\mathbf{E}e^{j\omega t} + \mathbf{E}^* e^{-j\omega t}\right) \times \tfrac{1}{2}\left(\mathbf{H}e^{j\omega t} + \mathbf{H}^* e^{-j\omega t}\right)$$
$$= \tfrac{1}{4}\left(\mathbf{E} \times \mathbf{H}^* + \mathbf{E}^* \times \mathbf{H} + e^{j2\omega t}\,\mathbf{E} \times \mathbf{H} + e^{-j2\omega t}\,\mathbf{E}^* \times \mathbf{H}^*\right). \tag{5.3-8}$$

The terms containing the factors $e^{j2\omega t}$ and $e^{-j2\omega t}$ oscillate at optical frequencies and are therefore washed out by the averaging process, which is slow in comparison with an optical cycle. Thus,

$$\langle \mathcal{S} \rangle = \tfrac{1}{4}(\mathbf{E} \times \mathbf{H}^* + \mathbf{E}^* \times \mathbf{H}) = \tfrac{1}{2}(\mathbf{S} + \mathbf{S}^*) = \text{Re}\{\mathbf{S}\}, \tag{5.3-9}$$

where the vector

$$\boxed{\mathbf{S} = \tfrac{1}{2}\mathbf{E} \times \mathbf{H}^*} \tag{5.3-10}$$

Complex Poynting Vector

may be regarded as a complex Poynting vector. The optical intensity is the magnitude of the vector $\text{Re}\{\mathbf{S}\}$.

Linear, Nondispersive, Homogeneous, and Isotropic Media

For monochromatic waves, the relations provided in (5.2-2) and (5.2-4) become the **material equations**

$$\mathbf{D} = \epsilon\mathbf{E} \qquad \text{and} \qquad \mathbf{B} = \mu\mathbf{H}, \tag{5.3-11}$$

so that Maxwell's equations, (5.3-2)–(5.3-5), depend solely on the interrelated complex-amplitude vectors $\mathbf{E}$ and $\mathbf{H}$:

$$\boxed{\begin{aligned} \nabla \times \mathbf{H} &= j\omega\epsilon\mathbf{E} \\ \nabla \times \mathbf{E} &= -j\omega\mu\mathbf{H} \\ \nabla \cdot \mathbf{E} &= 0 \\ \nabla \cdot \mathbf{H} &= 0. \end{aligned}} \qquad \begin{aligned} &(5.3\text{-}12)\\ &(5.3\text{-}13)\\ &(5.3\text{-}14)\\ &(5.3\text{-}15) \end{aligned}$$

Maxwell's Equations
(Linear, Nondispersive, Homogeneous, Isotropic, Source-Free Medium; Monochromatic Light)

Substituting the electric and magnetic fields $\mathcal{E}$ and $\mathcal{H}$ given in (5.3-1) into the wave equation (5.2-9) yields the Helmholtz equation

$$\boxed{\nabla^2 U + k^2 U = 0,} \qquad k = nk_o = \omega\sqrt{\epsilon\,\mu} \tag{5.3-16}$$

Helmholtz Equation

where the scalar function $U = U(\mathbf{r})$ represents the complex amplitude of any of the three components (E_x, E_y, E_z) of $\mathbf{E}$ or three components (H_x, H_y, H_z) of $\mathbf{H}$; and where $n = \sqrt{(\epsilon/\epsilon_o)(\mu/\mu_o)}$, $k_o = \omega/c_o$, and $c = c_o/n$. In the context of wave optics, the Helmholtz equation in (2.2-7) was written in terms of the complex amplitude $U(\mathbf{r})$ of the real wavefunction $u(\mathbf{r}, t)$.

Inhomogeneous Media

In an inhomogeneous nonmagnetic medium, Maxwell's equations (5.3-12)–(5.3-15) remain applicable, but the electric permittivity of the medium becomes position dependent, $\epsilon = \epsilon(\mathbf{r})$. For locally homogeneous media in which $\epsilon(\mathbf{r})$ varies slowly with respect to the wavelength, the Helmholtz equation (5.3-16) remains approximately valid, subject to the substitutions $k = n(\mathbf{r})k_o$ and $n(\mathbf{r}) = \sqrt{\epsilon(\mathbf{r})/\epsilon_o}$.

Dispersive Media

In a dispersive dielectric medium, $\mathcal{P}(t)$ and $\mathcal{E}(t)$ are connected by the dynamic relation provided in (5.2-23). To determine the corresponding relation between the complex-amplitude vectors $\mathbf{P}$ and $\mathbf{E}$, we substitute (5.3-1) into (5.2-23), which gives rise to

$$\boxed{\mathbf{P} = \epsilon_o \chi(\nu)\mathbf{E}} \tag{5.3-17}$$

where

$$\chi(\nu) = \int_{-\infty}^{\infty} \mathrm{x}(t) \exp(-j2\pi\nu t)\, dt \tag{5.3-18}$$

is the Fourier transform of $\mathrm{x}(t)$.

Equation (5.3-17) can also be directly inferred from (5.2-23) by invoking the convolution theorem: convolution in the time domain corresponds to multiplication in the frequency domain (see Secs. A.1 and B.1 of Appendices A and B, respectively), and recognizing $\mathbf{E}$ and $\mathbf{P}$ as the components of $\mathcal{E}$ and $\mathcal{P}$ of frequency ν. The function $\epsilon_o\chi(\nu)$ may therefore be regarded as the transfer function of the linear system that relates $\mathcal{P}(t)$ to $\mathcal{E}(t)$.

The relation between $\mathcal{D}$ and $\mathcal{E}$ is similar,

$$\mathbf{D} = \epsilon(\nu)\mathbf{E} \tag{5.3-19}$$

where

$$\epsilon(\nu) = \epsilon_o[1 + \chi(\nu)]. \tag{5.3-20}$$

Therefore, in dispersive media the susceptibility χ and the permittivity ϵ are frequency-dependent and, in general, complex-valued quantities. The Helmholtz equation (5.3-16) is thus readily adapted for use in dispersive nonmagnetic media by taking

$$k = \omega\sqrt{\epsilon(\nu)\,\mu_o}. \tag{5.3-21}$$

When $\chi(\nu)$ and $\epsilon(\nu)$ are approximately constant within the frequency band of interest, the medium may be treated as approximately nondispersive. The implications of the complex-valued nature of χ and k in dispersive media are discussed further in Sec. 5.5.

5.4 ELEMENTARY ELECTROMAGNETIC WAVES

A. Plane, Dipole, and Gaussian Electromagnetic Waves

We now examine three elementary solutions to Maxwell's equations that are of substantial importance in optics: plane waves and spherical (dipole) waves, which were discussed in Sec. 2.2B in the context of wave optics, and the Gaussian beam, which was studied in Chapter 3 using the wave-optics formalism. The medium is assumed to be linear, homogeneous, nondispersive, and isotropic, and the waves are assumed to be monochromatic.

The Transverse Electromagnetic (TEM) Plane Wave

Consider a monochromatic electromagnetic wave whose magnetic- and electric-field complex-amplitude vectors are plane waves with wavevector $\mathbf{k}$ (see Sec. 2.2B) so that

$$\mathbf{H}(\mathbf{r}) = \mathbf{H}_0 \exp(-j\mathbf{k}\cdot\mathbf{r}) \tag{5.4-1}$$

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_0 \exp(-j\mathbf{k}\cdot\mathbf{r}), \tag{5.4-2}$$

where the complex envelopes $\mathbf{H}_0$ and $\mathbf{E}_0$ are constant vectors. All components of $\mathbf{H}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r})$ satisfy the Helmholtz equation provided that the magnitude of $\mathbf{k}$ is $k = nk_o$, where n is the refractive index of the medium.

We now examine the conditions that must be obeyed by $\mathbf{H}_0$ and $\mathbf{E}_0$ in order that Maxwell's equations be satisfied. Substituting (5.4-1) and (5.4-2) into Maxwell's equations (5.3-12) and (5.3-13), respectively, leads to

$$\mathbf{k}\times\mathbf{H}_0 = -\omega\,\epsilon\,\mathbf{E}_0 \tag{5.4-3}$$

$$\mathbf{k}\times\mathbf{E}_0 = \omega\,\mu\,\mathbf{H}_0. \tag{5.4-4}$$

The other two Maxwell's equations, (5.3-14) and (5.3-15), are satisfied identically since the divergence of a uniform plane wave is zero.

It follows from (5.4-3) that $\mathbf{E}$ must be perpendicular to both $\mathbf{k}$ and $\mathbf{H}$ and from (5.4-4) that $\mathbf{H}$ must be perpendicular to both $\mathbf{k}$ and $\mathbf{E}$. Thus, $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{k}$ are mutually orthogonal, as illustrated in Fig. 5.4-1. Since $\mathbf{E}$ and $\mathbf{H}$ lie in a plane normal to the direction of propagation $\mathbf{k}$, the wave is called a **transverse electromagnetic (TEM)** wave.

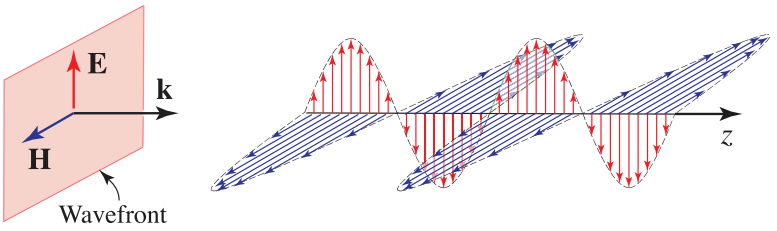


Figure 5.4-1 The TEM plane wave. The vectors $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{k}$ are mutually orthogonal. The wavefronts (surfaces of constant phase) are normal to the wavevector $\mathbf{k}$.

In accordance with (5.4-3), the magnitudes H_0 and E_0 are related by $H_0 = (\omega\epsilon/k)E_0$. Similarly, (5.4-4) yields $H_0 = (k/\omega\mu)E_0$. For these two equations to be consistent, we must have $\omega\epsilon/k = k/\omega\mu$, or $k = \omega\sqrt{\epsilon\,\mu} = \omega/c = n\omega/c_o = nk_o$.

This is, in fact, the same condition required in order that the wave satisfy the Helmholtz equation.

The ratio between the amplitudes of the electric and magnetic fields is $E_0/H_0 = \omega\mu/k = c\mu = \sqrt{\mu/\epsilon}$. This quantity is known as the **impedance** of the medium,

$$\boxed{\eta = \frac{E_0}{H_0} = \sqrt{\frac{\mu}{\epsilon}}\,.} \tag{5.4-5}$$

Impedance

For nonmagnetic media $\mu = \mu_o$, whereupon $\eta = \sqrt{\mu_o/\epsilon}$ may be defined in terms of the impedance of free space η_o via

$$\boxed{\eta = \frac{\eta_o}{n}\,,} \tag{5.4-6}$$

Impedance (Nonmagnetic Media)

where

$$\eta_o = \sqrt{\frac{\mu_o}{\epsilon_o}} \approx 120\pi \approx 377\,\Omega. \tag{5.4-7}$$

The complex Poynting vector $\mathbf{S} = \frac{1}{2}\mathbf{E} \times \mathbf{H}^*$ [see (5.3-10)] is parallel to the wavevector $\mathbf{k}$, so that the power flows along a direction normal to the wavefronts. Its magnitude is $\frac{1}{2}E_0H_0^* = |E_0|^2/2\eta$, and the intensity I is therefore given by

$$\boxed{I = \frac{|E_0|^2}{2\eta}\,.} \tag{5.4-8}$$

Intensity

The intensity of a TEM wave is thus seen to be proportional to the absolute square of the complex envelope of the electric field. As an example, an intensity of 10 W/cm^2 in free space corresponds to an electric field of $\approx$ 87 V/cm. Note the similarity between (5.4-8) and the relation $I = |U|^2$, which was defined for scalar waves in Sec. 2.2A.

Equation (5.2-15) provides that the time-averaged energy density $W = \langle \mathcal{W} \rangle$ of the plane wave is

$$W = \tfrac{1}{2}\epsilon|E_0|^2, \tag{5.4-9}$$

since the electric and magnetic contributions are equal, i.e., $\frac{1}{2}\epsilon|E_0|^2/2 = \frac{1}{2}\mu|H_0|^2/2$. The intensity in (5.4-8) and the time-averaged energy density in (5.4-9) are therefore related by

$$I = cW, \tag{5.4-10}$$

indicating that the time-averaged power density flow I results from the transport of the time-averaged energy density at the velocity of light c. This is readily visualized by considering a cylinder of area A and length c whose axis lies parallel to the direction of propagation. The energy stored in the cylinder, cAW, is transported across the area in one second, confirming that the intensity (power per unit area) is $I = cW$.

The linear momentum density (per unit volume) transported by a plane wave is $(1/c^2)\,\mathcal{S} = (1/c^2)I\,\widehat{k} = (W/c)\,\widehat{k}$.

The Dipole Wave

An oscillating **electric dipole** radiates a wave with features that resemble the scalar spherical wave discussed in Sec. 2.2B. The radiation frequency is determined by the frequency at which the dipole oscillates. This electromagnetic wave is readily constructed from an auxiliary vector field $\mathbf{A}(\mathbf{r})$, known as the **vector potential**, which is often used to facilitate the solution of Maxwell's equations in electromagnetics. For the case at hand we set

$$\mathbf{A}(\mathbf{r}) = \mathfrak{a}_0\, U(\mathbf{r})\, \widehat{\mathbf{x}}, \tag{5.4-11}$$

where $\mathfrak{a}_0$ is a constant and $\widehat{\mathbf{x}}$ is a unit vector in the direction of the dipole (the x direction). The quantity $U(\mathbf{r})$ represents a scalar spherical wave with the origin at $r = 0$:

$$U(\mathbf{r}) = \frac{1}{4\pi r} \exp(-jkr). \tag{5.4-12}$$

Because $U(\mathbf{r})$ satisfies the Helmholtz equation, as was established in Sec. 2.2B, $\mathbf{A}(\mathbf{r})$ will also satisfy the Helmholtz equation $\nabla^2\mathbf{A} + k^2\mathbf{A} = 0$.

We now define the magnetic field in terms of the curl of this vector

$$\mathbf{H} = \frac{1}{\mu} \nabla \times \mathbf{A}, \tag{5.4-13}$$

and determine the corresponding electric field from Maxwell's equation (5.3-12):

$$\mathbf{E} = \frac{1}{j\omega\epsilon} \nabla \times \mathbf{H}. \tag{5.4-14}$$

The form of (5.4-13) and (5.4-14) ensures that $\nabla \cdot \mathbf{E} = 0$ and $\nabla \cdot \mathbf{H} = 0$, as required by (5.3-14) and (5.3-15), since the divergence of the curl of any vector field vanishes. Because $\mathbf{A}(\mathbf{r})$ satisfies the Helmholtz equation, it can readily be shown that the remaining Maxwell's equation, $\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H}$, is also satisfied. It is therefore clear that (5.4-11)–(5.4-14) define a valid electromagnetic wave that satisfies Maxwell's equations.

Explicit expressions for $\mathbf{E}$ and $\mathbf{H}$ are obtained by carrying out the curl operations prescribed in (5.4-13) and (5.4-14). This is conveniently accomplished by making use of the spherical coordinate system (r, θ, ϕ) defined in Fig. 5.4-2(a), with unit vectors $(\widehat{\mathbf{r}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}})$. The exact results for $\mathbf{E}$ and $\mathbf{H}$ turn out to be

$$\mathbf{E}(\mathbf{r}) = 2\mathfrak{e}_0 \cos\theta \left[\frac{1}{jkr} + \frac{1}{(jkr)^2}\right] U(\mathbf{r})\, \widehat{\mathbf{r}} + \mathfrak{e}_0 \sin\theta \left[1 + \frac{1}{jkr} + \frac{1}{(jkr)^2}\right] U(\mathbf{r})\, \widehat{\boldsymbol{\theta}}, \tag{5.4-15}$$

$$\mathbf{H}(\mathbf{r}) = \mathfrak{h}_0 \sin\theta \left[1 + \frac{1}{jkr}\right] U(\mathbf{r})\, \widehat{\boldsymbol{\phi}}, \tag{5.4-16}$$

where $\mathfrak{h}_0 = (jk/\mu)A_0$ and $\mathfrak{e}_0 = \eta H_0$. It can be shown that an electric dipole moment $\mathfrak{p}$ pointing in the x direction radiates the wave described in (5.4-15) and (5.4-16) with $\mathfrak{a}_0 = j\mu\omega\mathfrak{p}$, so that $\mathfrak{h}_0 = (-\omega^2/c)\mathfrak{p}$ and $\mathfrak{e}_0 = -\mu\omega^2\mathfrak{p}$.

For points at distances from the origin that are much greater than a wavelength ($kr = 2\pi r/\lambda \gg 1$), the complex-amplitude vectors in (5.4-15) and (5.4-16) may be

approximated by

$$\mathbf{E}(\mathbf{r}) \approx e_0 \sin\theta\, U(\mathbf{r})\, \widehat{\boldsymbol{\theta}}, \tag{5.4-17}$$

$$\mathbf{H}(\mathbf{r}) \approx h_0 \sin\theta\, U(\mathbf{r})\, \widehat{\boldsymbol{\phi}}. \tag{5.4-18}$$

The wavefronts are then spherical (as for the scalar spherical wave) and, as illustrated in Fig. 5.4-2(b), the electric and magnetic fields are orthogonal to one another and to the radial direction $\widehat{\mathbf{r}}$, with the electric field pointing in the polar direction and the magnetic field pointing in the azimuthal direction. Since the field strength is proportional to $\sin\theta$, the radiation pattern is toroidal with a null in the direction of the dipole, as shown in 5.4-2(c). In the paraxial approximation, at points near the z axis and far from the origin, such that $\theta \approx \pi/2$ and $\phi \approx \pi/2$, the wavefront normals are nearly parallel to the z axis (corresponding to paraxial rays), and $\sin\theta \approx 1$.

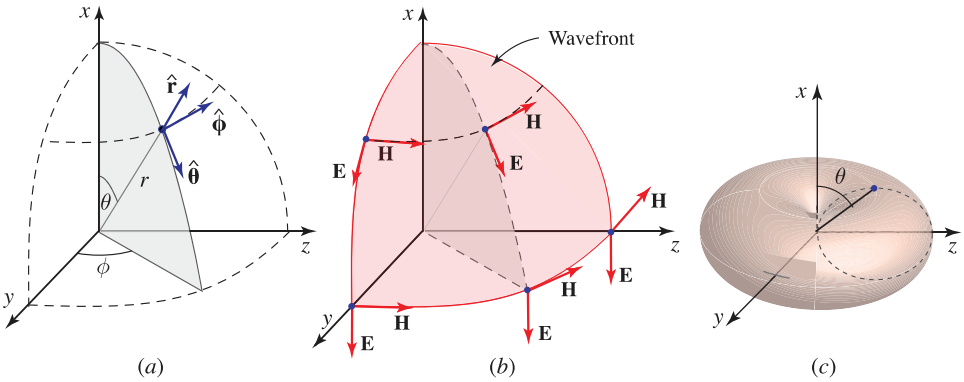


Figure 5.4-2 (a) Spherical coordinate system. (b) Electric- and magnetic-field vectors and wavefronts of the electromagnetic field, at distances $r \gg \lambda/2\pi$, radiated by an oscillating electric dipole. (c) The radiation pattern (field magnitude versus polar angle θ) is toroidal.

In a Cartesian coordinate system, $\widehat{\boldsymbol{\theta}} = -\sin\theta\, \widehat{\mathbf{x}} + \cos\theta\cos\phi\, \widehat{\mathbf{y}} + \cos\theta\sin\phi\, \widehat{\mathbf{z}} \approx -\widehat{\mathbf{x}} + (x/z)(y/z)\, \widehat{\mathbf{y}} + (x/z)\, \widehat{\mathbf{z}} \approx -\widehat{\mathbf{x}} + (x/z)\, \widehat{\mathbf{z}}$, so that

$$\mathbf{E}(\mathbf{r}) \approx e_0 \left(-\widehat{\mathbf{x}} + \frac{x}{z}\, \widehat{\mathbf{z}}\right) U(\mathbf{r}), \tag{5.4-19}$$

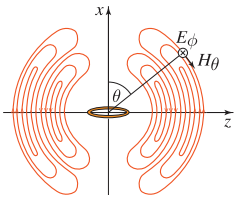
where $U(\mathbf{r})$ is the paraxial approximation of the spherical wave, i.e., the paraboloidal wave discussed in Sec. 2.2B. For sufficiently large values of z, the term (x/z) in (5.4-19) may also be neglected, whereupon

$$\mathbf{E}(\mathbf{r}) \approx -e_0\, U(\mathbf{r})\, \widehat{\mathbf{x}}, \tag{5.4-20}$$

$$\mathbf{H}(\mathbf{r}) \approx \quad h_0\, U(\mathbf{r})\, \widehat{\mathbf{y}}. \tag{5.4-21}$$

In this approximation $U(\mathbf{r})$ approaches $(1/4\pi z)\, e^{-jkz}$, so that a TEM plane wave ultimately emerges, as portrayed in Fig. 2.2-4.

An electromagnetic wave that is dual to the electric-dipole wave discussed above is radiated by a **magnetic dipole** with the magnetic dipole moment $\mathfrak{m}$ pointing in the x direction. In the far field ($kr \gg 1$), it has an electric field pointing in the azimuthal direction and an orthogonal magnetic field pointing in the polar direction, with complex-amplitude vectors given by



$$\mathbf{H}(\mathbf{r}) \approx \mathrm{h}_0 \sin\theta\, U(\mathbf{r})\, \widehat{\boldsymbol{\theta}}, \tag{5.4-22}$$

$$\mathbf{E}(\mathbf{r}) \approx \mathrm{e}_0 \sin\theta\, U(\mathbf{r})\, \widehat{\boldsymbol{\phi}}, \tag{5.4-23}$$

where $\mathrm{h}_0 = (\omega^2/c^2)\mathrm{m}$ and $\mathrm{e}_0 = \mu(\omega^2/c)\mathrm{m}$. At radio frequencies, this type of wave is radiated by electric current flowing in a loop antenna placed in a plane orthogonal to the x axis. At optical frequencies, tiny metal loops serve as optical antennas (Sec. 8.2D) and as important components in metamaterials (Sec. 8.3A).

The Gaussian Beam

It was demonstrated in Sec. 3.1 that a scalar Gaussian beam is readily obtained from a paraboloidal wave (the paraxial approximation to a spherical wave) by replacing the coordinate z by $z + jz_0$, where z_0 is a real constant.

The same transformation applied to the corresponding electromagnetic wave leads to the electromagnetic **vector Gaussian beam**. Replacing z in (5.4-19) by $z + jz_0$ yields

$$\mathbf{E}(\mathbf{r}) = \mathrm{e}_0 \left(-\widehat{\mathbf{x}} + \frac{x}{z + jz_0}\,\widehat{\mathbf{z}}\right) U(\mathbf{r}), \tag{5.4-24}$$

where $U(\mathbf{r})$ now represents the scalar complex amplitude of a Gaussian beam provided in (3.1-7). The wavefronts of the Gaussian beam are illustrated in Fig. 5.4-3(a) (these are also shown in Fig. 3.1-7), while the $\mathbf{E}$-field lines determined from (5.4-24) are displayed in Fig. 5.4-3(b). In this case, the direction of the $\mathbf{E}$ field is not spatially uniform.

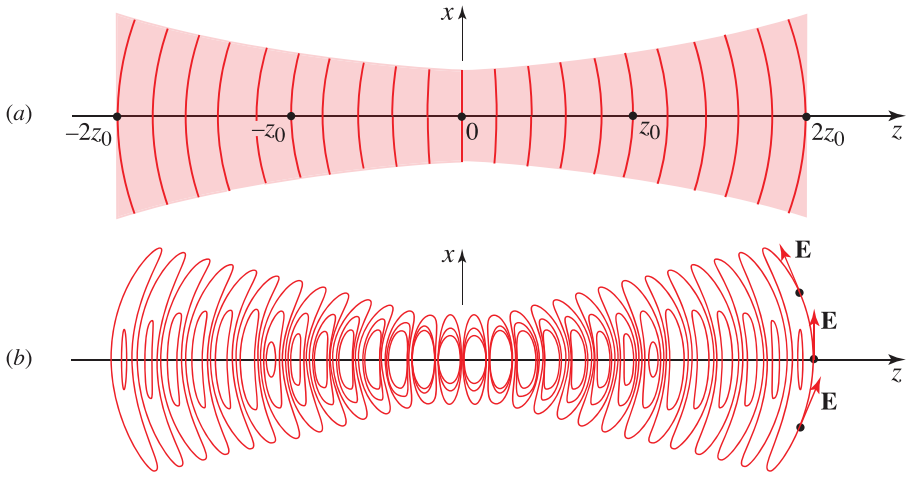


Figure 5.4-3 (a) Wavefronts of the scalar Gaussian beam $U(\mathbf{r})$ in the x–z plane. (b) Electric-field lines of the electromagnetic Gaussian beam in the x–z plane. (Adapted from H. A. Haus, *Waves and Fields in Optoelectronics*, Prentice Hall, 1984, Fig. 5.3a.)

B. Relation Between Electromagnetic Optics and Scalar Wave Optics

The paraxial scalar wave, defined in Sec. 2.2C, has wavefront normals that form small angles with respect to the axial coordinate z. The wavefronts behave locally as plane waves while the complex envelope and direction of propagation vary slowly with z.

This notion is also applicable to electromagnetic waves in linear isotropic media. A paraxial electromagnetic wave is locally approximated by a TEM plane wave. At each point, the vectors $\mathbf{E}$ and $\mathbf{H}$ lie in a plane that is tangential to the wavefront surfaces and normal to the wavevector $\mathbf{k}$ (Fig. 5.4-4). The optical power flows along the direction $\mathbf{E} \times \mathbf{H}$, which is parallel to $\mathbf{k}$ and approximately parallel to the coordinate z.

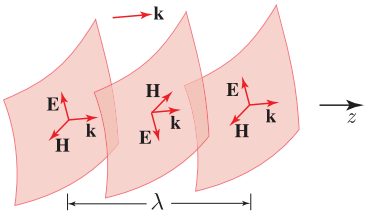


Figure 5.4-4 The paraxial electromagnetic wave. The vectors $\mathbf{E}$ and $\mathbf{H}$ reverse directions after propagation a distance of a half wavelength.

A paraxial scalar wave of intensity $I = |U|^2$ [see (2.2-10)] may be associated with a paraxial electromagnetic wave of the same intensity $I = |E|^2/2\eta$ [see (5.4-8)] by setting the complex amplitude to $U = E/\sqrt{2\eta}$ and matching the wavefronts. As attested to by the extensive development provided in Chapters 2–4, the scalar-wave description of light provides a very good approximation for solving a great many problems involving the interference, diffraction, propagation, and imaging of paraxial waves. The Gaussian beam with small divergence angle, considered in Chapter 3, provides a case in point. Most features of these beams, such as their intensity, focusing by a lens, reflection from a mirror, and interference, are addressed satisfactorily within the context of scalar wave optics. Of course, when polarization comes into play, wave optics is mute and we must appeal to electromagnetic optics.

It is of interest to note that U (as defined above) and E do not satisfy the same boundary conditions. For an electric field tangential to the boundary between two dielectric media, for example, E is continuous (Fig. 5.1-1), but $U = E/\sqrt{2\eta}$ is discontinuous since η changes value at the boundary. Thus, problems involving reflection and refraction at boundaries cannot be addressed completely within the scalar wave theory, although the matching of phase that leads to the law of reflection and Snell's law is adequately carried out within its confines (Sec. 2.4). Indeed, calculations of reflectance and transmittance at a boundary depend on the polarization state of the light and therefore require electromagnetic optics (see Sec. 6.2). Similarly, problems involving the transmission of light through dielectric waveguides require an analysis based on electromagnetic theory, as discussed in Chapters 9 and 10.

C. Vector Beams

Maxwell's equations in the paraxial approximation admit other cylindrically symmetric beam solutions for which the direction of the electric-field vector is spatially nonuniform. One example is a beam in which the electric field is aligned in an azimuthal orientation with respect to the beam axis, as illustrated in Fig. 5.4-5(a), i.e.,

$$\mathbf{E}(\mathbf{r}) = U(\rho, z) \exp(-jkz)\widehat{\boldsymbol{\phi}}. \tag{5.4-25}$$

The scalar function $U(\rho, z)$ turns out to be the Bessel–Gauss solution to the Helmholtz equation, as discussed in Sec. 3.5A. This beam vanishes on-axis ($\rho = 0$) and has a toroidal transverse spatial distribution. The beam diverges in the axial direction and the spot size increases, much like the Gaussian beam.

Yet another cylindrically-symmetric beam has an azimuthally oriented magnetic-field vector, so that the electric-field vector is radial, as illustrated schematically in Fig. 5.4-5(b). It also has a spatial distribution with an on-axis null. The distribution of the vector field of this beam bears some resemblance to the electromagnetic field radiated by a dipole oriented along the beam axis (see Fig. 5.4-2).

It has been shown that a vector beam with radial electric-field vector may be focused by a lens of large numerical aperture to a spot of significantly smaller size than is possible with a conventional scalar Gaussian beam. Clearly, applications for such beams find a place in high-resolution microscopy. There are, it turns out, many variations on the theme of optical vector beams and their uses.

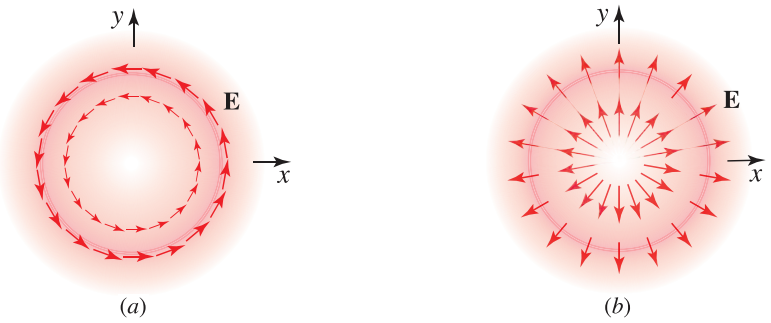


Figure 5.4-5 Vector beams with cylindrical symmetry. (a) Electric-field vectors oriented in the azimuthal direction. (b) Electric-field vectors oriented in the radial direction. The shading indicates the spatial distribution of the optical intensity in the transverse plane.

5.5 ABSORPTION AND DISPERSION

In this section, we consider absorption and dispersion in nonmagnetic media.

A. Absorption

The dielectric media considered thus far have been assumed to be fully transparent, i.e., not to absorb light. Glass is such a material in the visible region of the optical spectrum but it is, in fact, absorptive in the ultraviolet and infrared regions. Transmissive optical components in those bands are fabricated from other materials: examples are quartz and magnesium fluoride in the ultraviolet; and germanium and barium fluoride in the infrared. Figure 5.5-1 illustrates the spectral windows within which some commonly encountered optical materials are transparent (see Sec. 14.1D for further discussion).

In this section, we adopt a phenomenological approach to the absorption of light in linear media. Consider a complex electric susceptibility

$$\chi = \chi' + j\chi'', \tag{5.5-1}$$

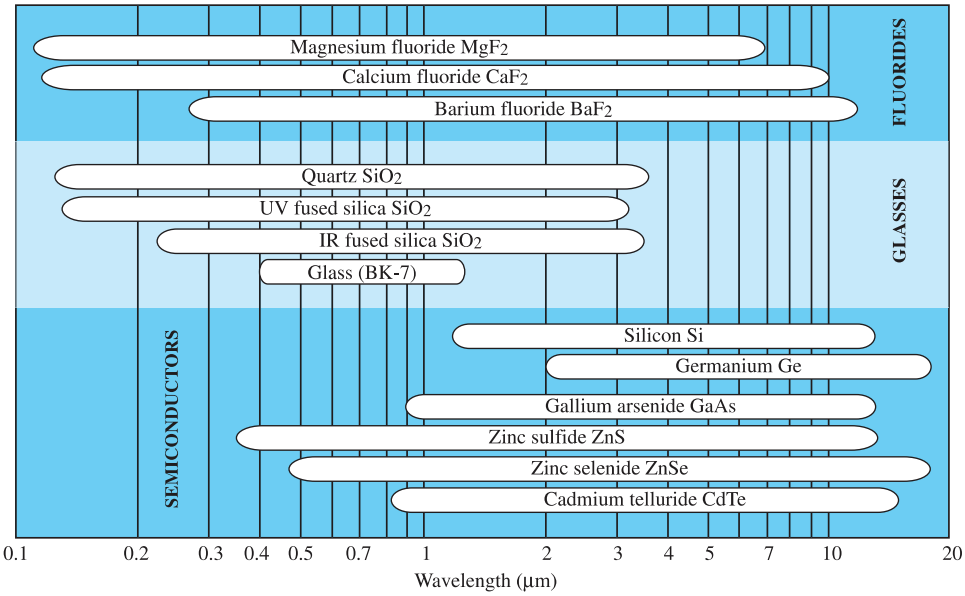


Figure 5.5-1 The white regions indicate the spectral bands within which the specified optical materials transmit light. Selected fluorides, glasses, and semiconductors are displayed.

corresponding to a complex electric permittivity $\epsilon = \epsilon_o(1+\chi)$ and a complex relative permittivity $\epsilon/\epsilon_o = (1+\chi)$. For monochromatic light, the Helmholtz equation (5.3-16) for the complex amplitude $U(\mathbf{r})$ remains valid, $\nabla^2 U + k^2 U = 0$, *but the wavenumber* k *itself becomes complex-valued*:

$$k = \omega\sqrt{\epsilon\,\mu_o} = k_o\sqrt{1+\chi} = k_o\sqrt{1+\chi'+j\chi''}, \tag{5.5-2}$$

where $k_o = \omega/c_o$ is the wavenumber in free space.

Writing k in terms of real and imaginary parts, $k = \beta - j\frac{1}{2}\alpha$, allows β and α to be related to the susceptibility components χ' and χ'':

$$k = \beta - j\tfrac{1}{2}\alpha = k_o\sqrt{1+\chi'+j\chi''}. \tag{5.5-3}$$

As a result of the imaginary part of k, a plane wave with complex amplitude $U = A\exp(-jkz)$ traveling through such a medium in the z-direction undergoes a change in magnitude (as well as the usual change in phase). Substituting $k = \beta - j\frac{1}{2}\alpha$ into the exponent of this plane wave yields $U = A\exp(-\frac{1}{2}\alpha z)\exp(-j\beta z)$. For $\alpha > 0$, which corresponds to absorption in the medium, the envelope A of the original plane wave is attenuated by the factor $\exp(-\frac{1}{2}\alpha z)$ so that the intensity, which is proportional to $|U|^2$, is attenuated by $|\exp(-\frac{1}{2}\alpha z)|^2 = \exp(-\alpha z)$. The coefficient α is therefore recognized as the **absorption coefficient** (also called the **attenuation coefficient**) of the medium. This simple exponential decay formula for the intensity provides the rationale for writing the imaginary part of k as $-\frac{1}{2}\alpha$. It will be seen in Sec. 15.1A that certain media, such as those used in lasers, can exhibit $\alpha < 0$, in which case $\gamma \equiv -\alpha$ is called the *gain coefficient* and the medium amplifies rather than attenuates light.

Since the parameter β is the rate at which the phase changes with z, it represents the propagation constant of the wave. The medium therefore has an effective refractive

index n defined by

$$\beta = nk_o, \tag{5.5-4}$$

and the wave travels with a phase velocity $c = c_o/n$.

Substituting (5.5-4) into (5.5-3) thus relates the refractive index n and the absorption coefficient α to the real and imaginary parts of the susceptibility χ' and χ'':

$$\boxed{n - j\frac{1}{2}\frac{\alpha}{k_o} = \sqrt{\epsilon/\epsilon_o} = \sqrt{1 + \chi' + j\chi''}.} \tag{5.5-5}$$

Absorption Coefficient and Refractive Index

Note that the square root in (5.5-5) provides two complex numbers with opposite signs (phase difference of π). The sign is selected such that if χ'' is negative, i.e., the medium is absorbing, then α is positive, i.e., the wave is attenuated. If $(1 + \chi')$ is positive, then the complex number $1+\chi'+j\chi''$ is in the fourth quadrant, and its square root can be in either the second or the fourth quadrant. By selecting the value in the fourth quadrant, we ensure that α is positive, and n is then also positive. Similarly, if $(1+\chi')$ is negative, then $1 + \chi' + j\chi''$ is in the third quadrant, and its square root is selected to be in the fourth quadrant so that both α and n are positive. The impedance associated with the complex susceptibility χ, which is also complex, is given by

$$\boxed{\eta = \sqrt{\frac{\mu_o}{\epsilon}} = \frac{\eta_o}{\sqrt{1+\chi}}.} \tag{5.5-6}$$

Impedance

Hence, in the context of our formulation, χ, k, ϵ, and η are complex quantities while α, β, and n are real.

Weakly Absorbing Media

In a weakly absorbing medium, we have the condition $\chi'' \ll 1 + \chi'$, so that $\sqrt{1 + \chi' + j\chi''} = \sqrt{1 + \chi'}\sqrt{1 + j\delta} \approx \sqrt{1 + \chi'}(1 + j\frac{1}{2}\delta)$, where $\delta = \chi''/(1 + \chi')$. It follows from (5.5-5) that

$$\boxed{\begin{aligned} n &\approx \sqrt{1+\chi'} \\ \alpha &\approx -\frac{k_o}{n}\chi''. \end{aligned}} \tag{5.5-7}$$

(5.5-8)

Weakly Absorbing Medium

Under these circumstances, the refractive index is determined by the real part of the susceptibility and the absorption coefficient is proportional to the imaginary part thereof. In an absorptive medium χ'' is negative so that α is positive whereas in an amplifying medium χ'' is positive and α is negative.

EXERCISE 5.5-1

Dilute Absorbing Medium. A nonabsorptive medium of refractive index n_0 serves as host to a dilute suspension of impurities characterized by susceptibility $\chi = \chi' + j\chi''$, where $\chi' \ll 1$ and

$\chi'' \ll 1$. Determine the overall susceptibility of the medium and demonstrate that the refractive index and absorption coefficient are given approximately by

$$n \approx n_0 + \frac{\chi'}{2n_0} \tag{5.5-9}$$

$$\alpha \approx -\frac{k_o \chi''}{n_0}. \tag{5.5-10}$$

Strongly Absorbing Media

In a strongly absorbing medium, $|\chi''| \gg |1 + \chi'|$, so that (5.5-5) yields $n - j\alpha/2k_o \approx \sqrt{j\chi''} = \sqrt{-j}\sqrt{(-\chi'')} = \pm\frac{1}{\sqrt{2}}(1-j)\sqrt{(-\chi'')}$, whereupon

$$n \approx \sqrt{(-\chi'')/2} \tag{5.5-11}$$

$$\alpha \approx 2k_o\sqrt{(-\chi'')/2}. \tag{5.5-12}$$

Strongly Absorbing Medium

Since χ'' is negative for an absorbing medium, the plus sign of the square root was selected to ensure that α is positive, and this yields a positive value for n as well.

B. Dispersion

Dispersive media are characterized by a frequency-dependent (and thus wavelength-dependent) susceptibility $\chi(\nu)$, electric permittivity $\epsilon(\nu)$, refractive index $n(\nu)$, and speed $c_o/n(\nu)$. Since the angle of refraction in Snell's law depends on refractive index, which is wavelength dependent, optical components fabricated from dispersive materials, such as prisms and lenses, bend light of different wavelengths by different angles. This accounts for the wavelength-resolving capabilities of refracting surfaces and for the wavelength-dependent focusing power of lenses (and the attendant chromatic aberration in imaging systems). Polychromatic light is therefore refracted into a range of directions. These effects are illustrated schematically in Fig. 5.5-2.

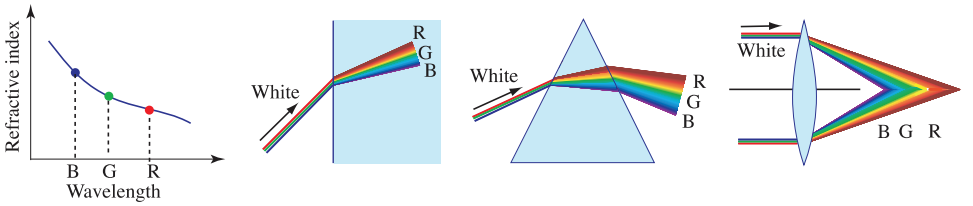


Figure 5.5-2 Optical components fabricated from dispersive materials refract waves of different wavelengths by different angles (B = blue, G = green, R = red).

Moreover, by virtue of the frequency-dependent speed of light in a dispersive medium, each of the frequency components comprising a short pulse of light experiences a different time delay. If the propagation distance through a medium is substantial, as is often the case in an optical fiber, for example, a brief light pulse at the input will be substantially dispersed in time so that its width at the output is increased, as illustrated in Fig. 5.5-3.

The wavelength dependence of the refractive index of some common optical materials is displayed in Fig. 5.5-4.

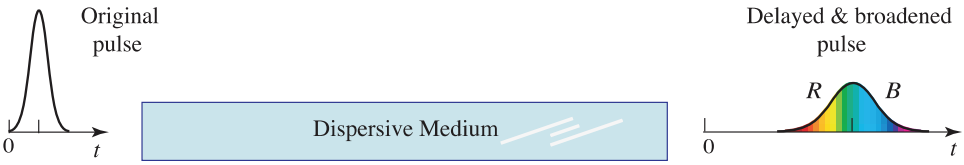


Figure 5.5-3 A dispersive medium serves to broaden a pulse of light because the different frequency components that constitute the pulse travel at different velocities. In this illustration, the low-frequency component (long wavelength, denoted R) travels faster than the high-frequency component (short wavelength, denoted B) and therefore arrives earlier.

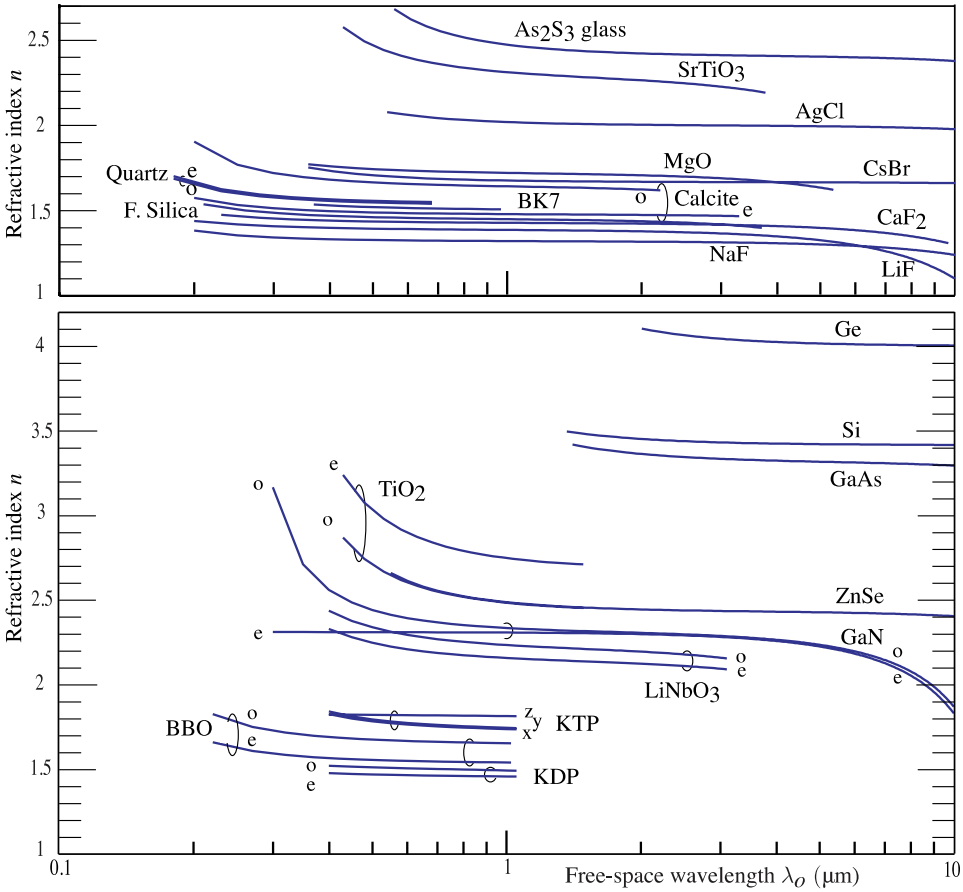


Figure 5.5-4 Wavelength dependence of the refractive index of selected optical materials, including glasses, crystals, and semiconductors. The designations 'e' and 'o' represent ordinary and extraordinary refraction, respectively, for anisotropic materials (see Sec. 6.3).

Measures of Dispersion

Material dispersion can be quantified in a number of different ways. For glass optical components and broad-spectrum light that covers the visible band (white light), a commonly used measure is the Abbe number $\mathbb{V} = (n_d - 1)/(n_F - n_C)$, where n_F, n_d, and n_C are the refractive indices of the glass at three standard wavelengths: blue at 486.1 nm, yellow at 587.6 nm, and red at 656.3 nm, respectively. For flint glass $\mathbb{V} \approx 38$

whereas for fused silica $\mathbb{V} \approx 68$.

On the other hand, if dispersion in the vicinity of a particular wavelength λ_o is of interest, an often used measure is the magnitude of the derivative $dn/d\lambda_o$ at that wavelength. This measure is appropriate for prisms, for example, in which the ray deflection angle θ_d is a function of n [see (1.2-6)]. The angular dispersion $d\theta_d/d\lambda_o = (d\theta_d/dn)(dn/d\lambda_o)$ is then a product of the material dispersion factor, $dn/d\lambda_o$, and another factor, $d\theta_d/dn$, that depends on the geometry of the prism and the refractive index of the material of which it is made.

The effect of material dispersion on the propagation of brief pulses of light is governed not only by the refractive index n and its first derivative $dn/d\lambda_o$, but also by the second derivative $d^2n/d\lambda_o^2$, as will be elucidated in Sec. 5.7 and Sec. 23.3.

Absorption and Dispersion: The Kramers–Kronig Relations

Absorption and dispersion are intimately related. Indeed, a dispersive material, i.e., a material whose refractive index is wavelength dependent, *must* be absorptive and must exhibit an absorption coefficient that is also wavelength dependent. The relation between the absorption coefficient and the refractive index is a result of the Kramers–Kronig relations, which relate the real and imaginary parts of the susceptibility of a medium, $\chi'(\nu)$ and $\chi''(\nu)$:

$$\chi'(\nu) = \frac{2}{\pi} \int_0^\infty \frac{s\chi''(s)}{s^2 - \nu^2}\, ds \tag{5.5-13}$$

$$\chi''(\nu) = \frac{2}{\pi} \int_0^\infty \frac{\nu\chi'(s)}{\nu^2 - s^2}\, ds. \tag{5.5-14}$$

Kramers–Kronig Relations

Given the real or the imaginary component of $\chi(\nu)$ for all ν, these powerful formulas allow the complementary component to be determined for all ν. The Kramers–Kronig relations connecting $\chi''(\nu)$ and $\chi'(\nu)$ translate into relations between the absorption coefficient $\alpha(\nu)$ and the refractive index $n(\nu)$ by virtue of (5.5-5), which relates α and n to χ'' and χ'.

The Kramers–Kronig relations are a special Hilbert-transform pair, as can be understood from linear systems theory (see Sec. B.1 of Appendix B). They are applicable for all linear, shift-invariant, causal systems with real impulse response functions. The linear system at hand is the polarization-density response of a medium $\mathcal{P}(t)$ to an applied electric field $\mathcal{E}(t)$ set forth in (5.2-23). Since $\mathcal{E}(t)$ and $\mathcal{P}(t)$ are real, so too is the impulse response function $\epsilon_o\chi(t)$. As a consequence, its Fourier transform, the transfer function $\epsilon_o\chi(\nu)$, exhibits Hermitian symmetry: $\chi(-\nu) = \chi^*(\nu)$ [see Sec. A.1 of Appendix A]. This system therefore obeys all of the conditions required for the Kramers–Kronig relations to apply. The real and imaginary parts of the transfer function $\epsilon_o\chi(\nu)$ are therefore related by (B.1-6) and (B.1-7) and, in particular, by (5.5-13) and (5.5-14).

C. The Resonant Medium

We now set forth a simple classical microscopic theory that leads to a complex susceptibility and provides an underlying rationale for the presence of frequency-dependent absorption and dispersion in an optical medium. The approach is known as the **Lorentz oscillator model**. A more thorough discussion of the interaction of light and matter is provided in Chapter 14.

Consider a dielectric medium such as a collection of resonant atoms, in which the dynamic relation between the polarization density $\mathcal{P}(t)$ and the electric field $\mathcal{E}(t)$, considered for a single polarization, is described by a linear second-order ordinary differential equation of the form

$$\frac{d^2\mathcal{P}}{dt^2} + \zeta \frac{d\mathcal{P}}{dt} + \omega_0^2 \mathcal{P} = \omega_0^2 \epsilon_o \chi_0 \mathcal{E}, \tag{5.5-15}$$

Resonant Dielectric Medium

where ζ, ω_0, and χ_0 are constants.

An equation of this form emerges when the motion of a bound charge associated with a resonant atom is modeled phenomenologically as a classical harmonic oscillator, in which the displacement of the charge $x(t)$ and the applied force $\mathcal{F}(t)$ are related by

$$\frac{d^2x}{dt^2} + \zeta \frac{dx}{dt} + \omega_0^2 x = \frac{\mathcal{F}}{m}. \tag{5.5-16}$$

Here m is the mass of the bound charge, $\omega_0 = \sqrt{\kappa/m}$ is its resonance angular frequency, κ is the elastic constant of the restoring force, and ζ is the damping coefficient.

If the dipole moment associated with each individual atom is $\mathrm{p} = -ex$, the polarization density of the medium as a whole is related to the displacement by $\mathcal{P} = N\mathrm{p} = -Nex$, where $-e$ is the electronic charge and N is the number of atoms per unit volume of the medium. The electric field and force are related by $\mathcal{E} = \mathcal{F}/(-e)$. The quantities $\mathcal{P}$ and $\mathcal{E}$ are therefore proportional to x and $\mathcal{F}$, respectively, and comparison of (5.5-15) and (5.5-16) provides

$$\chi_0 = \frac{Ne^2}{\epsilon_o m \omega_0^2}. \tag{5.5-17}$$

The applied electric field can thus be thought of as inducing a time-dependent electric dipole moment in each atom, as portrayed in Fig. 5.5-5, and hence a time-dependent polarization density in the medium as a whole.

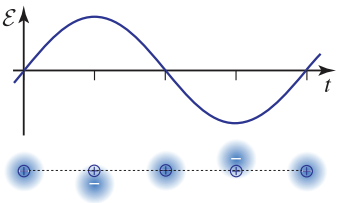


Figure 5.5-5 A time-varying electric field $\mathcal{E}$ applied to a Lorentz-oscillator atom induces a time-varying dipole moment p that contributes to the overall polarization density $\mathcal{P}$.

The medium is completely characterized by its impulse response function $\epsilon_o x(t)$, an exponentially decaying harmonic function, or equivalently by its transfer function $\epsilon_o \chi(\nu)$, which is obtained by solving (5.5-15) one frequency at a time, as follows. Substituting $\mathcal{E}(t) = \mathrm{Re}\{E \exp(j\omega t)\}$ and $\mathcal{P}(t) = \mathrm{Re}\{P \exp(j\omega t)\}$ into (5.5-15) yields

$$(-\omega^2 + j\zeta\omega + \omega_0^2) P = \omega_0^2 \epsilon_o \chi_0 E, \tag{5.5-18}$$

from which $P = \epsilon_o [\chi_0 \omega_0^2 / (\omega_0^2 - \omega^2 + j\zeta\omega)] E$. Writing this relation in the form $P = \epsilon_o \chi(\nu) E$, and substituting $\omega = 2\pi\nu$, yields an expression for the frequency-dependent

susceptibility,

$$\chi(\nu) = \chi_0 \frac{\nu_0^2}{\nu_0^2 - \nu^2 + j\nu\,\Delta\nu}\,, \tag{5.5-19}$$

Susceptibility (Resonant Medium)

where $\nu_0 = \omega_0/2\pi$ is the resonance frequency and $\Delta\nu = \zeta/2\pi$.

The real and imaginary parts of $\chi(\nu)$, denoted $\chi'(\nu)$ and $\chi''(\nu)$ respectively, are therefore given by

$$\chi'(\nu) = \chi_0 \frac{\nu_0^2\left(\nu_0^2 - \nu^2\right)}{\left(\nu_0^2 - \nu^2\right)^2 + (\nu\,\Delta\nu)^2} \tag{5.5-20}$$

$$\chi''(\nu) = -\chi_0 \frac{\nu_0^2\,\nu\,\Delta\nu}{\left(\nu_0^2 - \nu^2\right)^2 + (\nu\,\Delta\nu)^2}\,. \tag{5.5-21}$$

These equations are plotted in Fig. 5.5-6.

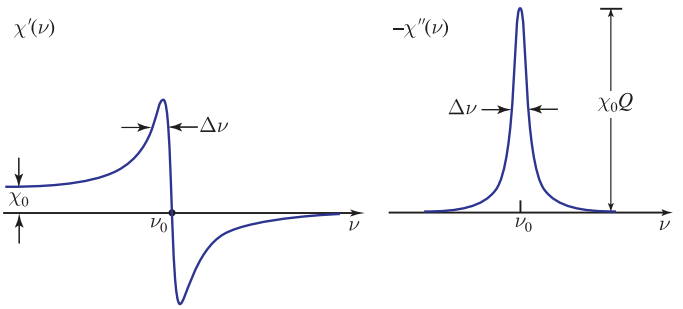


Figure 5.5-6 Real and imaginary parts of the susceptibility of a resonant dielectric medium. The real part $\chi'(\nu)$ is positive below resonance, zero at resonance, and negative above resonance. The imaginary part $\chi''(\nu)$ is negative so that $-\chi''(\nu)$ is positive everywhere and has a peak value $\chi_0 Q$ at $\nu = \nu_0$, where $Q = \nu_0/\Delta\nu$. The illustration portrays results for $Q = 10$.

At frequencies well below resonance ($\nu \ll \nu_0$), $\chi'(\nu) \approx \chi_0$ and $\chi''(\nu) \approx 0$, so that the low-frequency susceptibility is simply χ_0. At frequencies well above resonance ($\nu \gg \nu_0$), $\chi'(\nu) \approx \chi''(\nu) \approx 0$ so that the medium behaves like free space. Precisely at resonance ($\nu = \nu_0$), $\chi'(\nu_0) = 0$ and $-\chi''(\nu_0)$ reaches its peak value of $\chi_0 Q$, where $Q = \nu_0/\Delta\nu$. The resonance frequency ν_0 is usually much greater than $\Delta\nu$ so that $Q \gg 1$. Thus, the magnitude of the peak value of $-\chi''(\nu)$, which is $\chi_0 Q$, is much larger than the magnitude of the low-frequency value of $\chi'(\nu)$, which is χ_0. The maximum and minimum values of $\chi'(\nu)$ are $\pm\chi_0\,Q/(2 \mp 1/Q)$ and occur at frequencies $\nu_0\sqrt{1 \mp 1/Q}$, respectively. For large Q, χ' swings between positive and negative values with a magnitude approximately equal to $\chi_0 Q/2$, i.e., one half of the peak value of χ''. The signs of χ' and χ'' determine the phase of χ, which simply determines the angle between the phasors P and E.

The behavior of $\chi(\nu)$ in the vicinity of resonance ($\nu \sim \nu_0$) is often of particular interest. In this region, we may use the approximation $(\nu_0^2 - \nu^2) = (\nu_0 + \nu)(\nu_0 - \nu) \approx$

$2\nu_0(\nu_0 - \nu)$ in the real part of the denominator of (5.5-19), and replace ν with ν_0 in the imaginary part thereof, to obtain

$$\chi(\nu \sim \nu_0) \approx \chi_0 \frac{\nu_0/2}{(\nu_0 - \nu) + j\Delta\nu/2}, \tag{5.5-22}$$

from which

$$\chi''(\nu) \approx -\chi_0 \frac{\nu_0 \Delta\nu}{4} \frac{1}{(\nu_0 - \nu)^2 + (\Delta\nu/2)^2} \tag{5.5-23}$$

$$\chi'(\nu) \approx 2 \frac{\nu - \nu_0}{\Delta\nu} \chi''(\nu). \tag{5.5-24}$$

Susceptibility (Near Resonance)

The function $\chi''(\nu)$ in (5.5-23), known as the **Lorentzian function**, decreases to half its peak value when $|\nu - \nu_0| = \Delta\nu/2$. The parameter $\Delta\nu$ therefore represents the full-width at half-maximum (FWHM) value of $\chi''(\nu)$.

The behavior of $\chi(\nu)$ far from resonance is also of interest. In the limit $|(\nu - \nu_0)| \gg \triangle\nu$, the susceptibility given in (5.5-19) is approximately real,

$$\chi(\nu) \approx \chi_0 \frac{\nu_0^2}{\nu_0^2 - \nu^2}, \tag{5.5-25}$$

Susceptibility (Far from Resonance)

so that the medium exhibits negligible absorption.

The absorption coefficient and the refractive index of a resonant medium may be determined by substituting the expressions for $\chi'(\nu)$ and $\chi''(\nu)$, e.g., (5.5-23) and (5.5-24) into (5.5-5). Each of these parameters generally depends on both $\chi'(\nu)$ and $\chi''(\nu)$. However, in the special case for which the resonant atoms are embedded in a nondispersive host medium of refractive index n_0, and are sufficiently dilute so that $\chi''(\nu)$ and $\chi'(\nu)$ are both $\ll 1$, this dependence is much simpler, namely, the refractive index and the absorption coefficient are dependent on χ' and χ'', respectively. Using the results of Exercise 5.5-1, it can be shown that these parameters are related by:

$$\alpha(\nu) \approx -\left(\frac{2\pi\nu}{n_0 c_o}\right) \chi''(\nu) \tag{5.5-26}$$

$$n(\nu) \approx n_0 + \frac{\chi'(\nu)}{2n_0}. \tag{5.5-27}$$

The dependence of these quantities on ν is illustrated in Fig. 5.5-7.

Media with Multiple Resonances

A typical dielectric medium contains multiple resonances corresponding to different lattice and electronic vibrations. The overall susceptibility arises from a superposition of contributions from these resonances. Whereas the imaginary part of the susceptibility is confined to frequencies near the resonance, the real part contributes at all frequencies near and *below* resonance, as shown in Fig. 5.5-6. This is exhibited in the frequency dependence of the absorption coefficient and the refractive index, as illustrated in Fig. 5.5-8. Absorption and dispersion are strongest near the resonance

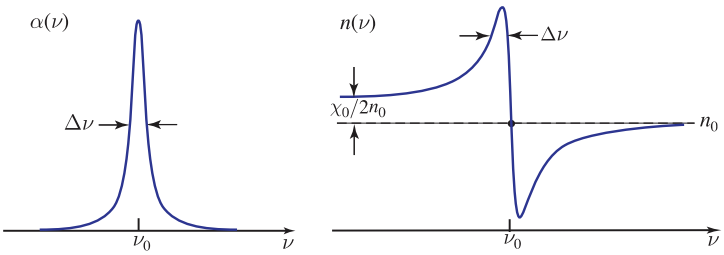


Figure 5.5-7 Absorption coefficient $\alpha(\nu)$ and refractive index $n(\nu)$ of a dielectric medium of refractive index n_0 containing a dilute concentration of atoms of resonance frequency ν_0.

frequencies. Away from the resonance frequencies, the refractive index is constant and the medium is approximately nondispersive and nonabsorptive. Each resonance does, however, contribute a constant value to the refractive index at all frequencies below its resonance frequency.

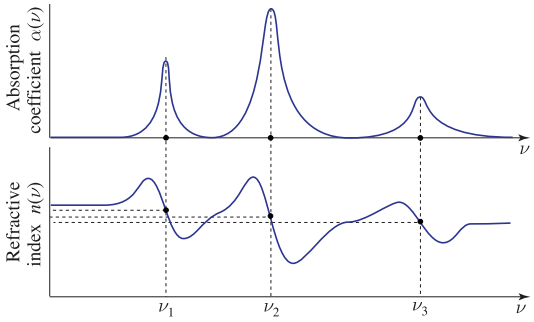


Figure 5.5-8 Frequency dependence of the absorption coefficient $\alpha(\nu)$ and the refractive index $n(\nu)$ for a medium with three resonances.

Other complex processes can also contribute to the absorption coefficient and the refractive index of a material, so that different patterns of frequency dependence emerge. Figure 5.5-9 shows an example of the *wavelength* dependence of the absorption coefficient and refractive index for a dielectric material that is essentially transparent at visible wavelengths. The illustration shows a decreasing refractive index with increasing wavelength in the visible region by virtue of a nearby ultraviolet resonance. The material is therefore more dispersive at shorter visible wavelengths where the rate of decrease of the index is greatest. This behavior is not unlike that exhibited in Fig. 5.5-1 and Fig. 5.5-4 for various real dielectric materials.

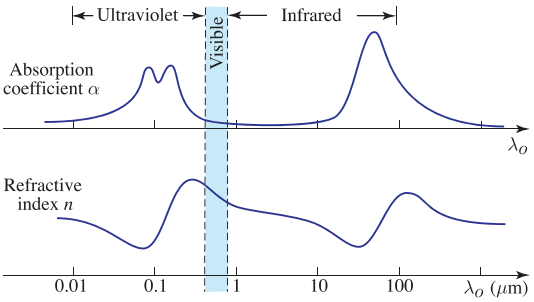


Figure 5.5-9 Typical wavelength dependence of the absorption coefficient and refractive index for a dielectric medium exhibiting resonant absorption in the ultraviolet and infrared bands, concomitant with low absorption in the visible band. In this diagram the abscissa is wavelength rather than frequency.

The Sellmeier Equation

In a medium with multiple resonances, labeled $i = 1, 2, \ldots,$ the susceptibility is approximately given by a sum of terms, each of the form of (5.5-25), for frequencies far from any of the resonances. Using the relation between the refractive index and the real susceptibility provided in (5.2-13), $n^2 = 1 + \chi$, the dependence of n on frequency and wavelength assumes a form known as the **Sellmeier equation**:

$$n^2 \approx 1 + \sum_i \chi_{0i} \frac{\nu_i^2}{\nu_i^2 - \nu^2} = 1 + \sum_i \chi_{0i} \frac{\lambda^2}{\lambda^2 - \lambda_i^2}. \tag{5.5-28}$$

Sellmeier Equation

The Sellmeier equation provides a good description of the refractive index for most optically transparent materials. At wavelengths for which $\lambda \ll \lambda_i$ the ith term becomes approximately proportional to λ^2, and for $\lambda \gg \lambda_i$ it becomes approximately constant. As an example, the dispersion in fused silica, illustrated in Example 5.7-1, is well described by three resonances. For some materials the Sellmeier equation is conveniently approximated by a power series.

The Sellmeier equations for a few selected materials, extracted from measured data using a least-squares fitting algorithm, are provided in Table 5.5-1.

Table 5.5-1 Sellmeier equations for the wavelength dependence of the refractive indices for selected materials at room temperature. The quantities n_o and n_e indicate the ordinary and extraordinary indices of refraction, respectively, for anisotropic materials (see Sec. 6.3). The range of wavelengths where the results are valid is indicated in the rightmost column.

Material	Sellmeier Equation (Wavelength λ in μm)	Wavelength Range (μm)
Fused silica	$n^2 = 1 + \frac{0.6962\lambda^2}{\lambda^2 - (0.06840)^2} + \frac{0.4079\lambda^2}{\lambda^2 - (0.1162)^2} + \frac{0.8975\lambda^2}{\lambda^2 - (9.8962)^2}$	0.21–3.71
Si	$n^2 = 1 + \frac{10.6684\lambda^2}{\lambda^2 - (0.3015)^2} + \frac{0.0030\lambda^2}{\lambda^2 - (1.1347)^2} + \frac{1.5413\lambda^2}{\lambda^2 - (1104.0)^2}$	1.36–11
GaAs	$n^2 = 3.5 + \frac{7.4969\lambda^2}{\lambda^2 - (0.4082)^2} + \frac{1.9347\lambda^2}{\lambda^2 - (37.17)^2}$	1.4–11
BBO	$n_o^2 = 2.7359 + \frac{0.01878}{\lambda^2 - 0.01822} - 0.01354\lambda^2$	0.22–1.06
	$n_e^2 = 2.3753 + \frac{0.01224}{\lambda^2 - 0.01667} - 0.01516\lambda^2$	
KDP	$n_o^2 = 1 + \frac{1.2566\lambda^2}{\lambda^2 - (0.09191)^2} + \frac{33.8991\lambda^2}{\lambda^2 - (33.3752)^2}$	0.4–1.06
	$n_e^2 = 1 + \frac{1.1311\lambda^2}{\lambda^2 - (0.09026)^2} + \frac{5.7568\lambda^2}{\lambda^2 - (28.4913)^2}$	
$LiNbO_3$	$n_o^2 = 2.3920 + \frac{2.5112\lambda^2}{\lambda^2 - (0.217)^2} + \frac{7.1333\lambda^2}{\lambda^2 - (16.502)^2}$	0.4–3.1
	$n_e^2 = 2.3247 + \frac{2.2565\lambda^2}{\lambda^2 - (0.210)^2} + \frac{14.503\lambda^2}{\lambda^2 - (25.915)^2}$	

5.6 SCATTERING OF ELECTROMAGNETIC WAVES

Previous chapters have described the propagation of optical waves through homogeneous media, the reflection and refraction of light at dielectric boundaries, wave transmission through optical components, and diffraction through apertures. In Sec. 5.5, we considered the absorption and dispersion of light. We turn now to the scattering of light, which plays an important role in various domains of optics, including nanophotonics.

In particular, we examine light scattering from a homogeneous medium containing localized inhomogeneities, irregularities, material defects, grains, or suspended particles. Both the medium and the scatterers are assumed to be dielectrics with linear and isotropic optical properties. The scattering from a small metal sphere is considered in Sec. 8.2C and various forms of light scattering are discussed in Sec. 14.5C.

A. Born Approximation

When an optical wave traveling in a given direction in a homogeneous medium encounters an object with different optical properties, the wave is scattered into other directions. This effect may be analyzed by solving Maxwell's equations and applying the appropriate boundary conditions. However, analytical solutions of this problem exist only in few ideal cases. We therefore resort to a commonly used approximate approach for solving such problems, known as the **Born approximation**. It is applicable for *weak scattering*, i.e., when the scattering object may be regarded as a small perturbation to the relative permittivity (or other optical properties) of the medium.

To introduce the Born approximation, it is convenient to first address the scattering of a scalar wave and then to subsequently consider an electromagnetic wave. The scalar complex amplitude $U(\mathbf{r})$ obeys the Helmholtz equation (2.2-7),

$$\left[\nabla^2 + k^2(\mathbf{r})\right] U = 0, \tag{5.6-1}$$

where inside the scattering object the wavenumber is $k(\mathbf{r}) = k_{\mathrm{s}}(\mathbf{r})$ and in the host medium, which is taken to be uniform, the wavenumber is $k(\mathbf{r}) = k$. By writing $k^2(\mathbf{r}) = k^2 + [k^2(\mathbf{r}) - k^2]$, (5.6-1) may be rewritten as the Helmholtz equation for the scattered complex amplitude $U_{\mathrm{s}}(\mathbf{r})$,

$$\left(\nabla^2 + k^2\right) U_{\mathrm{s}} = -S, \tag{5.6-2}$$

with a source

$$S(\mathbf{r}) = [k_{\mathrm{s}}^2(\mathbf{r}) - k^2]\, U_{\mathrm{s}}(\mathbf{r}) \tag{5.6-3}$$

that is localized within the volume V of the scatterer, and is zero outside of it. As will be justified shortly, the solution to (5.6-2) is

$$U_{\mathrm{s}}(\mathbf{r}) = \int_V S(\mathbf{r}') \frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi\,|\mathbf{r}-\mathbf{r}'|}\, \mathrm{d}\mathbf{r}' \tag{5.6-4}$$

at positions $\mathbf{r}$ outside the volume V. However, the integral in (5.6-4) cannot be readily evaluated to determine $U_{\mathrm{s}}(\mathbf{r})$ since, in accordance with (5.6-3), the source $S(\mathbf{r}')$ itself depends on the wave $U_{\mathrm{s}}(\mathbf{r})$, which is unknown.

If the scattering is weak, however, it is safe to assume that the incident wave $U_0(\mathbf{r})$ is essentially unaffected by the process of scattering within the volume V, in which

case the complex amplitude $U_s(\mathbf{r})$ in the expression for the scattering source (5.6-3) may be approximated by the incident complex amplitude $U_0(\mathbf{r})$, whereupon

$$S(\mathbf{r}) \approx [k_s^2(\mathbf{r}) - k^2]\, U_0(\mathbf{r}). \tag{5.6-5}$$

This expression may then be used in (5.6-4) to determine the scattered complex amplitude $U_s(\mathbf{r})$. Implicit in the assumption of weak scattering is the condition that a wave scattered from one point in the scattering volume V is not subsequently scattered from another point, i.e., multiple scattering is a negligible second-order effect.

It is evident from (5.6-4) that the scattered wave $U_s(\mathbf{r})$ is then approximately a superposition of spherical waves generated by a continuum of point sources within the scatterer, as schematized in Fig. 5.6-1. Each point at position $\mathbf{r}'$ creates a spherical wave with amplitude $S(\mathbf{r}')$ given by the approximate expression (5.6-5). The concept is similar to that of the Huygens–Fresnel principle of diffraction described in Sec. 4.1D (see Fig. 4.1-13). In this type of scattering, known as **elastic scattering**, the frequency of the scattered light remains the same as that of the incident light.

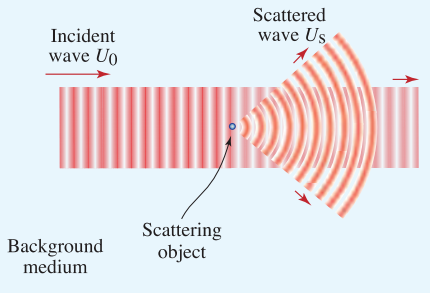


Figure 5.6-1 Under the Born approximation, the scattered wave $U_s(\mathbf{r})$ is a superposition of spherical waves, each generated by a point in the scatterer.

B. Rayleigh Scattering

Rayleigh scattering involves *small scatterers*. It is engendered by variations in a medium that are introduced, for example, by the presence of particles whose sizes are much smaller than a wavelength or by random inhomogeneities at a scale much finer than a wavelength.

Weak Scattering: Scalar Waves

If the contrast between the optical properties of the scattering and surrounding media is low, i.e., if the scattering is weak, then the Born approximation is applicable.

If we consider a single scattering object, much smaller than the wavelength of light and located at $\mathbf{r} = \mathbf{0}$, the source distribution in (5.6-5) may be approximated as $S(\mathbf{r}) \approx (k_s^2 - k^2)\, U_0\, V\, \delta(\mathbf{r})$, where $\delta(\mathbf{r})$ is the delta function and k_s is the wavenumber within the small scatterer. Substituting this in the integral provided in (5.6-4) yields

$$U_s \approx (k_s^2 - k^2)\, V\, U_0\, \frac{e^{-jkr}}{4\pi r}, \tag{5.6-6}$$

which represents a single spherical wave centered about $\mathbf{r} = \mathbf{0}$ (the location of the scatterer), with an amplitude proportional to that of the incident wave U_0.

In accordance with (2.2-10), the intensity of the scattered wave is therefore

$$I_s = |U_s|^2 \approx \left(k_s^2 - k^2\right)^2 \frac{V^2}{(4\pi r)^2}\, I_0, \tag{5.6-7}$$

where $I_0 = |U_0|^2$. Since the scalar scattered wave is isotropic, the total scattered power $P_s = 4\pi r^2 I_s$ becomes

$$P_s \approx \frac{1}{4\pi} \left(k_s^2 - k^2\right)^2 V^2 I_0, \tag{5.6-8}$$

which reveals that the scattered power is proportional to the square of the scatterer volume V.

Since k_s and k are both proportional to ω, it is clear from (5.6-8) that the scattered power is proportional to ω^4 or, in terms of wavelength, to $1/\lambda_o^4$. Known as the **Rayleigh inverse fourth-power law**, this indicates that incident waves of short wavelength undergo greater scattering than those of long wavelength. As an example, the Rayleigh scattering of light at a wavelength of $\lambda_o = 400$ nm exceeds that of light at a wavelength of $\lambda_o = 800$ nm by the factor $2^4 = 16$. Rayleigh scattering from the density fluctuations of air, which are finer than the wavelengths of light in the visible spectral band, is responsible for the blue color of the sky. The short-wavelength (blue) light is preferentially scattered over a large range of angles, whereas the light arriving directly from the sun is reduced in blue and therefore appears to have a yellowish tint. In silica-glass optical fibers, Rayleigh scattering is responsible for the greater attenuation of visible than infrared light, as discussed in Sec. 10.3A.

Weak Scattering: Electromagnetic Waves

The derivation of the scattered wave considered above was predicated on a scalar complex amplitude that obeys the Helmholtz equation (5.6-1). The scattering of an electromagnetic wave may be formulated in a similar manner by beginning with the vector potential $\mathbf{A}$, which also satisfies the Helmholtz equation. Applying the Born approximation, the vector potential of the scattered wave may be expressed as a superposition of dipole waves centered at points within the scatterer, in analogy with (5.6-4). The vector potential $\mathbf{A}$ for an oscillating dipole has the distribution of a spherical wave, with the associated electric and magnetic complex amplitudes $\mathbf{E}$ and $\mathbf{H}$ described in Sec. 5.4A.

From an electromagnetic point-of-view, scattering can thus be viewed as the creation, by the incident field, of a collection of oscillating electric dipoles at all points within the scatterer, each radiating a dipole wave.

For a single small scatterer at the origin, the scattered electromagnetic wave is identical to that radiated by a single electric dipole pointing along the direction of the electric field $\mathbf{E}_0$ of the incident wave, as illustrated in Fig. 5.6-2. In the far zone ($r \gg \lambda$), the electric and magnetic fields of the scattered wave point in the polar and azimuthal directions, respectively, as provided in (5.4-17) and (5.4-18), as well as in Fig. 5.4-2. The electric-field complex amplitude of the scattered wave is thus given by

$$\mathbf{E}_s \approx E_{s0} \sin\theta \, \frac{e^{-jkr}}{4\pi r} \, \widehat{\boldsymbol{\theta}}, \qquad E_{s0} = -(k_s^2 - k^2) V E_0, \tag{5.6-9}$$

so that the scattered-wave intensity is

$$I_s \approx \left|k_s^2 - k^2\right|^2 \frac{V^2}{(4\pi r)^2} \, I_0 \, \sin^2\theta, \tag{5.6-10}$$

where I_0 is given by (5.4-8). The angular distribution of the scattered wave is thus independent of ϕ and assumes the toroidal pattern illustrated in Fig. 5.6-2. The scattering is at a maximum when $\theta = \pi/2$, i.e., when the direction of the scattered wave is orthogonal to the direction of the electric field of the incident wave. In particular, back-scattering has the same intensity as forward-scattering.

The expression for the electromagnetic intensity given in (5.6-10) differs from that for the scalar-wave intensity provided in (5.6-7) by the factor $\sin^2\theta$. This distinction arises because the oscillating dipole radiates a transverse electromagnetic wave, which precludes scattering in a direction parallel to the incident electric field, whereas scalar wave optics does not take polarization into account.

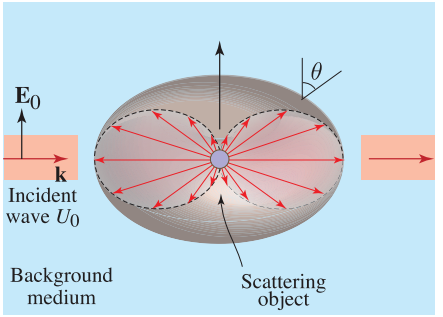


Figure 5.6-2 A transverse electromagnetic plane wave with electric field $\mathbf{E}_0$ scattered from a point object (blue circle at center) creates a scattered electric-dipole wave $\mathbf{E}_s$ with a toroidal directional pattern. The scattered intensity $I_s \propto \sin^2\theta$, where θ is the scattering angle.

The total scattered power is calculated by integrating (5.6-10) over the surface of a sphere. Using the incremental integration area in spherical coordinates, $r^2 \sin\theta\, d\theta\, d\phi$, and noting that $\int_0^\pi \sin^3\theta\, d\theta = 4/3$, leads to

$$P_s \approx \frac{1}{6\pi} \left|k_s^2 - k^2\right|^2 V^2 I_0. \tag{5.6-11}$$

The electromagnetic scattered power is thus $2/3$ of that obtained in the scalar-wave case, as provided in (5.6-8). The results differ because of the distinction in the integration over θ in the two cases. In the isotropic case the appropriate integration is $\int_0^\pi \sin\theta\, d\theta = 2$, whereas in the electromagnetic case the integration yields $4/3$, as indicated above, which is a factor of $2/3$ smaller.

It is commonplace to characterize the strength of scattering in terms of a **scattering cross section** σ_s. Writing the total scattered power P_s as the product

$$P_s = \sigma_s I_0, \tag{5.6-12}$$

where I_0 is the incident light intensity [W/m^2], it is evident that σ_s may be regarded as the area of an aperture [m^2] that intercepts the incident wave and collects an amount of power equal to the actual scattered power. Based on (5.6-11), the scattering cross section under the Born approximation (weak scattering) and the small-scatterer approximation (Rayleigh scattering) is therefore

$$\sigma_s = \frac{1}{6\pi} \left|k_s^2 - k^2\right|^2 V^2. \tag{5.6-13}$$

Let us consider a specific example: the scattering cross section of a spherical dielectric scatterer of radius a and permittivity ϵ_s embedded in a dielectric medium of permittivity ϵ, under the assumption that both media have the same magnetic permeability μ. Substituting $k = \omega\sqrt{\epsilon\mu} = 2\pi/\lambda$, $k_s = \omega\sqrt{\epsilon_s\mu} = \sqrt{\epsilon_s/\epsilon} \cdot 2\pi/\lambda$, and $V = \frac{4}{3}\pi a^3$ into (5.6-13), we obtain

$$\sigma_s = \pi a^2 Q_s, \qquad Q_s = \frac{8}{3} \left|\frac{\epsilon_s - \epsilon}{3\epsilon}\right|^2 \left(2\pi \frac{a}{\lambda}\right)^4. \tag{5.6-14}$$

The scattering cross section of the spherical scatterer is thus given by the product of its geometrical area, πa^2, and a small dimensionless factor Q_s, known as the **scattering efficiency**. The quantity Q_s is proportional to the fourth power of the ratio a/λ, where λ is the wavelength of light in the background medium, and to the square of the contrast factor $(\epsilon_s - \epsilon)/\epsilon = (n_s^2 - n^2)/n^2$, where n_s and n are the refractive indices of the scatterer and the medium, respectively. Rayleigh scattering is evidently highly dependent on the size of the scatterer; the scattered power is proportional to the sixth power of the radius of a spherical scatterer. Of course the validity of these results requires that the radius of the scatterer be small in comparison with a wavelength.

EXAMPLE 5.6-1. ***Rayleigh Scattering from a Dielectric Nanosphere.*** Light of wavelength $\lambda = 600$ nm is scattered from a spherical nanoparticle of radius $a = 60$ nm and a relative permittivity that is 10% greater than the background value. Since the $a/\lambda = 0.1$, the small-scatterer condition is satisfied. Also, since the contrast $(\epsilon_s - \epsilon)/\epsilon = 0.1$, the weak-scattering condition is satisfied. In accordance with (5.6-14), the scattering efficiency is $Q_s \approx 4.6 \times 10^{-4}$ and the scattering cross section is $\sigma_s \approx 5.2$ nm^2. Hence, if the intensity of the incident light is $I_0 \approx 10^5$ W/m^2 (corresponding to a 3-mW laser beam of 100-μm radius), the scattered power is $P_s \approx 0.52$ pW.

Strong Scattering: Nanosphere

The Born approximation is not applicable in the case of strong scattering, i.e., when the contrast $(\epsilon_s - \epsilon)/\epsilon$ between the relative permittivities of the scatterer and the background is not small. However, an alternative method, known as the **quasi-static approximation**, may be used to determine the Rayleigh scattered field if the scatterer is spherical and its a radius much smaller than the optical wavelength, i.e., a nanosphere.

Again, the scattered electric-field complex amplitude $\mathbf{E}_s$ is that radiated by an electric dipole, as in (5.4-15) and (5.4-16). As explained below, in the far zone the field is described by

$$\mathbf{E}_s \approx E_{s0} \sin\theta \, \frac{e^{-jkr}}{4\pi r} \, \widehat{\boldsymbol{\theta}}, \qquad E_{s0} = -4\pi \left(\frac{\epsilon_s - \epsilon}{\epsilon_s + 2\epsilon} \right) k^2 a^3 \, E_0, \tag{5.6-15}$$

and the associated scattering cross section turns out to be approximately given by

$$\boxed{\sigma_s = \pi a^2 \, Q_s, \qquad Q_s = \frac{8}{3} \left| \frac{\epsilon_s - \epsilon}{\epsilon_s + 2\epsilon} \right|^2 \left(2\pi \frac{a}{\lambda} \right)^4.} \tag{5.6-16}$$

Nanosphere Cross Section

If $\epsilon_s \approx \epsilon$, then $\epsilon_s + 2\epsilon \approx 3\epsilon$, whereupon the weak-scattering results are recovered from the above equations, i.e., (5.6-15) reproduces (5.6-9), and (5.6-16) reproduces (5.6-14).

These results may be confirmed by applying appropriate boundary conditions at the surface of the scattering sphere ($r = a$), namely, matching the tangential components of the external and internal electric fields $\mathbf{E}$, as well as the normal components of the displacement fields $\mathbf{D}$, which are products of the permittivities and the electric fields in each medium (see Fig. 5.1-1). The *internal electric field* $\mathbf{E}_i$ within the scattering sphere is uniformly distributed, with amplitude

$$\boxed{E_i = \frac{3\epsilon}{\epsilon_s + 2\epsilon} E_0,} \tag{5.6-17}$$

Nanosphere Internal Field

and with a direction that is parallel to the electric field of the incident wave, as shown in Fig. 5.6-3.

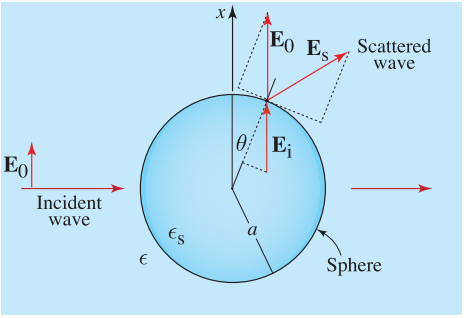


Figure 5.6-3 Scattering of a plane wave with electric field $\mathbf{E}_0$ from a dielectric nanosphere of radius $a \ll \lambda$. The scattered wave $\mathbf{E}_s$ is identical to the wave radiated by an electric dipole, and the internal field $\mathbf{E}_i$ is uniform within the sphere. Boundary conditions dictate that the polar components of $\mathbf{E}_0 + \mathbf{E}_s$ and $\mathbf{E}_i$ are equal, and the radial components of $\epsilon(\mathbf{E}_0 + \mathbf{E}_s)$ and $\epsilon_s \mathbf{E}_i$ are also equal. Scattering from a metal nanosphere is considered in Sec. 8.2C.

The external field is the sum of the incident field $\mathbf{E}_0$ and the scattered field $\mathbf{E}_s$, which is a dipole wave. Since the radius of the sphere is taken to be much smaller than the wavelength of the light ($r \ll \lambda$), at the boundary $r = a$ we have $kr \ll 1$. It follows that points on the sphere lie in the near-field zone of the dipole wave. As a consequence, the full expression for the electric field of the dipole wave provided in (5.4-15) may be approximated by the $1/(jkr)^2$ terms. At $r = a$, the radial and polar components of $\mathbf{E}_s$ are therefore $2(jka)^{-2}(E_{s0}\cos\theta)(4\pi a)^{-1}e^{-jka}$ and $(jka)^{-2}(E_{s0}\sin\theta)(4\pi a)^{-1}e^{-jka}$, respectively. Inserting these expressions in the boundary conditions results in (5.6-16) and (5.6-17). This solution, which is valid for long wavelengths ($\lambda \gg a$), i.e., low frequencies, may also be obtained by solving the electrostatic problem of a dielectric sphere in an applied steady electric field, which explains the appellation *quasi-static approximation*.

C. Mie Scattering

For weak scattering, the Born approximation is applicable for scatterers of all sizes, including those with dimensions comparable to, or larger than, the wavelength of the incident light. The scattered wave is formulated as an integral of dipole waves centered at points within the scatterer, and with amplitudes proportional to the local value of $k_s^2(\mathbf{r}) - k^2$, as set forth in Sec. 5.6A. The resultant scattering pattern is sensitive to the size and shape of the scatterer.

If the Born approximation cannot be used because the scattering is not sufficiently weak, the problem can be solved analytically for a few special shapes, such as spheres. This is known as **Mie scattering**. Quadrupole solutions, which are terms of order higher than the dipole solutions to the Helmholtz equation that we have considered thus far, become important for large spheres, which renders the mathematical analysis more complicated. The directional pattern of the scattering assumes complex, and often asymmetric, shapes so that scattering in the forward direction can become stronger than that in the backward direction. For spheres that are large in comparison with the wavelength, the scattered power turns out to be proportional to the square of the particle diameter, rather than to the sixth power as for Rayleigh scattering.

Moveover, the strength of Mie scattering is roughly independent of wavelength, in contrast to Rayleigh scattering, so all wavelengths in white light are scattered approximately equally. Mie scattering from the water droplets suspended in clouds, which are comparable in size to the visible wavelengths comprising sunlight, is responsible for their white (or gray) color. It is also responsible for the white glare around light sources (such as automobile headlights) in the presence of mist and fog.

D. Attenuation in a Medium with Scatterers

Although the intensity that is Rayleigh scattered from a single scatterer is very small, the cumulative effect of a large number of scatterers distributed within a medium can result in significant attenuation. A wave propagating through a homogeneous medium with an average of N_s identical scatterers per unit volume, each with scattering cross section σ_s, is attenuated exponentially at a rate α_s, known as the **scattering coefficient**:

$$\boxed{\alpha_s = N_s \sigma_s.} \qquad (5.6\text{-}18)$$

Scattering Coefficient

This result is derived by considering a plane wave of intensity I traveling along the z axis of a cylinder with unit cross-sectional area and incremental length Δz, as illustrated in Fig. 5.6-4. The incremental slice contains $N_s\,\Delta z$ scatterers, each of which scatters a small amount of power $\sigma_s I$ away from the z direction. In passing through this slice, the intensity therefore decreases by the increment $\Delta I = -(N_s\,\Delta z)\sigma_s I$. In the limit as $\Delta z \to 0$, this yields $dI/dz = -\alpha_s I$, where $\alpha_s = N_s \sigma_s$, so that the intensity of the wave decays exponentially at the rate α_s, thereby decreasing by the factor $\exp(-\alpha_s z)$ upon traveling a distance z.

Medium with Absorbing Scatterers

If the scatterers are absorptive as well, then additional attenuation is encountered as the wave passes through the medium. The overall attenuation coefficient, also called the intensity **extinction coefficient**,[†] is the sum of the **absorption coefficient** α_a and the **scattering coefficient** α_s, i.e., $\alpha = \alpha_a + \alpha_s$.

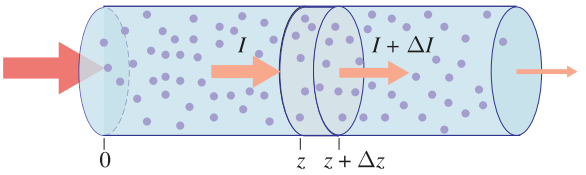


Figure 5.6-4 Scattering and absorption from scatterers embedded in a nonabsorbing homogenous medium results in wave extinction.

We proceed to derive an expression for the absorption coefficient α_a for a nonabsorbing homogeneous medium of real permittivity ϵ in which a concentration of N_s spherical scatterers per unit volume, each of complex permittivity ϵ_s and volume V, is embedded. The absorptive nature of the scatterers is embodied in the imaginary part of ϵ_s. The complex effective permittivity of the composite medium (the host medium and the embedded scatterers) is denoted ϵ_e. The wavenumbers of the composite and host media are $k_e = \omega\sqrt{\epsilon_e \mu}$ and $k = \omega\sqrt{\epsilon\mu}$, which are complex and real, respectively. Based on (5.5-3) we can therefore write $\alpha_a/2 = -\operatorname{Im}\{k_e\} = -k \operatorname{Im}\{\sqrt{\epsilon_e/\epsilon}\}$.

An approximate expression for ϵ_e is provided by the weighted average based on the volume fraction occupied by the scatterers, $\mathrm{f} = N_s V$:

$$\epsilon_e \approx (1 - \mathrm{f})\epsilon + \mathrm{f}\epsilon_s = \epsilon + (\epsilon_s - \epsilon)\mathrm{f}. \qquad (5.6\text{-}19)$$

This expression is valid for a dilute medium ($\mathrm{f} \ll 1$) with weak scattering ($\epsilon_s \approx \epsilon$) and small scatterers. For small spherical scatterers of arbitrary concentration, and for

[†] The field extinction coefficient, in contrast, is the rate at which the field, rather than the intensity, decreases.

arbitrary values of ϵ_s and ϵ, the Maxwell-Garnett mixing rule is applicable:[‡]

$$\epsilon_e \approx \epsilon + 3f\epsilon \frac{\epsilon_s - \epsilon}{\epsilon_s + 2\epsilon - (\epsilon_s - \epsilon)f} = \epsilon \frac{2(1-f)\epsilon + (1+2f)\epsilon_s}{(2+f)\epsilon + (1-f)\epsilon_s}. \tag{5.6-20}$$

This relation may be derived by noting that the average electric field is $\overline{E} = fE_i + (1-f)E_0$ and the average displacement is $\overline{D} = f\epsilon_s E_i + (1-f)\epsilon E_0$, where E_i and E_0 are the internal and external fields, respectively. Using (5.6-17), which provides $E_0/E_i = (\epsilon_s + 2\epsilon)/3\epsilon$, and forming the ratio $\epsilon_e \approx \overline{D}/\overline{E}$, yields the desired result. In a dilute medium of small spherical scatterers, for which $f \ll 1$, (5.6-20) takes the simpler form

$$\epsilon_e \approx \epsilon + 3f\epsilon \frac{\epsilon_s - \epsilon}{\epsilon_s + 2\epsilon}. \tag{5.6-21}$$

If, additionally, the scattering is weak ($\epsilon_s \approx \epsilon$), then (5.6-21) reduces to the weighted-average formula (5.6-19).

If the scatterers are small and dilute, but not necessarily weak, then $\epsilon_e = \epsilon + \Delta\epsilon$, where $\Delta\epsilon \ll \epsilon$. In this case, $\alpha_a/2 = -k\,\mathrm{Im}\{\sqrt{\epsilon_e/\epsilon}\} = -k\,\mathrm{Im}\{(1+\Delta\epsilon/\epsilon)^{1/2}\} \approx -k\,\mathrm{Im}\{1+\Delta\epsilon/2\epsilon\} = -k\,\mathrm{Im}\{\Delta\epsilon/2\epsilon\}$ so that $\alpha_a \approx -k\,\mathrm{Im}\{(\epsilon_e - \epsilon)/\epsilon\}$. With $k = 2\pi/\lambda$ and $f = N_s V = N_s 4\pi a^3/3$ for spherical scatterers of radius a, use of (5.6-21) leads to the approximate result

$$\alpha_a = N_s \sigma_a, \qquad \sigma_a = \pi a^2 Q_a, \qquad Q_a \approx -4\,\mathrm{Im}\left\{\frac{\epsilon_s - \epsilon}{\epsilon_s + 2\epsilon}\right\}\left(2\pi\frac{a}{\lambda}\right), \tag{5.6-22}$$

Absorption Coefficient

where σ_a is the absorption cross section and the dimensionless factor Q_a is the **absorption efficiency**. Note that if we use the weighted-average formula (5.6-19) for ϵ_e, which is valid for dilute, weak, and small scatterers, we obtain an expression for α_a identical to that provided in (5.6-22) except that the factor $\epsilon_s + 2\epsilon$ in Q_a is replaced by 3ϵ. The overall attenuation coefficient $\alpha = \alpha_a + \alpha_s$ is obtained by combining the results provided in (5.6-22), (5.6-16), and (5.6-18).

5.7 PULSE PROPAGATION IN DISPERSIVE MEDIA

The propagation of pulses of light in dispersive media is important in many applications including optical fiber communication systems, as will be discussed in detail in Chapters 10, 23, and 25. As indicated above, a dispersive medium is characterized by a frequency-dependent refractive index and absorption coefficient, so that monochromatic waves of different frequencies travel through the medium with different velocities and undergo different attenuations. Since a pulse of light comprises a sum of many monochromatic waves, each of which is modified differently, the pulse is delayed and broadened (dispersed in time); in general its shape is also altered. In this section we provide a simplified analysis of these effects; a detailed description is deferred to Chapter 23.

[‡] See J. C. Maxwell Garnett, XII. Colours in Metal Glasses and in Metallic Films, *Philosophical Transactions of the Royal Society A*, vol. 203, pp. 385–420, 1904; M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th expanded and corrected ed. 2002.

Group Velocity

Consider a pulsed plane wave traveling in the z direction through a lossless dispersive medium with refractive index $n(\omega)$. Following the example set forth in Sec. 2.6, assume that the initial complex wavefunction at $z = 0$ is $U(0,t) = \mathcal{A}(t)\exp(j\omega_0 t)$, where ω_0 is the central angular frequency and $\mathcal{A}(t)$ is the complex envelope of the wave. It will be shown below that if the dispersion is weak, i.e., if n varies slowly within the spectral bandwidth of the wave, then the complex wavefunction at a distance z is approximately $U(z,t) = \mathcal{A}(t - z/v)\exp[j\omega_0(t - z/c)]$, where $c = c_o/n(\omega_0)$ is the speed of light in the medium at the central frequency, and v is the velocity at which the envelope travels (see Fig. 5.7-1). The parameter v, called the **group velocity**, is given by

$$\frac{1}{v} = \beta' = \frac{d\beta}{d\omega}, \tag{5.7-1}$$

Group Velocity

where $\beta = \omega n(\omega)/c_o$ is the frequency-dependent propagation constant and the derivative in (5.7-1), which is often denoted β', is evaluated at the central frequency ω_0. The group velocity is a characteristic of the dispersive medium, and generally varies with the central frequency. The corresponding time delay $\tau_d = z/v$ is called the **group delay**.

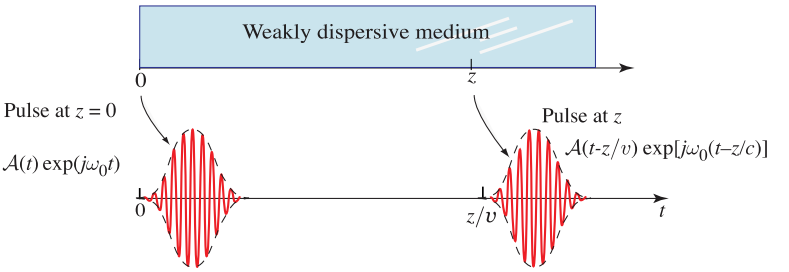


Figure 5.7-1 An optical pulse traveling in a dispersive medium that is weak enough so that its group velocity is frequency independent. The envelope travels with group velocity v while the underlying wave travels with phase velocity c.

Since the phase factor $\exp[j\omega_0(t - z/c)]$ is a function of $t - z/c$, the speed of light c, given by $1/c = \beta(\omega_0)/\omega_0$, is often called the **phase velocity**. In an ideal (nondispersive) medium, $\beta(\omega) = \omega/c$ whereupon $v = c$ and the group and phase velocities are identical.

□ **Derivation of the Formula for the Group Velocity.** The proof of (5.7-1) relies on a Fourier decomposition of the envelope $\mathcal{A}(t)$ into its constituent harmonic functions. A component of frequency Ω, assumed to have a Fourier amplitude $A(\Omega)$, corresponds to a monochromatic wave of frequency $\omega = \omega_0 + \Omega$ traveling with propagation constant $\beta(\omega_0 + \Omega)$. This component of the pulsed plane wave therefore travels as $A(\Omega)\exp\{-j[\beta(\omega_0 + \Omega)]z\}\exp[j(\omega_0 + \Omega)t]$. If $\beta(\omega)$ varies slowly near the central frequency ω_0, it may be approximately linearized via a two-term Taylor series expansion: $\beta(\omega_0 + \Omega) \approx \beta(\omega_0) + \Omega\, d\beta/d\omega = \omega_0/c + \Omega/v$. The Ω component of the complex wavefunction may therefore be approximated by $A(\Omega)\exp[j\Omega(t - z/v)]\exp[j\omega_0(t - z/c)]$. It follows that, upon traveling a distance z, the envelope of the Fourier component $A(\Omega)\exp(j\Omega t)$ becomes $A(\Omega)\exp[j\Omega(t - z/v)]$ for every value of Ω; thus the pulse envelope $\mathcal{A}(t)$ becomes $\mathcal{A}(t - z/v)$. The pulse therefore travels at the group velocity $v = 1/(d\beta/d\omega)$, in accordance with (5.7-1). ■

Since the index of refraction of most materials is typically measured and tabulated as a function of optical wavelength rather than frequency, it is convenient to express the group velocity v in terms of $n(\lambda_o)$. Using the relations $\beta = \omega n(\lambda_o)/c_o = 2\pi n(\lambda_o)/\lambda_o$ and $\lambda_o = 2\pi c_o/\omega$ in (5.7-1), along with the chain rule $d\beta/d\omega = (d\beta/d\lambda_o)(d\lambda_o/d\omega)$, provides

$$\boxed{v = \frac{c_o}{N}} \qquad \boxed{N = n(\lambda_o) - \lambda_o \frac{dn}{d\lambda_o}.} \tag{5.7-2}$$

Group Velocity and Group Index

The derivative $dn/d\lambda_o$ in (5.7-2) is evaluated at the central wavelength λ_0. The parameter N is often called the **group index**.

Group Velocity Dispersion (GVD)

Since the group velocity $v = 1/(d\beta/d\omega)$ is itself often frequency dependent, different frequency components of the pulse undergo different delays $\tau_d = z/v$. As a result, the pulse spreads in time. This phenomenon is known as **group velocity dispersion** (GVD). To estimate the spread associated with GVD it suffices to note that, upon traveling a distance z, two identical pulses of central frequencies ν and $\nu + \delta\nu$ suffer a differential group delay

$$\delta\tau = \frac{d\tau_d}{d\nu}\,\delta\nu = \frac{d}{d\nu}\left(\frac{z}{v}\right)\delta\nu = D_\nu\, z\, \delta\nu, \tag{5.7-3}$$

where the quantity

$$\boxed{D_\nu = \frac{d}{d\nu}\left(\frac{1}{v}\right) = 2\pi\beta''} \tag{5.7-4}$$

Dispersion Coefficient

is called the **dispersion coefficient** and $\beta'' \equiv d^2\beta/d\omega^2|_{\omega_0}$. This effect is actually associated with the higher-order terms in the Taylor series expansion of $\beta(\omega)$ that were neglected in the derivation of the group velocity carried out above; a more complete treatment will be provided in Chapter 23.

If the pulse has an initial spectral width σ_ν (Hz), in accordance with (5.7-3) a good estimate of its temporal spread is then provided by

$$\boxed{\sigma_\tau = |D_\nu|\sigma_\nu z.} \tag{5.7-5}$$

Pulse Spread

The dispersion coefficient D_ν is a measure of the pulse-time broadening per unit distance per unit spectral width (s/m-Hz). This temporal broadening process is illustrated schematically in Fig. 5.7-2. If the refractive index is specified in terms of the wavelength, $n(\lambda_o)$, then (5.7-2) and (5.7-4) give

$$\boxed{D_\nu = \frac{\lambda_o^3}{c_o^2}\frac{d^2n}{d\lambda_o^2}.} \tag{5.7-6}$$

Dispersion Coefficient (s/m-Hz)

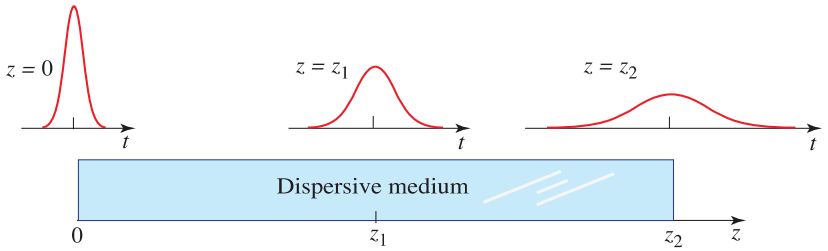


Figure 5.7-2 The temporal spread of an optical pulse traveling in a dispersive medium is proportional to the product of the dispersion coefficient D_ν, the spectral width σ_ν, and the distance traveled z.

It is also common to define a dispersion coefficient D_λ in terms of the wavelength[†] instead of the frequency. Using $D_\lambda\, d\lambda = D_\nu\, d\nu$ yields $D_\lambda = D_\nu\, d\nu/d\lambda_o = D_\nu\,(-c_o/\lambda_o^2)$, which leads directly to

$$D_\lambda = -\frac{\lambda_o}{c_o}\frac{d^2 n}{d\lambda_o^2}. \tag{5.7-7}$$

Dispersion Coefficient (s/m-nm)

In analogy with (5.7-5), for a source of spectral width σ_λ the temporal broadening of a pulse of light is

$$\sigma_\tau = |D_\lambda|\sigma_\lambda z. \tag{5.7-8}$$

Pulse Spread

As discussed in Secs. 10.3, 23.3, and 25.1, D_λ is usually specified in units of ps/km-nm in fiber-optics applications: the pulse broadening is measured in picoseconds, the length of the medium in kilometers, and the source spectral width in nanometers.

Normal and Anomalous Dispersion

Although the sign of the dispersion coefficient D_ν (or D_λ) does not affect the pulse-broadening rate, it does affect the phase of the complex envelope of the optical pulse. As such, the sign can play an important role in pulse propagation through media consisting of cascades of materials with different dispersion properties, as examined in Chapter 23. If $D_\nu > 0$ ($D_\lambda < 0$), the medium is said to exhibit **normal dispersion**. In this case, the travel time for higher-frequency components is greater than the travel time for lower-frequency components so that shorter-wavelength components of the pulse arrive later than longer-wavelength components, as illustrated schematically in Fig. 5.7-3. If $D_\nu < 0$ ($D_\lambda > 0$), the medium is said to exhibit **anomalous dispersion**, in which case the shorter-wavelength components travel faster and arrive earlier. Most glasses exhibit normal dispersion in the visible region of the spectrum; at longer wavelengths, however, the dispersion often becomes anomalous.

Single-Resonance Medium

The group velocity and dispersion coefficient for an optical pulse propagating through a single-resonance medium is determined by substituting (5.5-20) and (5.5-21) into

[†] An alternative definition of the dispersion coefficient, $M = -D_\lambda$, is also widely used in the literature.

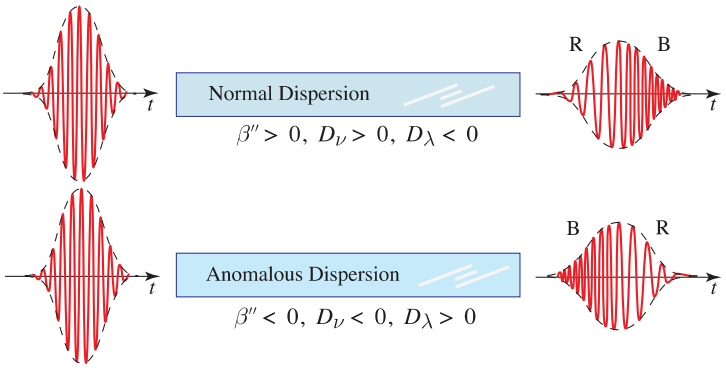


Figure 5.7-3 Propagation of an optical pulse through media with normal and anomalous dispersion. In a medium with normal dispersion the shorter-wavelength components of the pulse (B) arrive later that those with longer wavelength (R). A medium with anomalous dispersion exhibits the opposite behavior. The pulses are said to be chirped since the instantaneous frequency is time varying.

(5.5-5) and making use of (5.7-2) and (5.7-7). To illustrate the behavior of the pulse in this medium, the wavelength dependence of the refractive index n, the group index N, and the dispersion coefficient D_λ, are plotted in Fig. 5.7-4 as a function of normalized wavelength λ/λ_0, for a medium with parameters $\chi_0 = 0.05$ and $\Delta\nu/\nu_0 = 0.1$.

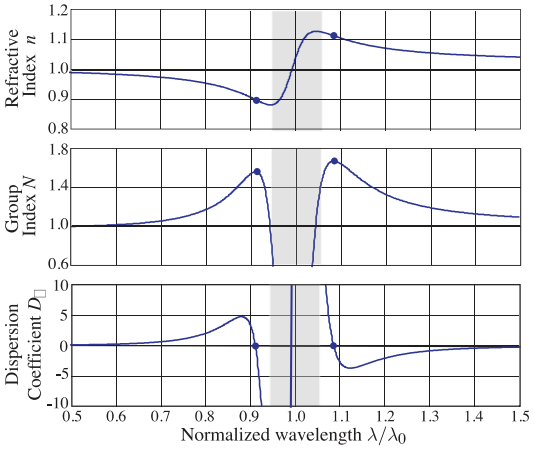


Figure 5.7-4 Wavelength dependence of the optical parameters associated with a single-resonance medium plotted as a function of the wavelength normalized to the wavelength at the resonance frequency, λ/λ_0: the refractive index $n = c_o/c$ (dots indicate points of inflection), the group index $N = c_o/v$ (dots indicate maxima), and the dispersion coefficient D_λ (dots indicate zeros). The parameters N and D_λ are not meaningful near resonance (shaded area).

In the vicinity of the resonance (shaded area in figure), n varies sufficiently rapidly with wavelength that the parameters N and D_λ, which are defined on the basis of a Taylor-series approximation comprising a few terms, cease to be meaningful. Away from the resonance, on both sides thereof, the refractive index decreases monotonically with increasing wavelength and exhibits points of inflection (indicated by dots). The first derivative of the refractive index achieves local maxima at these locations so that the group index N attains its maximum values there. Moreover, the second derivative vanishes at these points so that the dispersion coefficient changes sign. As the wavelength approaches the resonance wavelength from below, the dispersion changes from anomalous to normal; the reverse is true as the wavelength approaches resonance from above, as is evident in Fig. 5.7-4.

Fast and Slow Light in Resonant Media

As is evident in Fig. 5.7-4, in a resonant medium the refractive index n and the group index N undergo rapid changes near the resonance frequency, and may be substantially greater or smaller than unity. Consequently, the phase velocity $c = c_o/n$ and the group velocity $v = c_o/N$ may be significantly less than, or greater than, the velocity of light in free space, c_o. The group index, and hence the group velocity, may even be negative. This raises the question of a potential conflict with causality and the special theory of relativity, which provides that information cannot be transmitted at a velocity greater than c_o. It turns out that there is no such conflict since neither the group velocity nor the phase velocity corresponds to the **information velocity**, which is the speed at which information is transmitted between two points. The information velocity may be determined by tracing the propagation of a nonanalytic point on the pulse, for example, the onset of a rectangular pulse. It cannot exceed c_o.

The concepts of phase and group velocity were considered earlier in the context of an optical pulse traveling in a weakly dispersive medium, i.e., a medium with propagation constant $\beta(\omega)$ that is approximately linear in the vicinity of the central frequency of the pulse, ω_0. After traveling a distance z, the pulse is delayed by a time z/v and is modulated by a phase factor $\exp(-j\omega_0 z/c)$. This phase, which travels at the phase velocity c, carries no information. It is the group velocity v that governs the time of "arrival" of the pulse. Since, in this approximation, the pulse envelope maintains its shape as it travels (Fig. 5.7-1), the group velocity is a good approximation of the information velocity. In the resonant medium, this occurs at wavelengths far from resonance, where the group index is greater than unity and the group velocity is less than c_o.

At frequencies closer to resonance, higher-order dispersion terms become appreciable. In the presence of second-order dispersion (GVD), but negligible higher-order dispersion, a Gaussian pulse, for example, remains Gaussian, albeit with an increased width; its peak travels at the group velocity v. However, since the Gaussian pulse has a continuous profile and infinite support (i.e., extends over all time), the velocity at which the peak travels is not necessarily the information velocity; indeed, it can be greater than the free-space speed of light.

In the immediate vicinity of resonance, where the group velocity can be significantly greater than c_o and can even be negative (Fig. 5.7-4), higher-order dispersion terms must be considered. The pulse shape can then be significantly altered and the group velocity can no longer be considered as a possible information velocity. For sufficiently short distances, however, the pulse may travel without a significant alteration in shape, and this may occur at a group velocity significantly higher than c_o. The pulse can also travel at a negative group velocity, signifying that a point on the pulse, identified by a peak for example, arrives at the end of the medium before the corresponding point on the input pulse even enters the medium! In the opposite limit of slow light, certain special resonance media permit the group velocity of light to be made exceedingly small so that a light pulse may be substantially slowed or even halted. It should be emphasized, however, that in none of these situations does the information velocity exceed c_o.

Since fast- and slow-light phenomena can only be observed near resonance, where the absorption coefficient is large (and frequency dependent), a mechanism for optical amplification is necessary, and nonlinear effects are often exploited to enhance this phenomenon.

EXAMPLE 5.7-1. ***Dispersion in a Multi-Resonance Medium: Fused Silica.*** In the region between 0.21 and 3.71 μm, the wavelength dependence of the refractive index n for fused silica at room temperature is well characterized by the Sellmeier equation (5.5-28). Expressing all wavelengths in μm, this is achieved using three resonance wavelengths at $\lambda_1 = 0.06840$ μm, $\lambda_2 =$

0.1162 μm, and λ_3 = 9.8962 μm, with weights χ_{01} = 0.6962, χ_{02} = 0.4079, and χ_{03} = 0.8975, respectively. Expressions for the group index N and the dispersion coefficient D_λ are readily derived from this equation by means of (5.7-2) and (5.7-7). The results of this calculation in the 600–1600-nm wavelength range are presented in Fig. 5.7-5.

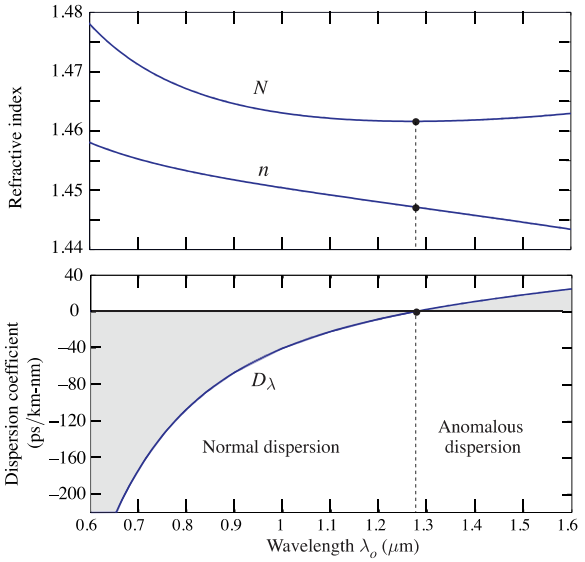


Figure 5.7-5 Wavelength dependence of optical parameters for fused silica calculated on the basis of the Sellmeier equation (5.5-28): the refractive index $n = c_o/c$ (dot indicates point of inflection), the group index $N = c_o/v$ (dot indicates minimum), and the dispersion coefficient D_λ (dot indicates zero).

The refractive index n is seen to decrease monotonically with increasing wavelength, and to exhibit a point of inflection at $\lambda_o = 1.276$ μm. At this wavelength, the group index N is minimum so that the group velocity $v = c_o/N$ is maximum. Since the dispersion coefficient D_λ is proportional to the second derivative of n with respect to λ_o, it vanishes at this wavelength. Zero dispersion coefficient signifies minimal pulse broadening. At wavelengths shorter than 1.276 μm, $D_\lambda < 0$ and the medium exhibits normal dispersion whereas at longer wavelengths, $D_\lambda > 0$ and the dispersion is anomalous. The presence of a zero-dispersion wavelength offers significant advantages in the design of optical fiber communication systems in which optical pulses carry information, as will become evident in Secs. 10.3, 25.1, and 23.3. The silica-glass fibers used in such systems are doped and exhibit zero dispersion close to 1.312 μm. The group index N is seen to be larger than the refractive index n by $\sim 2\%$, as is common in many materials.

READING LIST

Electromagnetics

See also the reading list on general optics in Chapter 1.

J.-M. Liu, *Principles of Photonics*, Cambridge University Press, 2016.

N. Ida, *Engineering Electromagnetics*, Springer-Verlag, 3rd ed. 2015.

J. C. Rautio, The Long Road to Maxwell's Equations, *IEEE Spectrum*, vol. 51, no. 12, pp. 36–40 & 54–56, 2014.

F. T. Ulaby and U. Ravaioli, *Fundamentals of Applied Electromagnetics*, Prentice Hall, 7th ed. 2014.

M. N. O. Sadiku, *Elements of Electromagnetics*, Oxford University Press, 6th ed. 2014.

U. S. Inan, A. Inan, and R. Said, *Engineering Electromagnetics and Waves*, Prentice Hall, 2nd ed. 2014.

A. Zangwill, *Modern Electrodynamics*, Cambridge University Press, 2013.

D. Fleisch, *A Student's Guide to Maxwell's Equations*, Cambridge University Press, 2013.

A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, 2nd ed. 2012.

V. Lucarini, J. J. Saarinen, K.-E. Peiponen, and E. M. Vartiainen, *Kramers–Kronig Relations in Optical Materials Research*, Springer-Verlag, 2005.

S. A. Akhmanov and S. Yu. Nikitin, *Physical Optics*, Oxford University Press, 1997.

Electromagnetics Classics

J. D. Jackson, *Classical Electrodynamics*, Wiley, 3rd ed. 1999.

S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, Wiley, 3rd ed. 1994.

H. A. Haus and J. R. Melcher, *Electromagnetic Fields and Energy*, Prentice Hall, 1989.

H. A. Haus, *Waves and Fields in Optoelectronics*, Prentice Hall, 1984.

L. D. Landau, E. M. Lifshitz, and L. P. Pitaevskii, *Electrodynamics of Continuous Media*, Nauka (Moscow), 2nd ed. 1982; Butterworth–Heinemann, 2nd English ed. 1984, reprinted 2004.

L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, Nauka (Moscow), 6th revised ed. 1973; Butterworth–Heinemann, 4th revised English ed. 1975, reprinted with corrections 2000.

J. A. Stratton, *Electromagnetic Theory*, McGraw–Hill, 1941; Wiley–IEEE Classic reissue 2007.

H. A. Lorentz, *The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat*, Teubner, 1906; Dover, reissued 2004.

O. Heaviside, *Electromagnetic Theory*, "The Electrician" Printing and Publishing (London), 1893.

J. C. Maxwell, *A Treatise on Electricity and Magnetism*, Macmillan/Clarendon Press (Oxford), 1873.

Vector Beams

Q. Zhan, ed., *Vectorial Optical Fields*, World Scientific, 2014.

R. Dorn, S. Quabis, and G. Leuchs, Sharper Focus for a Radially Polarized Light Beam, *Physical Review Letters*, vol. 91, 233901, 2003.

D. G. Hall, Vector-Beam Solutions of Maxwell's Wave Equation, *Optics Letters*, vol. 21, pp. 9–11, 1996.

Optical Constants

P. Hartmann, *Optical Glass*, SPIE Optical Engineering Press, 2014.

M. Bass and V. N. Mahajan, eds., *Handbook of Optics*, McGraw–Hill, 3rd ed. 2010.

A. R. Hilton, Sr., *Chalcogenide Glasses for Infrared Optics*, McGraw–Hill, 2010.

E. D. Palik, ed., *Handbook of Optical Constants of Solids III*, Academic Press, 1998.

Y.-M. Chiang, D. Birnie III, and W. D. Kingery, *Physical Ceramics: Principles for Ceramic Science and Engineering*, Wiley, 1997.

W. L. Wolfe and G. J. Zissis, eds., *The Infrared Handbook*, Environmental Research Institute of Michigan, revised ed. 1993.

Fast and Slow Light

J. B. Khurgin and R. S. Tucker, eds., *Slow Light: Science and Applications*, CRC Press/Taylor & Francis, 2009.

P. W. Milonni, *Fast Light, Slow Light and Left-Handed Light*, Taylor & Francis, 2005.

L. Brillouin, *Wave Propagation and Group Velocity*, Academic Press, 1960.

Light Scattering

J. A. Adam, *Rays, Waves, and Scattering: Topics in Classical Mathematical Physics*, Princeton University Press, 2017.

G. Gouesbet and G. Gréhan, *Generalized Lorenz-Mie Theories*, Springer-Verlag, 2nd ed. 2017.

B. Chu, *Laser Light Scattering: Basic Principles and Practice*, Dover, paperback 2nd ed. 2007.

M. Nieto-Vesperinas, *Scattering and Diffraction in Physical Optics*, World Scientific, 2nd ed. 2006.

H. M. Nussenzveig, *Diffraction Effects in Semiclassical Scattering*, Cambridge University Press, 1992, paperback ed. 2006.

L. Tsang, J. A. Kong and K-H. Ding, *Scattering of Electromagnetic Waves: Theories and Applications*, Wiley, 2000.

B. J. Berne and R. Pecora, *Dynamic Light Scattering: With Applications to Chemistry, Biology, and Physics*, Wiley, 1976; Dover, reissued 2000.

M. I. Mishchenko, J. W. Hovenier, and L. D. Travis, eds., *Light Scattering by Nonspherical Particles: Theory, Measurements, and Applications*, Academic Press, 2000.

M. Kerker, ed., *Selected Papers on Light Scattering*, SPIE Optical Engineering Press (Milestone Series Volume 4), 1988.

C. F. Bohren and D. R. Huffman, *Absorption and Scattering of Light by Small Particles*, Wiley, 1983, paperback ed. 1998.

H. C. van de Hulst, *Light Scattering by Small Particles*, Wiley, 1957; Dover, reissued 1981.

Optical Pulse Propagation

K. E. Oughstun, *Electromagnetic and Optical Pulse Propagation 1: Spectral Representations in Temporally Dispersive Media*, Springer-Verlag, 2007, paperback ed. 2010.

PROBLEMS

5.1-1 **An Electromagnetic Wave.** An electromagnetic wave in free space has an electric-field vector $\mathcal{E} = f(t - z/c_o)\widehat{\mathbf{x}}$, where $\widehat{\mathbf{x}}$ is a unit vector in the x direction, and $f(t) = \exp(-t^2/\tau^2)\exp(j2\pi\nu_0 t)$, where τ is a constant. Describe the physical nature of this wave and determine an expression for the magnetic-field vector.

5.2-1 **Dielectric Media.** Identify the media described by the following equations, with respect to linearity, dispersiveness, spatial dispersiveness, and homogeneity. Assume that all media are isotropic.

(a) $\mathcal{P} = \epsilon_o \chi \mathcal{E} - a\nabla \times \mathcal{E}$,

(b) $\mathcal{P} + a\mathcal{P}^2 = \epsilon_o \mathcal{E}$,

(c) $a_1\, \partial^2\mathcal{P}/\partial t^2 + a_2\, \partial\mathcal{P}/\partial t + \mathcal{P} = \epsilon_o \chi \mathcal{E}$,

(d) $\mathcal{P} = \epsilon_o\{a_1 + a_2 \exp[-(x^2 + y^2)]\}\mathcal{E}$,

where χ, a, a_1, and a_2 are constants.

5.3-1 **Traveling Standing Wave.** The electric-field complex-amplitude vector for a monochromatic wave of wavelength λ_o traveling in free space is $\mathbf{E}(\mathbf{r}) = E_0 \sin(\beta y) \exp(-j\beta z)\,\widehat{\mathbf{x}}$.

(a) Determine a relation between β and λ_o.

(b) Derive an expression for the magnetic-field complex-amplitude vector $\mathbf{H}(\mathbf{r})$.

(c) Determine the direction of the flow of optical power.

(d) This wave may be regarded as the sum of two TEM plane waves. Determine their directions of propagation.

5.4-1 **Electric Field of Focused Light.**

(a) 1 W of optical power is focused uniformly on a flat target of size 0.1×0.1 mm^2 placed in free space. Determine the peak value of the electric field E_0 (V/m). Assume that the optical wave is approximated as a TEM plane wave within the area of the target.

(b) Determine the electric field at the center of a Gaussian beam (the point on the beam axis located at the beam waist) if the beam power is 1 W and the beam waist radius $W_0 = 0.1$ mm (refer to Sec. 3.1).

5.5-2 **Amplitude-Modulated Wave in a Dispersive Medium.** An amplitude-modulated wave whose complex wavefunction takes the form $A(t) = [1 + m\cos(2\pi f_s t)]\exp(j2\pi\nu_0 t)$ at $z = 0$, where $f_s \ll \nu_0$, travels a distance z through a dispersive medium of propagation constant $\beta(\nu)$ and negligible attenuation. If $\beta(\nu_0) = \beta_0$, $\beta(\nu_0 - f_s) = \beta_1$, and $\beta(\nu_0 + f_s) = \beta_2$, derive an expression for the complex envelope of the transmitted wave as a function of β_0, β_1, β_2, and z. Show that at certain distances z the wave is amplitude modulated with no phase modulation.

5.7-1 **Group Velocity Dispersion in a Medium Described by the Sellmeier Equation.**

(a) Derive expressions for the group index N and the group velocity dispersion coefficient D_λ for a medium whose refractive index is described by the Sellmeier equation (5.5-28).

(b) Plot the wavelength dependence of n, N, and D_λ for fused silica in the region between 0.25 and 3.5 μm. Make use of the parameters provided in Table 5.5-1 (and in Example 5.7-1). Verify the curves provided in Fig. 5.7-5.

(c) Construct a similar collection of plots for GaAs in the region between 1.5 and 10.5 μm. As indicated in Table 5.5-1, in the wavelength region between 1.4 and 11 μm, at room temperature, GaAs is characterized by a 3-term Sellmeier equation with resonance wavelengths at 0 μm, 0.4082 μm, and 37.17 μm, with associated weights given by 3.5, 7.4969, and 1.9347, respectively. Compare and contrast the behavior of the dispersion properties of fused silica with those of GaAs.

5.7-2 **Refractive Index of Air.** The refractive index of air can be precisely measured with the help of a Michelson interferometer and a tunable light source. At atmospheric pressure, and a temperature of 20° C, the refractive index of air differs from unity by $n - 1 = 2.672 \times 10^{-4}$ at a wavelength of 0.76 μm, by $n - 1 = 2.669 \times 10^{-4}$ at a wavelength of 0.81 μm, and by $n - 1 = 2.665 \times 10^{-4}$ at a wavelength of 0.86 μm.

(a) Using a quadratic fit to these data, determine the wavelength dependence of the group velocity.

(b) Obtain an expression for the dispersion coefficient D_λ in ps/km-nm and compare your result with that for a silica optical fiber.

CHAPTER

6

POLARIZATION OPTICS

6.1 POLARIZATION OF LIGHT 211
- A. Polarization
- B. Matrix Representation

6.2 REFLECTION AND REFRACTION 221

6.3 OPTICS OF ANISOTROPIC MEDIA 227
- A. Refractive Indices
- B. Propagation Along a Principal Axis
- C. Propagation in an Arbitrary Direction
- D. Dispersion Relation, Rays, Wavefronts, and Energy Transport
- E. Double Refraction

6.4 OPTICAL ACTIVITY AND MAGNETO-OPTICS 240
- A. Optical Activity
- B. Magneto-Optics: The Faraday Effect

6.5 OPTICS OF LIQUID CRYSTALS 244

6.6 POLARIZATION DEVICES 247
- A. Polarizers
- B. Wave Retarders
- C. Polarization Rotators
- D. Nonreciprocal Polarization Devices

The French physicist **Augustin-Jean Fresnel (1788–1827)** put forth a transverse wave theory of light. Equations describing the partial reflection and refraction of light are named in his honor. Fresnel also made important contributions to the theory of light diffraction.

Sir George Gabriel Stokes (1819–1903), an Irish mathematician and physicist, developed a description of light that encompasses intensity as well as state of polarization. He also made seminal contributions to wave optics, fluorescence, and optical aberrations.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

The polarization of light at a given position in space is determined by the path taken by its electric-field vector $\mathcal{E}(\mathbf{r}, t)$ in time. In a simple medium, this vector lies in a plane tangential to the wavefront at that position. For monochromatic light, any two orthogonal components of the complex-amplitude vector $\mathbf{E}(\mathbf{r})$ in that plane vary sinusoidally in time, with amplitudes and phases that generally differ, so that the endpoint of the vector $\mathbf{E}(\mathbf{r})$ traces out an ellipse. Since the directions of the wavefront normals are generally position-dependent, so too are the tangential planes, along with the orientations and shapes of the ellipses, as illustrated in Fig. 6.0-1(a).

For a plane wave, however, the wavefronts are parallel transverse planes and the polarization ellipses are the same everywhere, as illustrated in Fig. 6.0-1(b), although the field vectors need not be parallel at any given time. A plane wave is therefore described by a single ellipse and is said to be **elliptically polarized**. The orientation and ellipticity of the polarization ellipse determine the state of polarization of the plane wave, while its size is established by the optical intensity. When the ellipse reduces to a straight line or becomes a circle, the wave is said to be **linearly polarized** or **circularly polarized**, respectively.

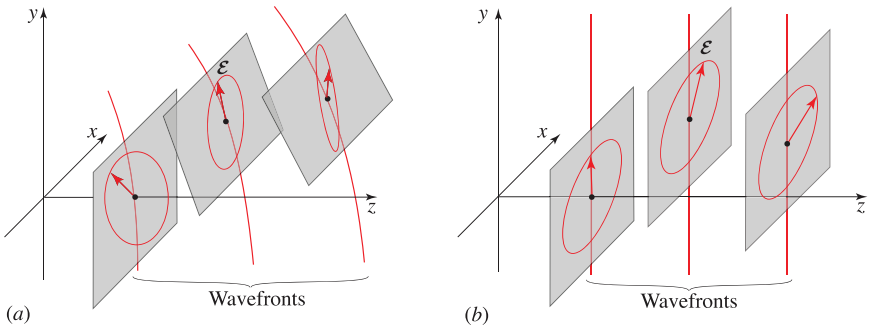


Figure 6.0-1 Trace of the time course of the electric-field vector endpoint for monochromatic light at several positions. (a) Arbitrary wave. (b) Plane or paraxial wave traveling in the z direction.

In paraxial optics, light propagates along directions that lie within a narrow cone centered about the optical axis (the z axis). Waves are approximately transverse electromagnetic (TEM) in character and the electric-field vectors therefore lie approximately in transverse planes, with negligible axial components. From the perspective of polarization, paraxial waves may be approximated by plane waves and described by a single polarization ellipse (or line or circle).

Polarization plays an important role in the interaction of light with matter as attested to by the following examples:

- The proportion of light reflected at the boundary between two materials depends on the polarization of the incident wave.
- The proportion of light absorbed by certain materials is polarization-dependent.
- Light scattering from matter is generally polarization-sensitive.
- The refractive index of anisotropic materials depends on the polarization. Waves with different polarizations travel at different velocities and undergo different phase shifts, so that the polarization ellipse is modified as the wave advances (linearly polarized light can be transformed into circularly polarized light, as a simple example). This feature finds use in the design of many optical devices.

- The polarization plane of linearly polarized light is rotated by passage through certain materials, including optically active media, liquid crystals, and some substances in the presence of an external magnetic field.

This Chapter

This chapter is devoted to a description of elementary polarization phenomena and a number of their applications. Elliptically polarized light is introduced in Sec. 6.1 using a matrix formalism that is convenient for describing polarization devices. Section 6.2 describes the effect of polarization on the reflection and refraction of light at boundaries between dielectric media. The propagation of light through anisotropic media (crystals), optically active media, and liquid crystals are the topics of Secs. 6.3, 6.4, and 6.5, respectively. Finally, elementary polarization devices (such as polarizers, retarders, rotators, and isolators) are discussed in Sec. 6.6. Polarization is important for understanding how light behaves in photonic crystals, metals, and metamaterials (Chapters 7 and 8), and how light is guided (Chapters 9 and 10) and stored (Chapter 11). The polarization properties of random light are considered in Sec.12.4. Polarization clearly plays a central role in many areas of optics and photonics.

6.1 POLARIZATION OF LIGHT

A. Polarization

Consider a monochromatic plane wave of frequency ν and angular frequency $\omega = 2\pi\nu$ traveling in the z direction with velocity c. The electric field lies in the x–y plane and is generally described by

$$\boldsymbol{\mathcal{E}}(z,t) = \text{Re}\left\{\mathbf{A}\exp\left[j\,\omega\left(t-\frac{z}{c}\right)\right]\right\}, \tag{6.1-1}$$

where the complex envelope

$$\mathbf{A} = A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}}, \tag{6.1-2}$$

is a vector with complex components A_x and A_y. To describe the polarization of this wave, we trace the endpoint of the vector $\boldsymbol{\mathcal{E}}(z,t)$ at each position z as a function of time.

Polarization Ellipse

Expressing A_x and A_y in terms of their magnitudes and phases, $A_x = \mathrm{a}_x\exp(j\varphi_x)$ and $A_y = \mathrm{a}_y\exp(j\varphi_y)$, and substituting into (6.1-2) and (6.1-1) we obtain

$$\boldsymbol{\mathcal{E}}(z,t) = \mathcal{E}_x\hat{\mathbf{x}} + \mathcal{E}_y\hat{\mathbf{y}}, \tag{6.1-3}$$

where

$$\mathcal{E}_x = \mathrm{a}_x\cos\left[\omega\left(t-\frac{z}{c}\right)+\varphi_x\right] \tag{6.1-4a}$$

$$\mathcal{E}_y = \mathrm{a}_y\cos\left[\omega\left(t-\frac{z}{c}\right)+\varphi_y\right] \tag{6.1-4b}$$

are the x and y components of the electric-field vector $\boldsymbol{\mathcal{E}}(z,t)$. The components $\mathcal{E}_x$ and $\mathcal{E}_y$ are periodic functions of $(t-z/c)$ that oscillate at frequency ν. Equations (6.1-4) are the parametric equations of the ellipse

$$\frac{\mathcal{E}_x^2}{\mathrm{a}_x^2} + \frac{\mathcal{E}_y^2}{\mathrm{a}_y^2} - 2\cos\varphi\,\frac{\mathcal{E}_x\mathcal{E}_y}{\mathrm{a}_x\mathrm{a}_y} = \sin^2\varphi, \tag{6.1-5}$$

where $\varphi = \varphi_y - \varphi_x$ is the phase difference.

At a fixed value of z, the tip of the electric-field vector rotates periodically in the x–y plane, tracing out this ellipse. At a fixed time t, the locus of the tip of the electric-field vector traces out a helical trajectory in space that lies on the surface of an elliptical cylinder (see Fig. 6.1-1). The electric field rotates as the wave advances, repeating its motion periodically for each distance corresponding to a wavelength $\lambda = c/\nu$.

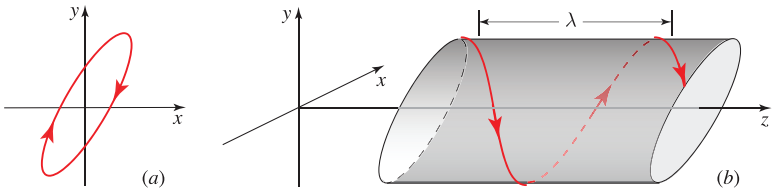
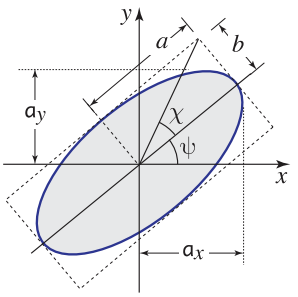


Figure 6.1-1 (a) Rotation of the endpoint of the electric-field vector in the x–y plane at a fixed position z. (b) Trajectory of the endpoint of the electric-field vector as the wave advances.

The state of polarization of the wave is determined by the orientation and shape of the **polarization ellipse**, which is characterized by the two angles defined in Fig. 6.1-2: the angle ψ determines the direction of the major axis, whereas the angle χ determines the ellipticity, namely the ratio of the minor to major axes of the ellipse b/a. These angles depend on the ratio of the complex-envelope magnitudes $\mathrm{R} = \mathrm{a}_y/\mathrm{a}_x$, and on the phase difference $\varphi = \varphi_y - \varphi_x$, in accordance with the following relations:



$$\tan 2\psi = \frac{2\mathrm{R}}{1-\mathrm{R}^2}\cos\varphi\,, \qquad \mathrm{R} = \frac{\mathrm{a}_y}{\mathrm{a}_x}\,, \tag{6.1-6}$$

$$\sin 2\chi = \frac{2\mathrm{R}}{1+\mathrm{R}^2}\sin\varphi\,, \qquad \varphi = \varphi_y - \varphi_x\,. \tag{6.1-7}$$

Figure 6.1-2
Polarization ellipse.

Equations (6.1-6) and (6.1-7) may be derived by finding the angle ψ that achieves a transformation of the coordinate system of $\mathcal{E}_x$ and $\mathcal{E}_y$ in (6.1-5) such that the rotated ellipse has no cross term. The size of the ellipse is determined by the intensity of the wave, which is proportional to $|A_x|^2 + |A_y|^2 = \mathrm{a}_x^2 + \mathrm{a}_y^2$.

Linearly Polarized Light

If one of the components vanishes ($\mathrm{a}_x = 0$, for example), the light is **linearly polarized** (**LP**) in the direction of the other component (the y direction). The wave is also linearly polarized if the phase difference $\varphi = 0$ or π, since (6.1-5) then yields $\mathcal{E}_y = \pm(\mathrm{a}_y/\mathrm{a}_x)\mathcal{E}_x$, which is the equation of a straight line of slope $\pm\mathrm{a}_y/\mathrm{a}_x$ (the $+$ and $-$ signs correspond to $\varphi = 0$ and π, respectively). In these cases, the elliptical cylinder in Fig. 6.1-1(b) collapses to a plane, as illustrated in Fig. 6.1-3. The wave is therefore also said to have **planar polarization**. As an example, if $\mathrm{a}_x = \mathrm{a}_y$, the plane of polarization makes an angle of $45°$ with respect to the x axis. If $\mathrm{a}_x = 0$, on the other hand, the plane of polarization is the y–z plane.

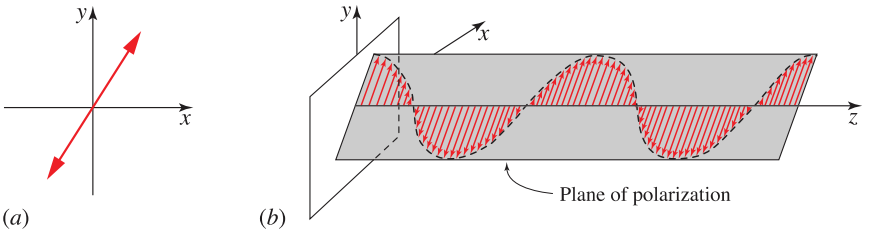


Figure 6.1-3 Linearly polarized light (also called plane polarized light). (a) Time course of the field at a fixed position z. (b) Endpoint of the electric-field vector at position z at a fixed time t.

Circularly Polarized Light

If $\varphi = \pm\pi/2$ and $\mathcal{A}_x = \mathcal{A}_y = \mathcal{A}_0$, (6.1-4) gives $\mathcal{E}_x = \mathcal{A}_0 \cos[\omega(t - z/c) + \varphi_x]$ and $\mathcal{E}_y = \mp\mathcal{A}_0 \sin[\omega(t - z/c) + \varphi_x]$, from which $\mathcal{E}_x^2 + \mathcal{E}_y^2 = \mathcal{A}_0^2$, which is the equation of a circle. The elliptical cylinder in Fig. 6.1-1(b) becomes a circular cylinder and the wave is said to be circularly polarized. In the case $\varphi = -\pi/2$, the electric field at a fixed position z rotates in a counterclockwise direction when viewed from the direction toward which the wave is approaching. The light is then said to be **left circularly polarized** (**LCP**) [Fig. 6.1-4(a)]. The case $\varphi = +\pi/2$ corresponds to clockwise rotation and **right circularly polarized** (**RCP**) light.[†] In the left circular case, the locus traced by the endpoint of the electric-field vector at different positions is a left-handed helix, as illustrated in Fig. 6.1-4(b). For right circular polarization, it is a right-handed helix. The helical path of the *endpoint* of the electric-field vector for the plane-wave circularly polarized light considered here is to be distinguished from the helical *wavefront* of the Laguerre–Gaussian beam discussed in Sec. 3.4 [Fig. 3.4-1(c)].

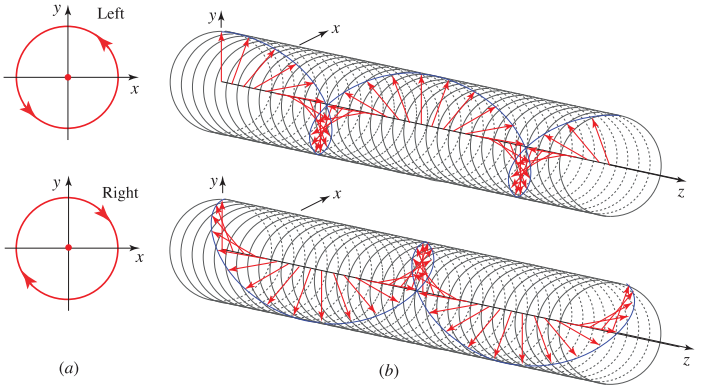


Figure 6.1-4 Motion of the endpoints of the electric-field vectors for left and right circularly polarized plane waves. (a) Time course at a fixed position z. (b) Trajectories of the endpoints.

Poincaré Sphere and Stokes Parameters

As indicated above, the state of polarization of a light wave can be described by two real parameters: the magnitude ratio $\mathrm{R} = \mathcal{A}_y/\mathcal{A}_x$ and the phase difference $\varphi = \varphi_y - \varphi_x$. These are sometimes lumped into a single complex number $\mathrm{R}\exp(j\varphi)$, called the **complex polarization ratio**. Alternatively, we may characterize the state of polarization by the two angles ψ and χ, which represent the orientation and ellipticity of the polarization ellipse, respectively, as defined in Fig. 6.1-2.

[†] This convention is used in most optics textbooks. The opposite designation is often used in the engineering literature: in the case of right (left) circularly polarized light, the electric-field vector at a fixed position rotates counterclockwise (clockwise) when viewed from the direction toward which the wave is approaching.

The **Poincaré sphere** (see Fig. 6.1-5) is a geometrical construct in which the state of polarization is represented by a point on the surface of a sphere of unit radius, with coordinates $(r = 1,\ \theta = 90^\circ - 2\chi,\ \phi = 2\psi)$ in a spherical coordinate system. Each point on the sphere represents a polarization state. For example, points on the equator $(\chi = 0^\circ)$ represent states of linear polarization, with the two points $2\psi = 0^\circ$ and $2\psi = 180^\circ$ representing linear polarization along the x and y axes, respectively. The north and south poles $(2\chi = \pm 90^\circ)$ represent right-handed and left-handed circular polarization, respectively. Other points on the surface of the sphere represent states of elliptical polarization.

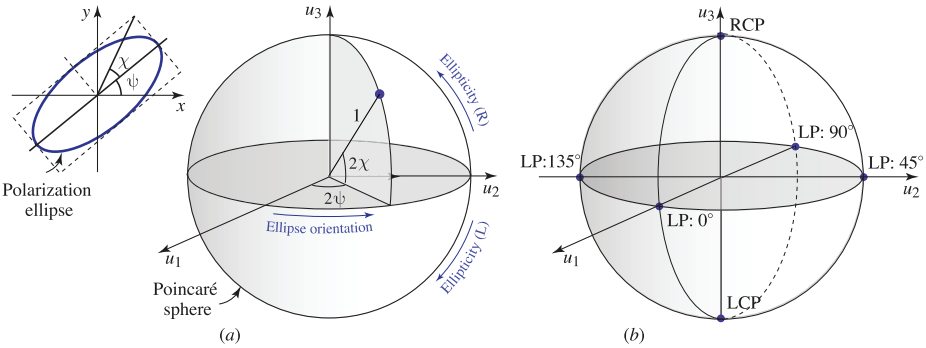


Figure 6.1-5 (a) The orientation and ellipticity of the polarization ellipse are represented geometrically as a point on the surface of the Poincaré sphere. (b) Points on the Poincaré sphere representing linearly polarized (LP) light at various angles with the x direction, as well as right-circularly polarized (RCP) and left-circularly polarized (LCP) light. Points in the interior of the sphere represent partially polarized light (the point at the origin of the sphere represents unpolarized light), as illustrated in Fig. 12.4-1.

The two real quantities (R, φ), or equivalently the angles (χ, ψ), describe the state of polarization but contain no information about the intensity of the wave. Another representation that does contain such information is the **Stokes vector**. This is a set of four real numbers $(\mathrm{S}_0, \mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3)$, called the **Stokes parameters**. The first of these, $\mathrm{S}_0 = \mathrm{a}_x^2 + \mathrm{a}_y^2$, is proportional to the optical intensity whereas the other three, $(\mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3)$, are the Cartesian coordinates of the point on the Poincaré sphere, $(u_1, u_2, u_3) = (\cos 2\chi \cos 2\psi,\ \cos 2\chi \sin 2\psi,\ \sin 2\chi)$, multiplied by S_0, so that

$$\mathrm{S}_1 = \mathrm{S}_0 \cos 2\chi \cos 2\psi \tag{6.1-8a}$$

$$\mathrm{S}_2 = \mathrm{S}_0 \cos 2\chi \sin 2\psi \tag{6.1-8b}$$

$$\mathrm{S}_3 = \mathrm{S}_0 \sin 2\chi. \tag{6.1-8c}$$

Using (6.1-6) and (6.1-7), together with a few trigonometric identities, the Stokes parameters in (6.1-8) may be expressed in terms of the field parameters $(\mathrm{a}_x, \mathrm{a}_y, \varphi)$, and in terms of the components of the complex envelope (A_x, A_y), as:

$$\mathrm{S}_0 = \mathrm{a}_x^2 + \mathrm{a}_y^2 \qquad = |A_x|^2 + |A_y|^2 \tag{6.1-9a}$$

$$\mathrm{S}_1 = \mathrm{a}_x^2 - \mathrm{a}_y^2 \qquad = |A_x|^2 - |A_y|^2 \tag{6.1-9b}$$

$$\mathrm{S}_2 = 2\mathrm{a}_x \mathrm{a}_y \cos\varphi = 2\,\mathrm{Re}\{A_x^* A_y\} \tag{6.1-9c}$$

$$\mathrm{S}_3 = 2\mathrm{a}_x \mathrm{a}_y \sin\varphi = 2\,\mathrm{Im}\{A_x^* A_y\}\,. \tag{6.1-9d}$$

Stokes Parameters

Since $\mathtt{S}_1^2 + \mathtt{S}_2^2 + \mathtt{S}_3^2 = \mathtt{S}_0^2$, only three of the four components of the Stokes vector are independent; they completely define the intensity and the state of polarization of the light. A generalization of the Stokes parameters suitable for describing partially coherent light is presented in Sec. 12.4.

We conclude that there are a number of equivalent representations for describing the state of polarization of an optical field: (1) the polarization ellipse; (2) the Poincaré sphere; and (3) the Stokes vector. Yet another equivalent representation, the Jones vector, is introduced in the following section.

B. Matrix Representation

The Jones Vector

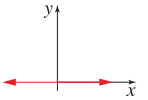
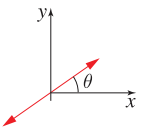
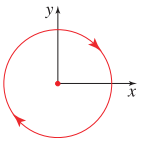
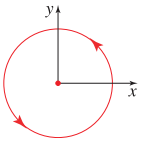
As indicated above, a monochromatic plane wave of frequency ν traveling in the z direction is completely characterized by the complex envelopes $A_x = \mathtt{a}_x \exp(j\varphi_x)$ and $A_y = \mathtt{a}_y \exp(j\varphi_y)$ of the x and y components of the electric-field vector. These complex quantities may be written in the form of a column matrix known as the **Jones vector**:

$$\mathbf{J} = \begin{bmatrix} A_x \\ A_y \end{bmatrix}. \tag{6.1-10}$$

Given $\mathbf{J}$, we can determine the total intensity of the wave, $I = (|A_x|^2 + |A_y|^2)/2\eta$, and use the ratio $\mathtt{R} = \mathtt{a}_y/\mathtt{a}_x = |A_y|/|A_x|$ and the phase difference $\varphi = \varphi_y - \varphi_x = \arg\{A_y\} - \arg\{A_x\}$ to determine the orientation and shape of the polarization ellipse, as well as the Poincaré sphere and the Stokes parameters.

The Jones vectors for some special polarization states are provided in Table 6.1-1. The intensity in each case has been normalized so that $|A_x|^2 + |A_y|^2 = 1$ and the phase of the x component is taken to be $\varphi_x = 0$.

Table 6.1-1 Jones vectors of linearly polarized (LP) and right- and left-circularly polarized (RCP, LCP) light.

LP in x direction	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$		LP at angle θ	$\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$	
RCP	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ j \end{bmatrix}$		LCP	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -j \end{bmatrix}$	

Orthogonal Polarizations

Two polarization states represented by the Jones vectors $\mathbf{J}_1$ and $\mathbf{J}_2$ are said to be orthogonal if the inner product between $\mathbf{J}_1$ and $\mathbf{J}_2$ is zero. The inner product is defined by

$$(\mathbf{J}_1, \mathbf{J}_2) = A_{1x}A_{2x}^* + A_{1y}A_{2y}^*, \tag{6.1-11}$$

where A_{1x} and A_{1y} are the elements of $\mathbf{J}_1$ and A_{2x} and A_{2y} are the elements of $\mathbf{J}_2$. An example of orthogonal Jones vectors are the linearly polarized waves in the x and

y directions, or any other pair of orthogonal directions. Another example is provided by right and left circularly polarized waves.

Expansion of Arbitrary Polarization as a Superposition of Two Orthogonal Polarizations

An arbitrary Jones vector $\mathbf{J}$ can always be analyzed as a weighted superposition of two orthogonal Jones vectors, say $\mathbf{J}_1$ and $\mathbf{J}_2$, that form the expansion basis; thus $\mathbf{J} = \alpha_1\mathbf{J}_1 + \alpha_2\mathbf{J}_2$. If $\mathbf{J}_1$ and $\mathbf{J}_2$ are normalized such that $(\mathbf{J}_1, \mathbf{J}_1) = (\mathbf{J}_2, \mathbf{J}_2) = 1$, the expansion coefficients are the inner products $\alpha_1 = (\mathbf{J}, \mathbf{J}_1)$ and $\alpha_2 = (\mathbf{J}, \mathbf{J}_2)$.

EXAMPLE 6.1-1. ***Expansions in Linearly Polarized and Circularly Polarized Bases.*** Using the x and y linearly polarized vectors $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ as an expansion basis, the expansion coefficients for a Jones vector of components A_x and A_y with $|A_x|^2 + |A_y|^2 = 1$ are, by definition, $\alpha_1 = A_x$ and $\alpha_2 = A_y$. The same polarization state may be expanded in other bases.

- In a basis of linearly polarized vectors at angles $45°$ and $135°$, i.e., $\mathbf{J}_1 = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ and $\mathbf{J}_2 = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix}-1\\1\end{smallmatrix}\right]$, the expansion coefficients α_1 and α_2 are:

$$A_{45} = \frac{1}{\sqrt{2}}(A_x + A_y), \quad A_{135} = \frac{1}{\sqrt{2}}(A_y - A_x). \tag{6.1-12}$$

- Similarly, if the right and left circularly polarized waves $\frac{1}{\sqrt{2}}\left[\begin{smallmatrix}1\\j\end{smallmatrix}\right]$ and $\frac{1}{\sqrt{2}}\left[\begin{smallmatrix}1\\-j\end{smallmatrix}\right]$ are used as an expansion basis, the coefficients α_1 and α_2 are:

$$A_R = \frac{1}{\sqrt{2}}(A_x - jA_y), \quad A_L = \frac{1}{\sqrt{2}}(A_x + jA_y). \tag{6.1-13}$$

  For example, a linearly polarized wave with a plane of polarization that makes an angle θ with the x axis (i.e., $A_x = \cos\theta$ and $A_y = \sin\theta$) is equivalent to a superposition of right and left circularly polarized waves with coefficients $\frac{1}{\sqrt{2}}\,e^{-j\theta}$ and $\frac{1}{\sqrt{2}}\,e^{j\theta}$, respectively. A linearly polarized wave therefore equals a weighted sum of right and left circularly polarized waves.

EXERCISE 6.1-1

Measurement of the Stokes Parameters. Show that the Stokes parameters defined in (6.1-9) for light with Jones vector components A_x and A_y are given by

$$\mathrm{S}_0 = |A_x|^2 + |A_y|^2 \tag{6.1-14a}$$

$$\mathrm{S}_1 = |A_x|^2 - |A_y|^2 \tag{6.1-14b}$$

$$\mathrm{S}_2 = |A_{45}|^2 - |A_{135}|^2 \tag{6.1-14c}$$

$$\mathrm{S}_3 = |A_R|^2 - |A_L|^2, \tag{6.1-14d}$$

where A_{45} and A_{135} are the coefficients of expansion in a basis of linearly polarized vectors at angles $45°$ and $135°$ as in (6.1-12), and A_R and A_L are the coefficients of expansion in a basis of the right and left circularly polarized waves set forth in (6.1-13). Suggest a method of measuring the Stokes parameters for light of arbitrary polarization.

Matrix Representation of Polarization Devices

Consider the transmission of a plane wave of arbitrary polarization through an optical system that maintains the plane-wave nature of the wave, but alters its polarization, as illustrated schematically in Fig. 6.1-6. The system is assumed to be linear, so that the principle of superposition of optical fields is obeyed. Two examples of such systems are the reflection of light from a planar boundary between two media, and the transmission of light through a plate with anisotropic optical properties.

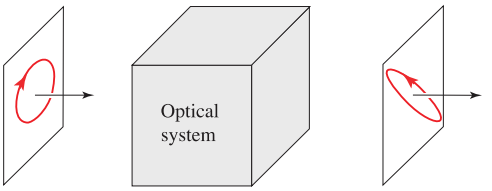


Figure 6.1-6 An optical system that alters the polarization of a plane wave.

The complex envelopes of the two electric-field components of the input (incident) wave, A_{1x} and A_{1y}, and those of the output (transmitted or reflected) wave, A_{2x} and A_{2y}, are in general related by the weighted superpositions

$$\begin{aligned} A_{2x} &= T_{11}A_{1x} + T_{12}A_{1y} \\ A_{2y} &= T_{21}A_{1x} + T_{22}A_{1y}, \end{aligned} \tag{6.1-15}$$

where T_{11}, T_{12}, T_{21}, and T_{22} are constants describing the device. Equations (6.1-15) are general relations that all linear optical polarization devices must satisfy.

The linear relations in (6.1-15) may conveniently be written in matrix notation by defining a 2×2 matrix $\mathbf{T}$ with elements T_{11}, T_{12}, T_{21}, and T_{22} so that

$$\begin{bmatrix} A_{2x} \\ A_{2y} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} A_{1x} \\ A_{1y} \end{bmatrix}. \tag{6.1-16}$$

If the input and output waves are described by the Jones vectors $\mathbf{J}_1$ and $\mathbf{J}_2$, respectively, then (6.1-16) may be written in the compact matrix form

$$\mathbf{J}_2 = \mathbf{T}\mathbf{J}_1. \tag{6.1-17}$$

The matrix $\mathbf{T}$, called the **Jones matrix**, describes the optical system, whereas the vectors $\mathbf{J}_1$ and $\mathbf{J}_2$ describe the input and output waves.

The structure of the Jones matrix $\mathbf{T}$ of a given optical system determines its effect on the polarization state and intensity of the wave. The following is a compilation of the Jones matrices of some systems with simple characteristics. Physical devices that have such characteristics will be discussed subsequently in this chapter.

Linear polarizers. The system represented by the Jones matrix

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{6.1-18}$$

Linear Polarizer Along x Direction

transforms a wave of components (A_{1x}, A_{1y}) into a wave of components $(A_{1x}, 0)$ by eliminating the y component, thereby yielding a wave polarized along the x direction, as illustrated in Fig. 6.1-7. The system is a **linear polarizer** with its transmission axis pointing in the x direction.

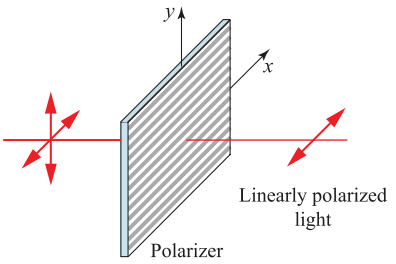


Figure 6.1-7 The linear polarizer. The lines in the polarizer represent the field direction that is permitted to pass.

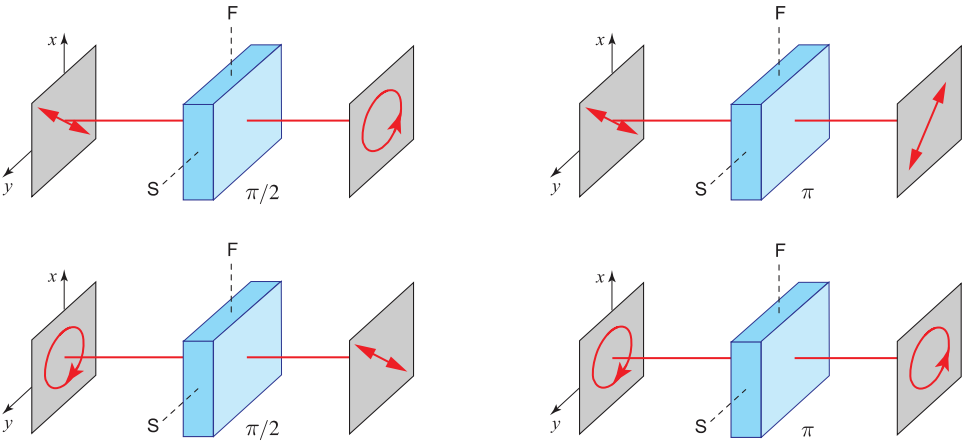
Wave retarders. The system represented by the matrix

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ 0 & e^{-j\Gamma} \end{bmatrix} \tag{6.1-19}$$

Wave-Retarder
(Fast Axis Along x Direction)

transforms a wave with field components (A_{1x}, A_{1y}) into another with components $(A_{1x}, e^{-j\Gamma}A_{1y})$, thereby delaying the y component by a phase Γ while leaving the x component unchanged. It is therefore called a **wave retarder**. The x and y axes are called the fast and slow axes of the retarder, respectively.

The simple application of matrix algebra permits the results illustrated in Fig. 6.1-8 to be understood:



(*a*) Quarter-wave retarder (*b*) Half-wave retarder

Figure 6.1-8 Operations of quarter-wave $(\pi/2)$ and half-wave (π) retarders on several particular states of polarization are shown in (a) and (b), respectively. F and S represent the fast and slow axes of the retarder, respectively.

- When $\Gamma = \pi/2$, the retarder (called a **quarter-wave retarder** and described by the Jones matrix $\begin{bmatrix} 1 & 0 \\ 0 & -j \end{bmatrix}$) converts the linearly polarized wave $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ into the left

circularly polarized wave $\begin{bmatrix} 1 \\ -j \end{bmatrix}$, and converts the right circularly polarized wave $\begin{bmatrix} 1 \\ j \end{bmatrix}$ into the linearly polarized wave $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

- When $\Gamma = \pi$, the retarder (called a **half-wave retarder** and described by the Jones matrix $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$) converts the linearly polarized wave $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ into the linearly polarized wave $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, thereby rotating the plane of polarization by $90°$. The half-wave retarder converts the right circularly polarized wave $\begin{bmatrix} 1 \\ j \end{bmatrix}$ into the left circularly polarized wave $\begin{bmatrix} 1 \\ -j \end{bmatrix}$.

Polarization rotators. While a wave retarder can transform a wave with one form of polarization into another, a **polarization rotator** always maintains the linear polarization of a wave but rotates the plane of polarization by a particular angle. The Jones matrix

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{6.1-20}$$

Polarization Rotator

represents a device that converts a linearly polarized wave $\begin{bmatrix} \cos\theta_1 \\ \sin\theta_1 \end{bmatrix}$ into another linearly polarized wave $\begin{bmatrix} \cos\theta_2 \\ \sin\theta_2 \end{bmatrix}$, where $\theta_2 = \theta_1 + \theta$. It therefore rotates the plane of polarization of a linearly polarized wave by an angle θ.

Cascaded Polarization Devices

The action of cascaded optical systems on polarized light may be conveniently determined by using conventional matrix multiplication formulas. A system characterized by the Jones matrix $\mathbf{T}_1$ followed by another characterized by $\mathbf{T}_2$ are equivalent to a single system characterized by the product matrix $\mathbf{T} = \mathbf{T}_2\mathbf{T}_1$. The matrix of the system through which light is first transmitted must stand to the right in the matrix product since it is the first to affect the input Jones vector.

EXERCISE 6.1-2

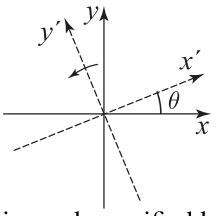
Cascaded Wave Retarders. Show that two cascaded quarter-wave retarders with parallel fast axes are equivalent to a half-wave retarder. What is the result if the fast axes are orthogonal?

Coordinate Transformation

The elements of the Jones vectors and Jones matrices are dependent on the choice of the coordinate system. However, if these elements are known in one coordinate system, they can be determined in another coordinate system by using matrix methods. If $\mathbf{J}$ is the Jones vector in the x–y coordinate system, then in a new coordinate system x'–y', with the x' direction making an angle θ with the x direction, the Jones vector $\mathbf{J}'$ is given by

$$\mathbf{J}' = \mathbf{R}(\theta)\,\mathbf{J}, \tag{6.1-21}$$

where $\mathbf{R}(\theta)$ is the coordinate-transformation matrix



$$\mathbf{R}(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{6.1-22}$$

(6.1-22) Coordinate Transformation

This can be verified by relating the components of the electric field in the two coordinate systems.

The Jones matrix $\mathbf{T}$, which represents an optical system, is similarly transformed into $\mathbf{T}'$, in accordance with the matrix relations

$$\mathbf{T}' = \mathbf{R}(\theta)\,\mathbf{T}\,\mathbf{R}(-\theta) \tag{6.1-23}$$

$$\mathbf{T} = \mathbf{R}(-\theta)\,\mathbf{T}'\,\mathbf{R}(\theta), \tag{6.1-24}$$

where $\mathbf{R}(-\theta)$ is given by (6.1-22) with $-\theta$ replacing θ. The matrix $\mathbf{R}(-\theta)$ is the inverse of $\mathbf{R}(\theta)$, so that $\mathbf{R}(-\theta)\,\mathbf{R}(\theta)$ is a unit matrix. Equation (6.1-23) can be obtained by using the relation $\mathbf{J}_2 = \mathbf{T}\mathbf{J}_1$ and the transformation $\mathbf{J}_2' = \mathbf{R}(\theta)\,\mathbf{J}_2 = \mathbf{R}(\theta)\,\mathbf{T}\,\mathbf{J}_1$. Since $\mathbf{J}_1 = \mathbf{R}(-\theta)\,\mathbf{J}_1'$, we have $\mathbf{J}_2' = \mathbf{R}(\theta)\,\mathbf{T}\,\mathbf{R}(-\theta)\,\mathbf{J}_1'$; since $\mathbf{J}_2' = \mathbf{T}'\,\mathbf{J}_1'$, (6.1-23) follows.

EXERCISE 6.1-3

Jones Matrix of a Rotated Half-Wave Retarder. Show that the Jones matrix of a half-wave retarder whose fast axis makes an angle θ with the x axis is

$$\mathbf{T} = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}. \tag{6.1-25}$$

(6.1-25) Half-Wave Retarder at Angle θ

Derive (6.1-25) using (6.1-19), (6.1-22), and (6.1-24). Demonstrate that if $\theta = 22.5°$, the output waves are proportional to the sum and difference of the input waves.

Normal Modes

The normal modes of a polarization system are the states of polarization that are not changed when the wave is transmitted through the system (see Appendix C). These states have Jones vectors satisfying

$$\mathbf{T}\mathbf{J} = \mu\mathbf{J}, \tag{6.1-26}$$

where μ is constant. The normal modes are therefore the eigenvectors of the Jones matrix $\mathbf{T}$, and the values of μ are the corresponding eigenvalues. Since the matrix $\mathbf{T}$ is of size 2×2 there are only two independent normal modes, $\mathbf{T}\mathbf{J}_1 = \mu_1\mathbf{J}_1$ and $\mathbf{T}\mathbf{J}_2 = \mu_2\mathbf{J}_2$. If the matrix $\mathbf{T}$ is a Hermitian, i.e., if $T_{12} = T_{21}^*$, the normal modes are orthogonal: $(\mathbf{J}_1, \mathbf{J}_2) = 0$. The normal modes are usually used as an expansion basis, so that an arbitrary input wave $\mathbf{J}$ may be expanded as a superposition of normal modes: $\mathbf{J} = \alpha_1\mathbf{J}_1 + \alpha_2\mathbf{J}_2$. The response of the system may then be easily evaluated since $\mathbf{T}\mathbf{J} = \mathbf{T}(\alpha_1\mathbf{J}_1 + \alpha_2\mathbf{J}_2) = \alpha_1\mathbf{T}\mathbf{J}_1 + \alpha_2\mathbf{T}\mathbf{J}_2 = \alpha_1\mu_1\mathbf{J}_1 + \alpha_2\mu_2\mathbf{J}_2$ (see Appendix C).

EXERCISE 6.1-4

Normal Modes of Simple Polarization Systems.

(a) Show that the normal modes of the linear polarizer are linearly polarized waves.

(b) Show that the normal modes of the wave retarder are linearly polarized waves.

(c) Show that the normal modes of the polarization rotator are right and left circularly polarized waves.

What are the eigenvalues of the systems described above?

6.2 REFLECTION AND REFRACTION

In this section we examine the reflection and refraction of a monochromatic plane wave of arbitrary polarization incident at a planar boundary between two dielectric media. The media are assumed to be linear, homogeneous, and isotropic with impedances η_1 and η_2, and refractive indices n_1 and n_2. The incident, refracted, and reflected waves are labeled with the subscripts 1, 2, and 3, respectively, as illustrated in Fig. 6.2-1.

As shown in Sec. 2.4A, the wavefronts of these waves are matched at the boundary if the angles of reflection and incidence are equal, $\theta_3 = \theta_1$, and if the angles of refraction and incidence satisfy Snell's law,

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \tag{6.2-1}$$

To relate the amplitudes and polarizations of the three waves, we associate with each wave an x–y coordinate system in a plane normal to the direction of propagation (Fig. 6.2-1). The electric-field envelopes of these waves are described by the Jones vectors

$$\mathbf{J}_1 = \begin{bmatrix} A_{1x} \\ A_{1y} \end{bmatrix}, \qquad \mathbf{J}_2 = \begin{bmatrix} A_{2x} \\ A_{2y} \end{bmatrix}, \qquad \mathbf{J}_3 = \begin{bmatrix} A_{3x} \\ A_{3y} \end{bmatrix}. \tag{6.2-2}$$

We proceed to determine the relations between $\mathbf{J}_2$ and $\mathbf{J}_1$ and between $\mathbf{J}_3$ and $\mathbf{J}_1$. These relations are written in the form of matrices $\mathbf{J}_2 = \mathbf{t}\mathbf{J}_1$ and $\mathbf{J}_3 = \mathbf{r}\mathbf{J}_1$, where $\mathbf{t}$ and $\mathbf{r}$ are 2×2 Jones matrices describing the transmission and reflection of the wave, respectively.

The elements of the transmission and reflection matrices may be determined by imposing the boundary conditions required by electromagnetic theory, namely the continuity at the boundary of the tangential components of $\mathbf{E}$ and $\mathbf{H}$ and the normal components of $\mathbf{D}$ and $\mathbf{B}$. The electric field associated with each wave is orthogonal to the magnetic field; the ratio of their envelopes is the characteristic impedance, which is η_1 for the incident and reflected waves and η_2 for the transmitted wave. The result is a set of equations that are solved to obtain relations between the components of the electric fields of the three waves.

The algebra involved is reduced substantially if we observe that the two normal modes for this system are linearly polarized waves with polarizations along the x and y directions. This may be proved if we show that an incident, a reflected, and a refracted wave with their electric-field vectors pointing in the x direction are self-consistent with the boundary conditions, and similarly for three waves linearly polarized in the y direction. This is indeed the case. The x and y polarized waves are therefore uncoupled.

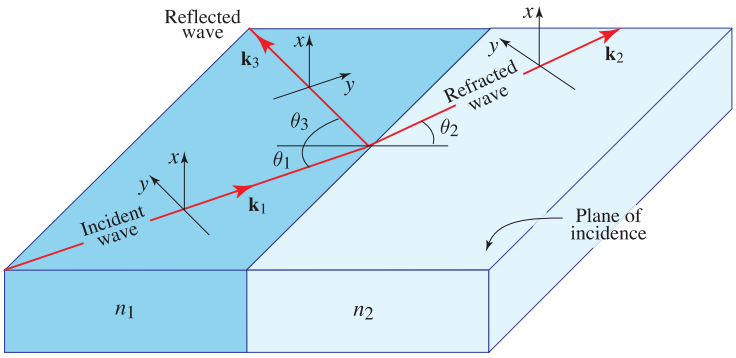


Figure 6.2-1 Reflection and refraction at the boundary between two dielectric media.

The x-polarized mode is called the **transverse electric (TE)** polarization or the **orthogonal polarization**, since the electric fields are orthogonal to the plane of incidence. The y-polarized mode is called the **transverse magnetic (TM)** polarization since the magnetic field is orthogonal to the plane of incidence, or the **parallel polarization** since the electric fields are parallel to the plane of incidence. The orthogonal and parallel polarizations are also called the s (for the German *senkrecht*, meaning "perpendicular") and p (for "parallel") polarizations, respectively. The y axes in Fig. 6.2-1 are arbitrarily defined such that their components parallel to the boundary between the dielectrics all point in the same direction.

The independence of the x and y polarizations implies that the Jones matrices $\mathbf{t}$ and $\mathbf{r}$ are diagonal,

$$\mathbf{t} = \begin{bmatrix} \mathrm{t}_x & 0 \\ 0 & \mathrm{t}_y \end{bmatrix}, \qquad \mathbf{r} = \begin{bmatrix} \mathrm{r}_x & 0 \\ 0 & \mathrm{r}_y \end{bmatrix} \tag{6.2-3}$$

so that

$$E_{2x} = \mathrm{t}_x E_{1x}, \qquad E_{2y} = \mathrm{t}_y E_{1y} \tag{6.2-4}$$

$$E_{3x} = \mathrm{r}_x E_{1x}, \qquad E_{3y} = \mathrm{r}_y E_{1y}. \tag{6.2-5}$$

The coefficients t_x and t_y are the complex amplitude transmittances for the TE and TM polarizations, respectively; r_x and r_y are the analogous complex amplitude reflectances.

Applying the boundary conditions (i.e., equating the tangential components of the electric fields and the tangential components of the magnetic fields at both sides of the boundary) in each of the TE and TM cases, we obtain the following expressions for the complex-amplitude reflectances and transmittances:

$$\mathrm{r}_x = \frac{\eta_2 \sec\theta_2 - \eta_1 \sec\theta_1}{\eta_2 \sec\theta_2 + \eta_1 \sec\theta_1}, \qquad \mathrm{t}_x = 1 + \mathrm{r}_x, \tag{6.2-6}$$

TE Polarization

$$\mathrm{r}_y = \frac{\eta_2 \cos\theta_2 - \eta_1 \cos\theta_1}{\eta_2 \cos\theta_2 + \eta_1 \cos\theta_1}, \qquad \mathrm{t}_y = (1 + \mathrm{r}_y)\frac{\cos\theta_1}{\cos\theta_2}. \tag{6.2-7}$$

TM Polarization

Reflection & Transmission

The characteristic impedance $\eta = \sqrt{\mu/\epsilon}$ is complex if ϵ and/or μ are complex, as is the case for lossy or conductive media. For nonlossy, nonmagnetic, dielectric media, $\eta = \eta_o/n$ is real, where $\eta_o = \sqrt{\mu_o/\epsilon_o}$ and n is the refractive index. In this case,

the reflection and transmission coefficients in (6.2-6) and (6.2-7) yield the **Fresnel equations**:

$$\mathrm{r}_x = \frac{n_1 \cos\theta_1 - n_2 \cos\theta_2}{n_1 \cos\theta_1 + n_2 \cos\theta_2}, \qquad \mathrm{t}_x = 1 + \mathrm{r}_x, \tag{6.2-8}$$

TE Polarization

$$\mathrm{r}_y = \frac{n_1 \sec\theta_1 - n_2 \sec\theta_2}{n_1 \sec\theta_1 + n_2 \sec\theta_2}, \qquad \mathrm{t}_y = (1 + \mathrm{r}_y)\frac{\cos\theta_1}{\cos\theta_2}. \tag{6.2-9}$$

TM Polarization

Fresnel Equations

Given n_1, n_2, and θ_1, the reflection coefficients can be determined from the Fresnel equations by first determining θ_2 using Snell's law, (6.2-1), from which

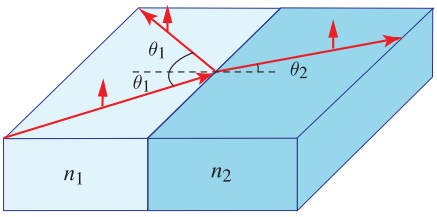
$$\cos\theta_2 = \sqrt{1 - \sin^2\theta_2} = \sqrt{1 - (n_1/n_2)^2 \sin^2\theta_1}\,. \tag{6.2-10}$$

Since the quantities under the square-root signs in (6.2-10) can be negative, the reflection and transmission coefficients are in general complex. The magnitudes $|\mathrm{r}_x|$ and $|\mathrm{r}_y|$, and the phase shifts $\varphi_x = \arg\{\mathrm{r}_x\}$ and $\varphi_y = \arg\{\mathrm{r}_y\}$, are plotted in Figs. 6.2-2 to 6.2-5 for the two polarizations, as functions of the angle of incidence θ_1. Plots are provided for external reflection $(n_1 < n_2)$ as well as for internal reflection $(n_1 > n_2)$.

TE Polarization

The dependence of the reflection coefficient r_x on θ_1 for the TE-polarized wave is given by (6.2-8):

External reflection $(n_1 < n_2)$**.** The reflection coefficient r_x is always real and negative, corresponding to a phase shift $\varphi_x = \pi$. The magnitude $|\mathrm{r}_x| = (n_2 - n_1)/(n_1 + n_2)$ at $\theta_1 = 0$ (normal incidence) and increases to unity at $\theta_1 = 90°$ (grazing incidence), as shown in Fig. 6.2-2.



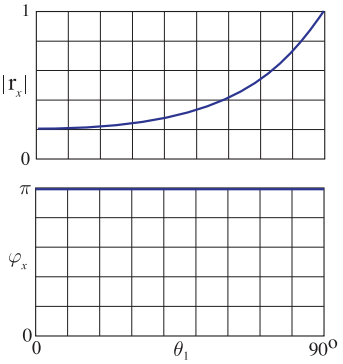


Figure 6.2-2 Magnitude and phase of the reflection coefficient as a function of the angle of incidence for *external reflection* of the *TE-polarized* wave $(n_2/n_1 = 1.5)$.

Internal reflection $(n_1 > n_2)$**.** For small θ_1 the reflection coefficient is real and positive. Its magnitude is $(n_1 - n_2)/(n_1 + n_2)$ when $\theta_1 = 0°$, and increases gradually to a value of unity, which is attained when θ_1 equals the critical angle $\theta_c = \sin^{-1}(n_2/n_1)$. For $\theta_1 > \theta_c$, the magnitude of r_x remains at unity, which corresponds to total internal reflection. This may be shown by using (6.2-10) to write[†] $\cos\theta_2 = -\sqrt{1 - \sin^2\theta_1/\sin^2\theta_c} = -j\sqrt{\sin^2\theta_1/\sin^2\theta_c - 1}$, and substituting into (6.2-8). Total internal reflection is accompanied by a phase shift $\varphi_x = \arg\{\mathrm{r}_x\}$ given by

[†] The choice of the minus sign for the square root is consistent with the derivation that leads to the Fresnel equation.

$$\tan\frac{\varphi_x}{2} = \sqrt{\frac{\cos^2\theta_c}{\cos^2\theta_1} - 1} \tag{6.2-11}$$

TE-Reflection Phase Shift

The phase shift φ_x increases from 0 at $\theta_1 = \theta_c$ to π at $\theta_1 = 90°$, as illustrated in Fig. 6.2-3. This phase plays an important role in dielectric waveguides (see Sec. 9.2). An evanescent wave is created in the vicinity of the boundary when total internal reflection occurs.

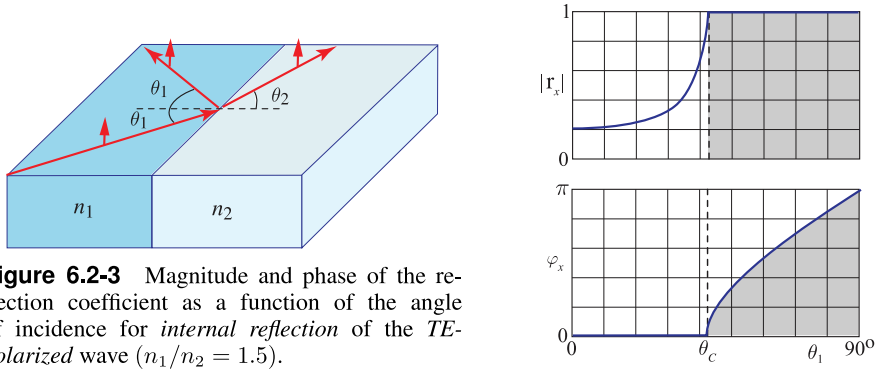


Figure 6.2-3 Magnitude and phase of the reflection coefficient as a function of the angle of incidence for *internal reflection* of the *TE-polarized* wave ($n_1/n_2 = 1.5$).

TM Polarization

Similarly, the dependence of the reflection coefficient r_y on θ_1 for the TM-polarized wave is provided by (6.2-9):

External reflection $(n_1 < n_2)$**.** The reflection coefficient r_y is always real, as shown in Fig. 6.2-4. It assumes a negative value of $(n_1 - n_2)/(n_1 + n_2)$ at $\theta_1 = 0$ (normal incidence). Its magnitude then decreases until it vanishes when $n_1 \sec\theta_1 = n_2 \sec\theta_2$, at an angle $\theta_1 = \theta_\mathrm{B}$, known as the **Brewster angle**:

$$\theta_\mathrm{B} = \tan^{-1}(n_2/n_1) \tag{6.2-12}$$

Brewster Angle

(see Prob. 6.2-4 for other properties of the Brewster angle). For $\theta_1 > \theta_\mathrm{B}$, r_y reverses sign (φ_y goes from π to 0) and its magnitude gradually increases until it approaches unity at $\theta_1 = 90°$. The absence of reflection of the TM wave at the Brewster angle is useful for making polarizers (see Sec. 6.6).

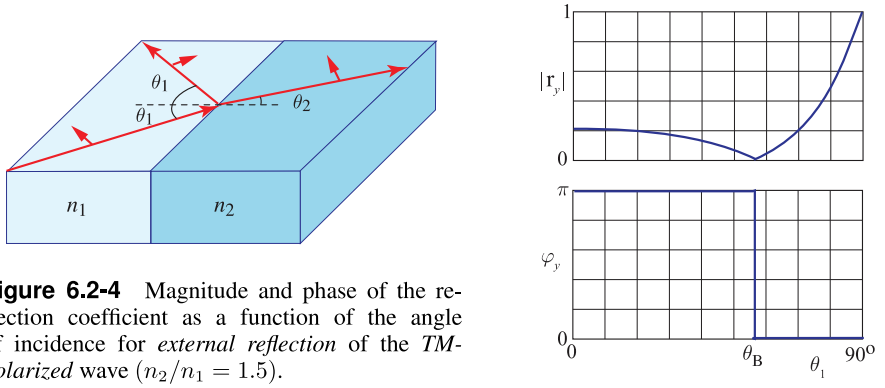


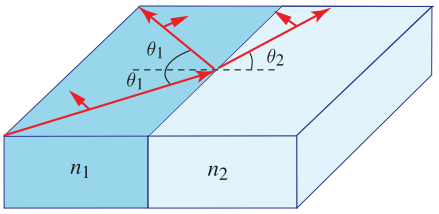
Figure 6.2-4 Magnitude and phase of the reflection coefficient as a function of the angle of incidence for *external reflection* of the *TM-polarized* wave ($n_2/n_1 = 1.5$).

Internal reflection $(n_1 > n_2)$**.** At $\theta_1 = 0°$, the reflection coefficient r_y is positive and has magnitude $(n_1 - n_2)/(n_1 + n_2)$, as illustrated in Fig. 6.2-5. As θ_1 increases, the magnitude decreases until it vanishes at the Brewster angle $\theta_\mathrm{B} = \tan^{-1}(n_2/n_1)$. As θ_1 increases beyond θ_B, r_y becomes negative and its magnitude increases until it reaches unity at the critical angle θ_c. For $\theta_1 > \theta_c$ the wave undergoes total internal reflection accompanied by a phase shift $\varphi_y = \arg\{\mathrm{r}_y\}$ given by

$$\tan\frac{\varphi_y}{2} = \frac{-1}{\sin^2\theta_c}\sqrt{\frac{\cos^2\theta_c}{\cos^2\theta_1} - 1}\,. \tag{6.2-13}$$

TM-Reflection Phase Shift

At normal incidence, evidently, the reflection coefficient is $\mathrm{r} = (n_1 - n_2)/(n_1 + n_2)$, whether the reflection is TE or TM, or external or internal.



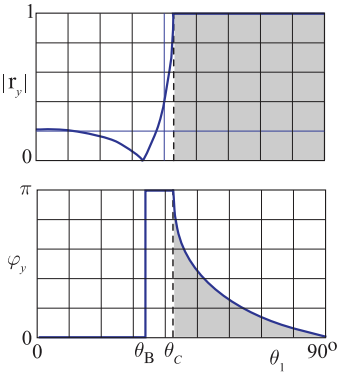


Figure 6.2-5 Magnitude and phase of the reflection coefficient as a function of the angle of incidence for *internal reflection* of the *TM-polarized wave* $(n_1/n_2 = 1.5)$.

EXERCISE 6.2-1

Brewster Windows. At what angle is a TM-polarized beam of light transmitted through a glass plate of refractive index $n = 1.5$ placed in air $(n = 1)$ without suffering reflection losses at either surface? Such plates, known as Brewster windows (Fig. 6.2-6), are used in lasers, as described in Sec. 16.2D.

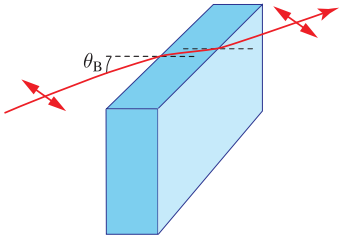


Figure 6.2-6 The Brewster window transmits TM-polarized light with no reflection loss.

Power Reflectance and Transmittance

The reflection and transmission coefficients r and t represent ratios of complex amplitudes. The power reflectance $\mathcal{R}$ and power transmittance $\mathcal{T}$ are defined as the ratios of power flow (along a direction normal to the boundary) of the reflected and transmitted waves to that of the incident wave. Because the reflected and incident waves propagate in the same medium and make the same angle with the normal to the surface, it follows that

$$\mathcal{R} = |\mathrm{r}|^2. \tag{6.2-14}$$

For both TE and TM polarizations, and for both external and internal reflection, the power reflectance at normal incidence is therefore

$$\boxed{\mathcal{R} = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2.} \tag{6.2-15}$$

Power Reflectance at Normal Incidence

At the boundary between glass $(n = 1.5)$ and air $(n = 1)$, for example, $\mathcal{R} = 0.04$, so that 4% of the light is reflected at normal incidence. At the boundary between GaAs $(n = 3.6)$ and air $(n = 1)$, $\mathcal{R} \approx 0.32$, so that 32% of the light is reflected at normal incidence. However, at oblique angles the reflectance can be much greater or much smaller than 32%, as illustrated in Fig. 6.2-7.

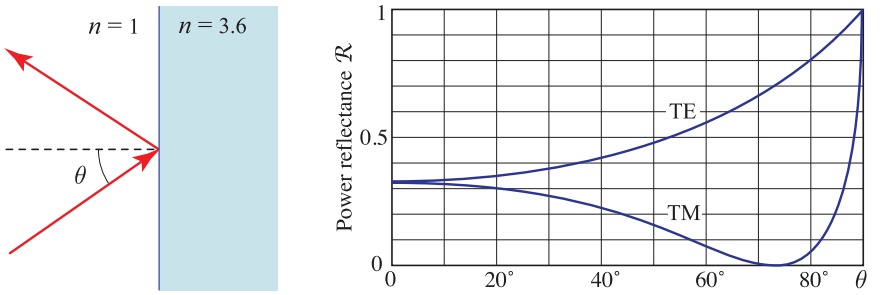


Figure 6.2-7 Power reflectance of TE- and TM-polarization plane waves at the boundary between air $(n = 1)$ and GaAs $(n = 3.6)$, as a function of the angle of incidence θ.

The power transmittance $\mathcal{T}$ is determined by invoking the conservation of power, so that in the absence of absorption loss the transmittance is simply

$$\mathcal{T} = 1 - \mathcal{R}. \tag{6.2-16}$$

It is important to note, however, that $\mathcal{T}$ is generally *not* equal to $|\mathrm{t}|^2$ since the power travels at different angles and with different impedances in the two media. For a wave traveling at an angle θ in a medium of refractive index n, the power flow in the direction normal to the boundary is $(|\mathcal{E}|^2/2\eta)\cos\theta = (|\mathcal{E}|^2/2\eta_o)\, n \cos\theta$. It follows that

$$\mathcal{T} = \frac{n_2 \cos\theta_2}{n_1 \cos\theta_1} |\mathrm{t}|^2. \tag{6.2-17}$$

Reflectance from a plate. The power reflectance at normal incidence from a plate with two surfaces is described by $\mathcal{R}(1 + \mathcal{T}^2)$ since the power reflected from the far surface involves a double transmission through the near surface. For a glass plate in air, the overall reflectance is $\mathcal{R}(1 + \mathcal{T}^2) = 0.04[1 + (0.96)^2] \approx 0.077$, so that about 7.7% of the incident light power is reflected. However, this calculation does not include interference effects, which are washed out when the light is incoherent (see Sec.12.2), nor does it account for multiple reflections inside the plate. Optical transmission and reflectance from multiple boundaries in layered media are described in detail in Sec. 7.1.

6.3 OPTICS OF ANISOTROPIC MEDIA

A dielectric medium is said to be anisotropic if its macroscopic optical properties depend on direction. The macroscopic properties of a material are, of course, ultimately governed by its microscopic properties: the shape and orientation of the individual molecules and the organization of their centers in space. Optical materials have different kinds of positional and orientational types of order, which may be described as follows (see Fig. 6.3-1):

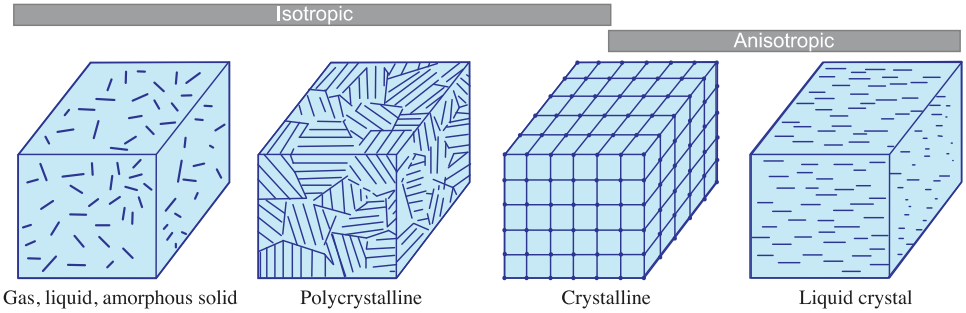


Figure 6.3-1 Positional and orientational order in different types of materials.

- If the molecules are located at totally random positions in space, and are themselves isotropic or oriented along random directions, the medium is isotropic. *Gases*, *liquids*, and *amorphous solids* follow this prescription.
- If the structure takes the form of disjointed crystalline grains that are randomly oriented with respect to each other, the material is said to be *polycrystalline*. The individual grains are, in general, anisotropic, but their averaged macroscopic behavior is isotropic.
- If the molecules are organized in space according to a regular periodic pattern and they are oriented in the same direction, as in *crystals*, the medium is, in general, anisotropic.
- If the molecules are anisotropic and their orientations are not totally random, the medium is anisotropic, even if their positions are totally random. This is the case for *liquid crystals*, which have orientational order but lack complete positional order.

A. Refractive Indices

Permittivity Tensor

In a linear anisotropic dielectric medium (a crystal, for example), each component of the electric flux density $\mathbf{D}$ is a linear combination of the three components of the electric field,

$$D_i = \sum_j \epsilon_{ij} E_j. \tag{6.3-1}$$

The indices $i, j = 1, 2, 3$ refer to the x, y, and z components, respectively, as described in Sec. 5.2B. The dielectric properties of the medium are therefore characterized by a 3×3 array of nine coefficients, $\{\epsilon_{ij}\}$, that form the **electric permittivity tensor** $\boldsymbol{\epsilon}$, which is a tensor of second rank. The **material equation** (6.3-1) is usually written in the symbolic form

$$\mathbf{D} = \boldsymbol{\epsilon}\mathbf{E}. \tag{6.3-2}$$

For most dielectric media, the electric permittivity tensor is symmetric, i.e., $\epsilon_{ij} = \epsilon_{ji}$. This means that the relation between the vectors $\mathbf{D}$ and $\mathbf{E}$ is reciprocal, i.e., their ratio remains the same if their directions are exchanged. This symmetry is obeyed for dielectric nonmagnetic materials that do not exhibit optical activity, and in the absence of an external magnetic field (Sec. 6.4). With this symmetry, the medium is characterized by only six independent numbers in an arbitrary coordinate system. For crystals of certain symmetries, even fewer coefficients suffice since some vanish and some are related.

Geometrical Representation of Vectors and Tensors

A *vector*, such as the electric field $\mathbf{E}$, for example, describes a physical variable with magnitude and direction. It is represented *geometrically* by an arrow pointing in that particular direction, whose length is proportional to the magnitude of the vector [Fig. 6.3-2(a)]. A vector, which is a tensor of first rank, is represented *numerically* by three numbers: its projections on the three axes of a particular coordinate system. Though these components depend on the choice of the coordinate system, the magnitude and direction of the vector in physical space are independent of the choice of the coordinate system. A scalar, which is described by a single number, is a tensor of zero rank.

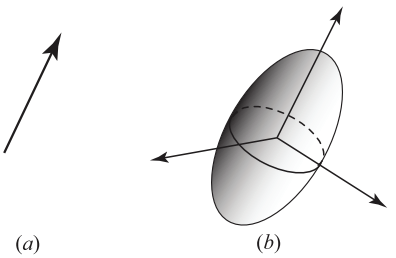


Figure 6.3-2 Geometrical representation of: (a) a vector; (b) a symmetric second-rank tensor.

A second-rank *tensor* is a rule that relates two vectors. In a given coordinate system, it is represented *numerically* by nine numbers. Changing the coordinate system yields

a different set of nine numbers, but the physical nature of the rule is unchanged. A useful *geometrical* representation [Fig. 6.3-2(*b*)] of a *symmetric* second-rank tensor (the permittivity tensor $\boldsymbol{\epsilon}$, for example), is a quadratic surface (an ellipsoid) defined by

$$\sum_{ij} \epsilon_{ij}\, x_i x_j = 1, \tag{6.3-3}$$

which is known as the **quadric representation**. This surface is invariant to the choice of the coordinate system; if the coordinate system is rotated, both x_i and ϵ_{ij} are altered but the ellipsoid remains intact in physical space. The ellipsoid has six degrees of freedom and carries all information about the symmetric second-rank tensor. In the principal coordinate system, ϵ_{ij} is diagonal and the ellipsoid assumes a particularly simple form:

$$\epsilon_1 x_1^2 + \epsilon_2 x_2^2 + \epsilon_3 x_3^2 = 1. \tag{6.3-4}$$

Its principal axes are those of the tensor, and its axes have half-lengths $1/\sqrt{\epsilon_1}$, $1/\sqrt{\epsilon_2}$, and $1/\sqrt{\epsilon_3}$.

Principal Axes and Principal Refractive Indices

The elements of the permittivity tensor depend on how the coordinate system is chosen relative to the crystal structure. However, a coordinate system can always be found for which the off-diagonal elements of ϵ_{ij} vanish, so that

$$D_1 = \epsilon_1 E_1, \qquad D_2 = \epsilon_2 E_2, \qquad D_3 = \epsilon_3 E_3, \tag{6.3-5}$$

where $\epsilon_1 = \epsilon_{11}$, $\epsilon_2 = \epsilon_{22}$, and $\epsilon_3 = \epsilon_{33}$. According to (6.3-1), $\mathbf{E}$ and $\mathbf{D}$ are parallel along these particular directions so that if, for example, $\mathbf{E}$ points in the x direction, then so too must $\mathbf{D}$. This coordinate system defines the **principal axes** and principal planes of the crystal. Throughout the remainder of this chapter, the coordinate system x, y, z, which is equivalently denoted x_1, x_2, x_3, is assumed to lie along the principal axes of the crystal. This choice simplifies all analyses without loss of generality. The permittivities ϵ_1, ϵ_2, and ϵ_3 correspond to refractive indices

$$n_1 = \sqrt{\epsilon_1/\epsilon_o}, \qquad n_2 = \sqrt{\epsilon_2/\epsilon_o}, \qquad n_3 = \sqrt{\epsilon_3/\epsilon_o}, \tag{6.3-6}$$

respectively, where ϵ_o is the permittivity of free space; these are known as the **principal refractive indices**.

Biaxial, Uniaxial, and Isotropic Crystals

Crystals in which the three principal refractive indices are different are termed **biaxial**. For crystals with certain symmetries, namely a single axis of threefold, fourfold, or sixfold symmetry, two of the refractive indices are equal ($n_1 = n_2$) and the crystal is called **uniaxial**. In this case, the indices are usually denoted $n_1 = n_2 = n_o$ and $n_3 = n_e$, which are known as the **ordinary** and **extraordinary** indices, respectively, for reasons that will become clear shortly. The crystal is said to be **positive uniaxial** if $n_e > n_o$, and **negative uniaxial** if $n_e < n_o$. The z axis of a uniaxial crystal is called the **optic axis**. In certain crystals with even greater symmetry (those with cubic unit cells, for example), all three indices are equal and the medium is optically isotropic.

Impermeability Tensor

The relation $\mathbf{D} = \boldsymbol{\epsilon}\mathbf{E}$ can be inverted and written in the form $\mathbf{E} = \boldsymbol{\epsilon}^{-1}\mathbf{D}$, where $\boldsymbol{\epsilon}^{-1}$ is the inverse of the tensor $\boldsymbol{\epsilon}$. It is also useful to define the **electric impermeability tensor** $\boldsymbol{\eta} = \epsilon_o\boldsymbol{\epsilon}^{-1}$ (not to be confused with the impedance of the medium η), so that $\epsilon_o\mathbf{E} = \boldsymbol{\eta}\mathbf{D}$. Since $\boldsymbol{\epsilon}$ is symmetric, so too is $\boldsymbol{\eta}$. Both tensors, $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$, share the same principal axes. In the principal coordinate system, $\boldsymbol{\eta}$ is diagonal with principal values $\epsilon_o/\epsilon_1 = 1/n_1^2$, $\epsilon_o/\epsilon_2 = 1/n_2^2$, and $\epsilon_o/\epsilon_3 = 1/n_3^2$. Either tensor, $\boldsymbol{\epsilon}$ or $\boldsymbol{\eta}$, fully describes the optical properties of the crystal.

Index Ellipsoid

The **index ellipsoid** (also called the **optical indicatrix**) is the quadric representation of the electric impermeability tensor $\boldsymbol{\eta} = \epsilon_o\boldsymbol{\epsilon}^{-1}$:

$$\sum_{ij} \eta_{ij}\, x_i x_j = 1, \qquad i, j = 1, 2, 3. \tag{6.3-7}$$

If the principal axes were to be used as the coordinate system, we would obtain

$$\boxed{\frac{x_1^2}{n_1^2} + \frac{x_2^2}{n_2^2} + \frac{x_3^2}{n_3^2} = 1\,,} \tag{6.3-8}$$

Index Ellipsoid

with principal values $1/n_1^2$, $1/n_2^2$, and $1/n_3^2$, and axes of half-lengths n_1, n_2, and n_3.

The optical properties of the crystal (the directions of the principal axes and the values of the principal refractive indices) are therefore completely described by the index ellipsoid (Fig. 6.3-3). For a uniaxial crystal, the index ellipsoid reduces to an ellipsoid of revolution; for an isotropic medium it becomes a sphere.

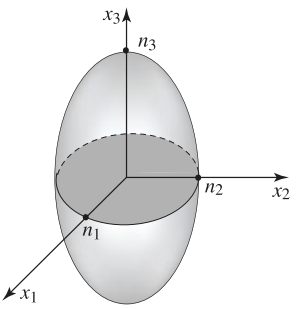


Figure 6.3-3 The index ellipsoid. The coordinates (x_1, x_2, x_3) are the principal axes while (n_1, n_2, n_3) are the principal refractive indices of the crystal.

B. Propagation Along a Principal Axis

The rules that govern the propagation of light in crystals under general conditions are rather complex. However, they become relatively simple if the light is a plane wave traveling along one of the principal axes of the crystal. We begin with this case.

Normal Modes

Let x–y–z be a coordinate system that coincides with the principal axes of a crystal. A plane wave traveling in the z direction and linearly polarized along the x direction [Fig. 6.3-4(a)] travels with phase velocity c_o/n_1 (wavenumber $k = n_1 k_o$) without changing its polarization. The reason for this is that the electric field has only one

component, E_1 pointed along the x direction, so that $\mathbf{D}$ is also in the x direction with $D_1 = \epsilon_1 E_1$; the wave equation derived from Maxwell's equations therefore provides a velocity of light given by $1/\sqrt{\mu_o \epsilon_1} = c_o/n_1$. Similarly, a plane wave traveling in the z direction and linearly polarized along the y direction [Fig. 6.3-4(b)] travels with phase velocity c_o/n_2, thereby experiencing a refractive index n_2. Thus, the normal modes for propagation in the z direction are linearly polarized waves in the x and y directions. These waves are said to be normal modes because their velocities and polarizations are maintained as they propagate (see Appendix C). Other cases in which the wave propagates along one of the principal axes and is linearly polarized along another are treated similarly [Fig. 6.3-4(c)].

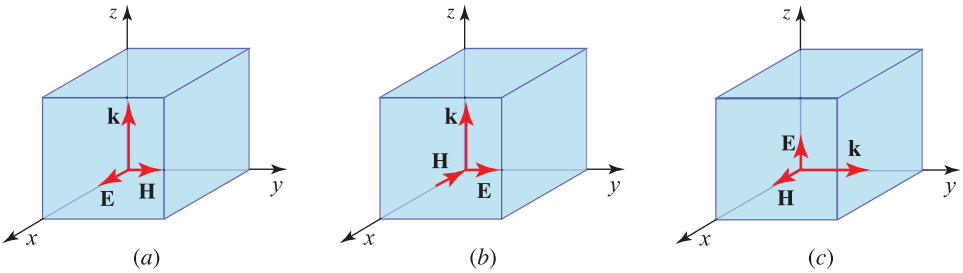


Figure 6.3-4 A wave traveling along a principal axis and polarized along another principal axis has phase velocity c_o/n_1, c_o/n_2, or c_o/n_3, when the electric-field vector points in the x, y, or z directions, respectively. (a) $k = n_1 k_o$; (b) $k = n_2 k_o$; (c) $k = n_3 k_o$.

Polarization Along an Arbitrary Direction

We now consider a wave traveling along one principal axis (the z axis, for example) that is linearly polarized along an arbitrary direction in the x–y plane. This case is addressed by analyzing the wave as a sum of the normal modes, namely the linearly polarized waves in the x and y directions. These two components travel with different phase velocities, c_o/n_1 and c_o/n_2, respectively. They therefore undergo different phase shifts, $\varphi_x = n_1 k_o d$ and $\varphi_y = n_2 k_o d$, respectively, after propagating a distance d. Their phase retardation is thus $\varphi = \varphi_y - \varphi_x = (n_2 - n_1) k_o d$. Recombination of the two components yields an elliptically polarized wave, as explained in Sec. 6.1 and illustrated in Fig. 6.3-5. Such a crystal can therefore serve as a **wave retarder**, a device in which two orthogonal polarizations travel at different phase velocities so that one is retarded with respect to the other (see Fig. 6.1-8).

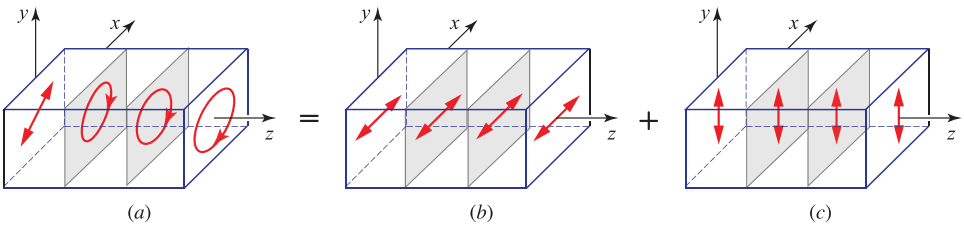


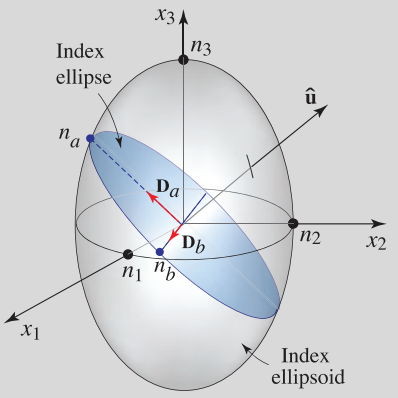
Figure 6.3-5 A linearly polarized wave at 45° in the $z = 0$ plane (a) is analyzed as a superposition of two linearly polarized components in the x and y directions (normal modes), which travel at velocities c_o/n_1 and c_o/n_2 [(b) and (c), respectively]. As a result of phase retardation, the wave is converted from plane polarization to elliptical polarization (a). It is therefore clear that the initial linearly polarized wave is not a normal mode of the system.

C. Propagation in an Arbitrary Direction

We now consider the general case of a plane wave traveling in an anisotropic crystal in an arbitrary direction defined by the unit vector $\hat{\mathbf{u}}$. We demonstrate that the two normal modes are linearly polarized waves. The refractive indices n_a and n_b, and the directions of polarization of these modes, may be determined by use of a procedure based on the index ellipsoid:

Index-Ellipsoid Construction for Determining Normal Modes

Figure 6.3-6 illustrates a geometrical construction for determining the polarizations and refractive indices n_a and n_b of the normal modes of a wave traveling in the direction of the unit vector $\hat{\mathbf{u}}$ in an anisotropic material characterized by the index ellipsoid:



$$\frac{x_1^2}{n_1^2}+\frac{x_2^2}{n_2^2}+\frac{x_3^2}{n_3^2}=1.$$

Figure 6.3-6 Determination of the normal modes from the index ellipsoid.

- Draw a plane passing through the origin of the index ellipsoid, normal to $\hat{\mathbf{u}}$. The intersection of the plane with the ellipsoid is an ellipse called the **index ellipse**.
- The half-lengths of the major and minor axes of the index ellipse are the refractive indices n_a and n_b of the two normal modes.
- The directions of the major and minor axes of the index ellipse are the directions of the vectors $\mathbf{D}_a$ and $\mathbf{D}_b$ for the normal modes. These directions are orthogonal.
- The vectors $\mathbf{E}_a$ and $\mathbf{E}_b$ may be determined from $\mathbf{D}_a$ and $\mathbf{D}_b$ with the help of (6.3-5).

□ **Proof of the Index-Ellipsoid Construction for Determining the Normal Modes.** To determine the normal modes (see Sec. 6.1B) for a plane wave traveling in the direction $\hat{\mathbf{u}}$, we cast Maxwell's equations (5.3-2)–(5.3-5), and the material equation $\mathbf{D}=\boldsymbol{\epsilon}\mathbf{E}$ given in (6.3-2), as an eigenvalue problem. Since all fields are assumed to vary with the position $\mathbf{r}$ as $\exp(-j\mathbf{k}\cdot\mathbf{r})$, where $\mathbf{k}=k\hat{\mathbf{u}}$, Maxwell's equations (5.4-3) and (5.4-4) reduce to

$$\mathbf{k}\times\mathbf{H}=-\omega\mathbf{D} \tag{6.3-9}$$

$$\mathbf{k}\times\mathbf{E}=\omega\mu_o\mathbf{H} \tag{6.3-10}$$

Substituting (6.3-10) into (6.3-9) leads to

$$\mathbf{k} \times (\mathbf{k} \times \mathbf{E}) = -\omega^2 \mu_o \mathbf{D}. \tag{6.3-11}$$

Using $\mathbf{E} = \boldsymbol{\epsilon}^{-1}\mathbf{D}$, we obtain

$$\mathbf{k} \times (\mathbf{k} \times \boldsymbol{\epsilon}^{-1}\mathbf{D}) = -\omega^2 \mu_o \mathbf{D}. \tag{6.3-12}$$

This is an eigenvalue equation that $\mathbf{D}$ must satisfy. Working with $\mathbf{D}$ is convenient since we know that it lies in a plane normal to the wave direction $\hat{\mathbf{u}}$.

We now simplify (6.3-12) by using $\boldsymbol{\eta} = \epsilon_o \boldsymbol{\epsilon}^{-1}$, $\mathbf{k} = k\hat{\mathbf{u}}$, $n = k/k_o$, and $k_o^2 = \omega^2 \mu_o \epsilon_o$ to obtain

$$-\hat{\mathbf{u}} \times (\hat{\mathbf{u}} \times \boldsymbol{\eta}\, \mathbf{D}) = \frac{1}{n^2}\, \mathbf{D}. \tag{6.3-13}$$

The operation $-\hat{\mathbf{u}} \times (\hat{\mathbf{u}} \times \boldsymbol{\eta}\mathbf{D})$ may be interpreted as a projection of the vector $\boldsymbol{\eta}\mathbf{D}$ onto a plane normal to $\hat{\mathbf{u}}$. We may therefore rewrite (6.3-13) in the form

$$\mathbf{P}_u \boldsymbol{\eta}\, \mathbf{D} = \frac{1}{n^2}\, \mathbf{D}, \tag{6.3-14}$$

where $\mathbf{P}_u$ is an operator representing projection. Equation (6.3-14) is an eigenvalue equation for the operator $\mathbf{P}_u \boldsymbol{\eta}$, with eigenvalue $1/n^2$ and eigenvector $\mathbf{D}$. The two eigenvalues, $1/n_a^2$ and $1/n_b^2$, and two corresponding eigenvectors, $\mathbf{D}_a$ and $\mathbf{D}_b$, which are orthogonal, represent the two normal modes.

The eigenvalue problem (6.3-14) has a simple geometrical interpretation. The tensor $\boldsymbol{\eta}$ is represented geometrically by its quadric representation, the index ellipsoid. The operator $\mathbf{P}_u \boldsymbol{\eta}$ represents projection onto a plane normal to $\hat{\mathbf{u}}$. Solving the eigenvalue problem in (6.3-14) is thus equivalent to finding the principal axes of the ellipse formed by the intersection of the plane normal to $\hat{\mathbf{u}}$ with the index ellipsoid. This is precisely the construction set forth in Fig. 6.3-6 for determining the normal modes. ■

Special Case: Uniaxial Crystals

In uniaxial crystals ($n_1 = n_2 = n_o$ and $n_3 = n_e$) the index ellipsoid of Fig. 6.3-6 is an ellipsoid of revolution. For a wave whose direction of travel $\hat{\mathbf{u}}$ forms an angle θ with the optic axis, the index ellipse has half-lengths n_o and $n(\theta)$, where $n(\theta)$ is determined from the index-ellipsoid equation by making the substitutions $x_1 = n(\theta)\cos\theta$, $x_2 = 0$, and $x_3 = -n(\theta)\sin\theta$. The result is

$$\boxed{\frac{1}{n^2(\theta)} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2},} \tag{6.3-15}$$

Refractive Index of Extraordinary Wave

so that the normal modes have refractive indices $n_b = n_o$ and $n_a = n(\theta)$. The first mode, called the **ordinary wave**, has a refractive index n_o regardless of θ. In accordance with the ellipse shown in Fig. 6.3-7, the second mode, called the **extraordinary wave**, has a refractive index $n(\theta)$ that varies from n_o when $\theta = 0°$, to n_e when $\theta = 90°$. The vector $\mathbf{D}$ of the ordinary wave is normal to the plane defined by the optic axis (z axis) and the direction of wave propagation $\mathbf{k}$, and the vectors $\mathbf{E}$ and $\mathbf{D}$ are parallel. The extraordinary wave, on the other hand, has a vector $\mathbf{D}$ that is normal to $\mathbf{k}$ and lies in the k–z plane, and $\mathbf{E}$ is not parallel to $\mathbf{D}$, as shown in Fig. 6.3-7.

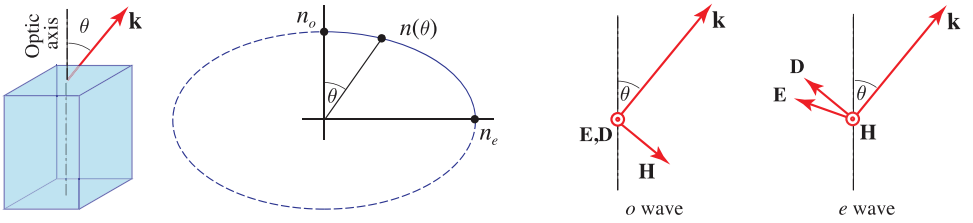


Figure 6.3-7 Variation of the refractive index $n(\theta)$ of the extraordinary wave with θ (the angle between the direction of propagation and the optic axis) in a uniaxial crystal, and directions of the electromagnetic fields of the ordinary (o) and extraordinary (e) waves. The circle with a dot at the center located at the origin signifies that the direction of the vector is out of the plane of the paper, toward the reader.

D. Dispersion Relation, Rays, Wavefronts, and Energy Transport

We now examine other properties of waves in anisotropic media including the dispersion relation (the relation between ω and $\mathbf{k}$).

The optical wave is characterized by the wavevector $\mathbf{k}$, the field vectors $\mathbf{E}$, $\mathbf{D}$, $\mathbf{H}$, and $\mathbf{B}$, and the complex Poynting vector $\mathbf{S} = \frac{1}{2}\mathbf{E} \times \mathbf{H}^*$ (direction of power flow). These vectors are related by (6.3-9) and (6.3-10). It follows from (6.3-9) that $\mathbf{D}$ is normal to both $\mathbf{k}$ and $\mathbf{H}$. Equation (6.3-10) similarly indicates that $\mathbf{H}$ is normal to both $\mathbf{k}$ and $\mathbf{E}$. These geometrical conditions are illustrated in Fig. 6.3-8, which also shows the complex Poynting vector $\mathbf{S}$, which is orthogonal to both $\mathbf{E}$ and $\mathbf{H}$. Thus, $\mathbf{D}$, $\mathbf{E}$, $\mathbf{k}$, and $\mathbf{S}$ lie in one plane to which $\mathbf{H}$ and $\mathbf{B}$ are normal. In this plane $\mathbf{D} \perp \mathbf{k}$ and $\mathbf{S} \perp \mathbf{E}$; but $\mathbf{D}$ is not necessarily parallel to $\mathbf{E}$, and $\mathbf{S}$ is not necessarily parallel to $\mathbf{k}$.

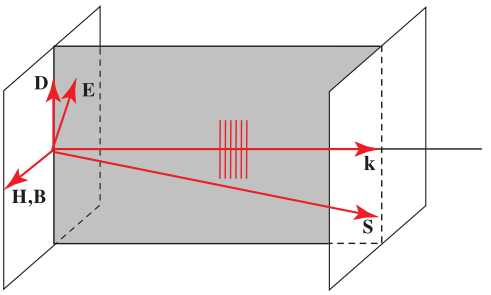


Figure 6.3-8 The vectors $\mathbf{D}$, $\mathbf{E}$, $\mathbf{k}$, and $\mathbf{S}$ all lie in a single plane, to which $\mathbf{H}$ and $\mathbf{B}$ are normal. Also $\mathbf{D} \perp \mathbf{k}$ and $\mathbf{E} \perp \mathbf{S}$. The wavefronts are orthogonal to $\mathbf{k}$.

Dispersion Relation: The k *Surface*

Using the relation $\mathbf{D} = \boldsymbol{\epsilon}\mathbf{E}$ in (6.3-11), we obtain

$$\mathbf{k} \times (\mathbf{k} \times \mathbf{E}) + \omega^2 \mu_o \boldsymbol{\epsilon} \mathbf{E} = \mathbf{0}. \tag{6.3-16}$$

This vector equation, which $\mathbf{E}$ must satisfy, translates to three linear homogeneous equations for the components E_1, E_2, and E_3 along the principal axes, written in the matrix form

$$\begin{bmatrix} n_1^2 k_o^2 - k_2^2 - k_3^2 & k_1 k_2 & k_1 k_3 \\ k_2 k_1 & n_2^2 k_o^2 - k_1^2 - k_3^2 & k_2 k_3 \\ k_3 k_1 & k_3 k_2 & n_3^2 k_o^2 - k_1^2 - k_2^2 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{6.3-17}$$

where (k_1, k_2, k_3) are the components of $\mathbf{k}$, $k_o = \omega/c_o$, and (n_1, n_2, n_3) are the principal refractive indices given by (6.3-6). A nontrivial solution to these equations obtains when the determinant of the matrix is set to zero, which yields

$$\sum_{j=1,2,3} \frac{k_j^2}{k^2 - n_j^2\, k_o^2} = 1, \tag{6.3-18}$$

Dispersion Relation
k Surface

where $k^2 = k_1^2 + k_2^2 + k_3^2$ and $k_o = \omega/c_o$. This relation is known as the **dispersion relation**. It is the equation of a surface $\omega = \omega(k_1, k_2, k_3)$ in the k_1, k_2, k_3 space, known as the **normal surface** or the **k** surface.

The **k** surface is a centrosymmetric surface comprising two sheets, each corresponding to a solution (a normal mode). It can be shown that the **k** surface intersects each of the principal planes in an ellipse and a circle, as illustrated in Fig. 6.3-9. For biaxial crystals ($n_1 < n_2 < n_3$), the two sheets meet at four points, defining two optic axes. In the uniaxial case ($n_1 = n_2 = n_o$, $n_3 = n_e$), the two sheets become a sphere and an ellipsoid of revolution that meet at only two points, thereby defining a single optic axis (the z axis). In the isotropic case ($n_1 = n_2 = n_3 = n$), the two sheets degenerate into a single sphere.

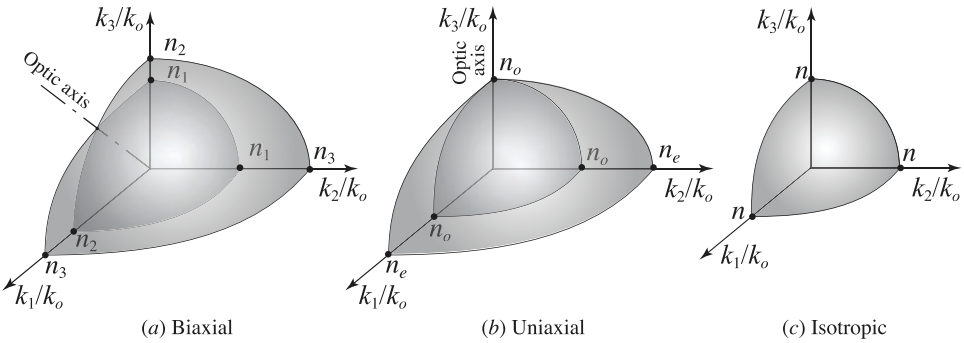


Figure 6.3-9 One octant of the **k** surface for (a) a biaxial crystal ($n_1 < n_2 < n_3$); (b) a uniaxial crystal ($n_1 = n_2 = n_o$, $n_3 = n_e$); and (c) an isotropic crystal ($n_1 = n_2 = n_3 = n$).

Determining Properties of the Normal Modes from the k *Surface*

Much like the index ellipsoid, the **k** surface can be used to determine the normal modes for waves propagating in any prescribed direction, but it also provides physical insight into a collection of other properties of these modes, as considered below:

Refractive indices. The intersection of the prescribed direction $\hat{\mathbf{u}} = (u_1, u_2, u_3)$ of the wave with the **k** surface is a point at a distance from the origin equal to the wavenumber $k = n\omega/c_o$, from which the refractive index n and the corresponding phase velocity $c = \omega/k$ may be determined. In fact, associated with each direction are two intersections corresponding to two wavenumbers and two refractive indices for the two normal modes. Substituting $(k_1, k_2, k_3) = (u_1 k, u_2 k, u_3 k)$ into (6.3-18) yields

$$\sum_{j=1,2,3} \frac{u_j^2\, k^2}{k^2 - n_j^2\, k_o^2} = 1. \tag{6.3-19}$$

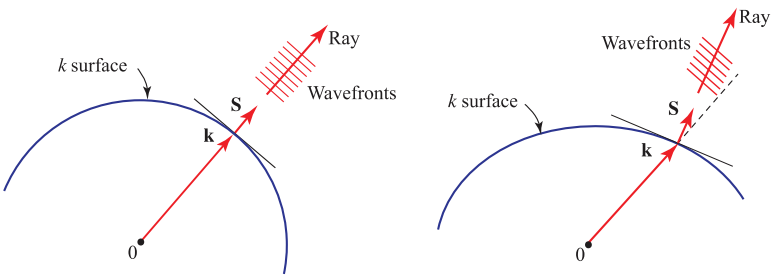
This is a fourth-order equation in k (or second order in k^2). It has four solutions, $\pm k_a$ and $\pm k_b$, of which only the two positive values are meaningful, since the negative values represent a reversed direction of propagation. The problem is therefore solved: the wavenumbers of the normal modes are k_a and k_b and the refractive indices are $n_a = k_a/k_o$ and $n_b = k_b/k_o$.

Polarization. To find the directions of polarization of the two normal modes, we determine the components $(k_1, k_2, k_3) = (ku_1, ku_2, ku_3)$ and the elements of the matrix in (6.3-17) for each of the two wavenumbers $k = k_a$ and $k = k_b$. We then solve two of the three equations in (6.3-17) to establish the ratios E_1/E_3 and E_2/E_3, from which we find the directions of the corresponding fields $\mathbf{E}_a$ and $\mathbf{E}_b$. Note that $\mathbf{E}_a$ and $\mathbf{E}_b$ for the two modes are not necessarily orthogonal, whereas $\mathbf{D}_a$ and $\mathbf{D}_b$ are.

Group velocity. The group velocity may also be determined from the $\mathbf{k}$ surface. In analogy with the group velocity $v = d\omega/dk$ that governs the propagation of light pulses (wavepackets), as discussed in Sec. 5.7, the group velocity for *rays* (localized beams or spatial wavepackets) is the vector $\mathbf{v} = \nabla_k \omega(\mathbf{k})$, the gradient of ω with respect to $\mathbf{k}$. Since the $\mathbf{k}$ surface is the surface $\omega(k_1, k_2, k_3) =$ constant, $\mathbf{v}$ must be normal to the $\mathbf{k}$ surface. Thus, rays travel along directions normal to the $\mathbf{k}$ surface. The wavefronts are perpendicular to the wavevector $\mathbf{k}$ since the phase of the wave is $\mathbf{k} \cdot \mathbf{r}$. The wavefront normals are therefore parallel to the wavevector $\mathbf{k}$.

Energy transport. The complex Poynting vector $\mathbf{S} = \frac{1}{2}\mathbf{E} \times \mathbf{H}^*$ is also normal to the $\mathbf{k}$ surface. This can be demonstrated by choosing a value for ω and considering two vectors $\mathbf{k}$ and $\mathbf{k} + \Delta\mathbf{k}$ that lie on the $\mathbf{k}$ surface. Taking the differentials of (6.3-9) and (6.3-10), and using certain vector identities, shows that $\Delta\mathbf{k} \cdot \mathbf{S} = 0$, so that $\mathbf{S}$ is normal to the $\mathbf{k}$ surface. Consequently, $\mathbf{S}$ is also parallel to the group velocity vector $\mathbf{v}$.

Optical rays. If the $\mathbf{k}$ surface is a sphere, as it is for isotropic media, the vectors $\mathbf{k}$, $\mathbf{S}$, and $\mathbf{v}$ are all parallel, indicating that rays are parallel to the wavevector $\mathbf{k}$ and energy flows in the same direction, as illustrated in Fig. 6.3-10(a). On the other hand, if the $\mathbf{k}$ surface is not normal to the wavevector $\mathbf{k}$, as illustrated in Fig. 6.3-10(b), the rays and the direction of energy transport are not orthogonal to the wavefronts. Rays then have the "extraordinary" property of traveling at an oblique angle to their wavefronts [Fig. 6.3-10(b)].



(a) Ordinary (b) Extraordinary

Figure 6.3-10 Rays and wavefronts for (a) a spherical $\mathbf{k}$ surface, and (b) a nonspherical $\mathbf{k}$ surface.

Special Case: Uniaxial Crystals

In uniaxial crystals ($n_1 = n_2 = n_o$ and $n_3 = n_e$), the equation of the $\mathbf{k}$ surface $\omega = \omega(k_1, k_2, k_3)$ simplifies to

$$(k^2 - n_o^2 k_o^2)\left(\frac{k_1^2 + k_2^2}{n_e^2} + \frac{k_3^2}{n_o^2} - k_o^2\right) = 0. \tag{6.3-20}$$

This equation has two solutions: a sphere, corresponding to the leftmost factor vanishing:

$$k = n_o k_o, \tag{6.3-21}$$

and an ellipsoid of revolution, corresponding to the rightmost factor vanishing:

$$\frac{k_1^2 + k_2^2}{n_e^2} + \frac{k_3^2}{n_o^2} = k_o^2. \tag{6.3-22}$$

Because of symmetry about the z axis (optic axis), there is no loss of generality in assuming that the vector $\mathbf{k}$ lies in the y–z plane. Its direction is then characterized by the angle θ it makes with the optic axis. It is thus convenient to draw the k-surfaces only in the y–z plane, as a circle and an ellipse, as shown in Fig. 6.3-11.

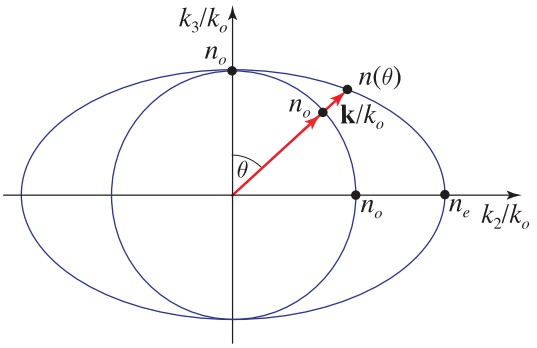
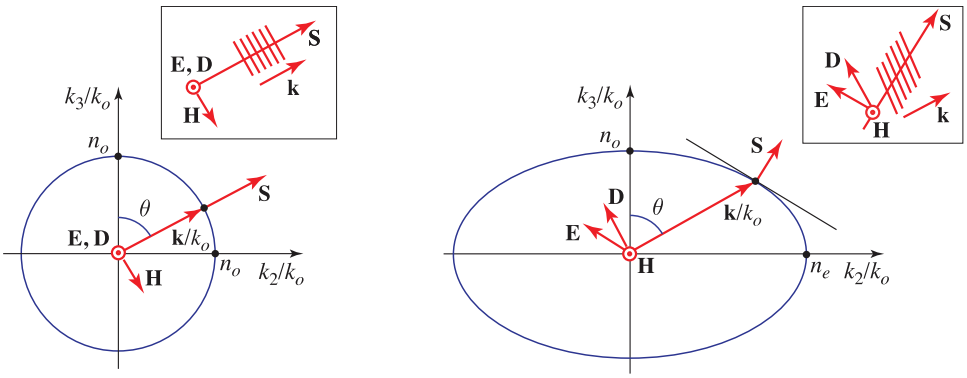


Figure 6.3-11 Intersection of the $\mathbf{k}$ surfaces with the y–z plane for a positive uniaxial crystal ($n_e > n_o$).

Given the direction $\hat{\mathbf{u}}$ of the vector $\mathbf{k}$, the wavenumber k is determined by finding the intersection with the $\mathbf{k}$ surfaces. The two solutions define the two normal modes, the ordinary and extraordinary waves. The ordinary wave has wavenumber $k = n_o k_o$ regardless of the direction of $\hat{\mathbf{u}}$, whereas the extraordinary wave has wavenumber $n(\theta)k_o$, where $n(\theta)$ is given by (6.3-15), thereby confirming earlier results obtained from the index-ellipsoid geometrical construction. The directions of the rays, wavefronts, energy flow, and field vectors $\mathbf{E}$ and $\mathbf{D}$ for the ordinary and extraordinary waves in a uniaxial crystal are illustrated in Fig. 6.3-12.



(*a*) Ordinary (*b*) Extraordinary

Figure 6.3-12 The normal modes for a plane wave traveling in a direction $\mathbf{k}$ that makes an angle θ with the optic axis z of a uniaxial crystal are: (*a*) An ordinary wave of refractive index n_o polarized in a direction normal to the k–z plane. (*b*) An extraordinary wave of refractive index $n(\theta)$ [given by (6.3-15)] polarized in the k–z plane along a direction tangential to the ellipse (the $\mathbf{k}$ surface) at the point of its intersection with $\mathbf{k}$. This wave is "extraordinary" in the following ways: $\mathbf{D}$ is not parallel to $\mathbf{E}$ but both lie in the k–z plane, and $\mathbf{S}$ is not parallel to $\mathbf{k}$ so that power does not flow along the direction of $\mathbf{k}$; the rays are therefore not normal to the wavefronts so that the wave travels "sideways."

E. Double Refraction

Refraction of Plane Waves

We now examine the refraction of a plane wave at the boundary between an isotropic medium (say air, $n = 1$) and an anisotropic medium (a crystal). The key principle that governs the refraction of waves for this configuration is that the wavefronts of the incident and refracted waves must be matched at the boundary. Because the anisotropic medium supports two modes with distinctly different phase velocities, and therefore different indices of refraction, an incident wave gives rise to two refracted waves with different directions and different polarizations. The effect is known as **double refraction** or **birefringence**.

The phase-matching condition requires that Snell's law be obeyed, i.e.,

$$k_o \sin\theta_1 = k \sin\theta, \tag{6.3-23}$$

where θ_1 and θ are the angles of incidence and refraction, respectively. In an anisotropic medium, however, the wavenumber $k = n(\theta)k_o$ is itself a function of θ, so that

$$\sin\theta_1 = n(\theta_a + \theta)\sin\theta, \tag{6.3-24}$$

where θ_a is the angle between the optic axis and the normal to the surface, so that $\theta_a + \theta$ is the angle the refracted ray makes with the optic axis. Equation (6.3-24) is a modified version of Snell's law. To solve (6.3-23), we draw the intersection of the $\mathbf{k}$ surface with the plane of incidence and search for an angle θ for which (6.3-23) is satisfied. Two solutions, corresponding to the two normal modes, are expected. The polarization state of the incident light governs the distribution of energy among the two refracted waves.

Take, for example, a uniaxial crystal and a plane of incidence parallel to the optic axis. The $\mathbf{k}$ surfaces intersect the plane of incidence in a circle and an ellipse (Fig. 6.3-13). The two refracted waves that satisfy the phase-matching condition are determined by satisfying (6.3-24):

- An ordinary wave of orthogonal polarization (TE) at an angle $\theta = \theta_o$, for which

$$\sin\theta_1 = n_o \sin\theta_o\,; \tag{6.3-25}$$

- An extraordinary wave of parallel polarization (TM) at an angle $\theta = \theta_e$, for which

$$\sin\theta_1 = n(\theta_a + \theta_e)\sin\theta_e\,, \tag{6.3-26}$$

  where $n(\theta)$ is given by (6.3-15).

If the incident wave carries the two polarizations, the two refracted waves will emerge, as shown in Fig. 6.3-13.

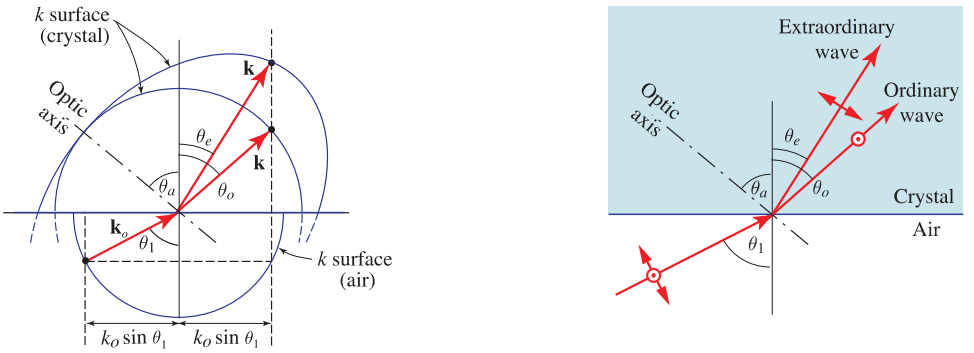


Figure 6.3-13 Determination of the angles of refraction by matching projections of the $\mathbf{k}$ vectors in air and in a uniaxial crystal.

Refraction of Rays

The analysis immediately above dealt with the refraction of plane waves. The refraction of rays is different in an anisotropic medium, since rays do not necessarily travel in directions normal to the wavefronts. In air, before entering the crystal, the wavefronts are normal to the rays. The refracted wave must have a wavevector that satisfies the phase-matching condition, so that Snell's law (6.3-24) is applicable, with the angle of refraction θ determining the direction of $\mathbf{k}$. However, since the direction of $\mathbf{k}$ is not the direction of the ray, Snell's law is not applicable to rays in anisotropic media.

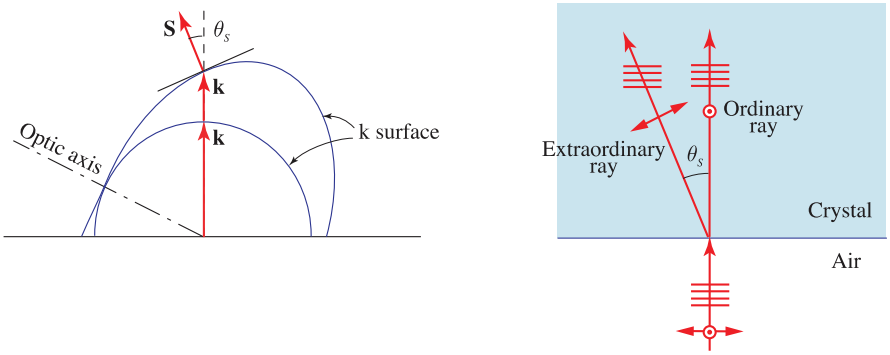


Figure 6.3-14 Double refraction at normal incidence.

An example that dramatizes the deviation from Snell's law is that of normal incidence into a uniaxial crystal whose optic axis is neither parallel nor perpendicular to the crystal boundary. The incident wave has a $\mathbf{k}$ vector normal to the boundary. To ensure phase matching, the refracted waves must also have wavevectors in the same direction. Intersections with the $\mathbf{k}$ surface yield two points corresponding to two waves. The ordinary ray is parallel to $\mathbf{k}$. But the extraordinary ray points in the direction of the normal to the $\mathbf{k}$ surface, at an angle θ_s with the normal to the crystal boundary, as illustrated in Fig. 6.3-14. Thus, normal incidence creates oblique refraction. The principle of phase matching is maintained, however: wavefronts of both refracted rays are parallel to the crystal boundary and to the wavefront of the incident ray.

When light rays are transmitted through a plate of anisotropic material as described above, the two rays refracted at the first surface refract again at the second surface, creating two laterally separated rays with orthogonal polarizations, as illustrated in Fig. 6.3-15.

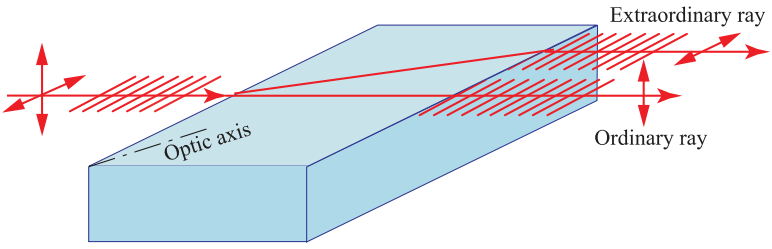


Figure 6.3-15 Double refraction through an anisotropic plate. The plate serves as a polarizing beamsplitter.

6.4 OPTICAL ACTIVITY AND MAGNETO-OPTICS

A. Optical Activity

Certain materials act as natural polarization rotators, a property known as **optical activity**. Their normal modes are waves that are circularly, rather than linearly polarized; waves with right- and left-circular polarizations travel at different phase velocities.

We demonstrate below that an optically active medium with right- and left-circular-polarization phase velocities c_o/n_+ and c_o/n_- acts as a polarization rotator with an angle of rotation $\pi(n_- - n_+)d/\lambda_o$ that is proportional to the thickness of the medium d. The rotatory power (rotation angle per unit length) of the optically active medium is therefore

$$\rho = \frac{\pi}{\lambda_o}(n_- - n_+). \tag{6.4-1}$$

Rotatory Power

The direction in which the polarization plane rotates is the same as that of the circularly polarized component with the greater phase velocity (smaller refractive index). If $n_+ < n_-$, ρ is positive and the rotation is in the same direction as the electric-field vector of the right circularly polarized wave [clockwise when viewed from the direction toward which the wave is approaching, as illustrated in Fig. 6.4-1(a)]. Such materials are said to be **dextrorotatory**, whereas those for which $n_+ > n_-$ are termed **levorotatory**.

□ **Derivation of the Rotatory Power.** Equation (6.4-1) may be derived by decomposing the incident linearly polarized wave into a sum of right and left circularly polarized components of equal amplitudes (see Example 6.1-1),

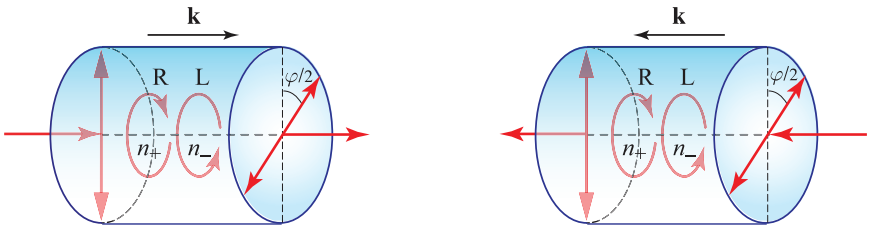
$$\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix} = \tfrac{1}{2}e^{-j\theta}\begin{bmatrix}1\\ j\end{bmatrix} + \tfrac{1}{2}e^{j\theta}\begin{bmatrix}1\\ -j\end{bmatrix}, \tag{6.4-2}$$

where θ is the initial angle of the plane of polarization. After propagating a distance d through the medium, the phase shifts encountered by the right and left circularly polarized waves are $\varphi_+ = 2\pi n_+ d/\lambda_o$ and $\varphi_- = 2\pi n_- d/\lambda_o$, respectively, resulting in a Jones vector

$$\tfrac{1}{2}e^{-j\theta}e^{-j\varphi_+}\begin{bmatrix}1\\ j\end{bmatrix} + \tfrac{1}{2}e^{j\theta}e^{-j\varphi_-}\begin{bmatrix}1\\ -j\end{bmatrix} = e^{-j\varphi_o}\begin{bmatrix}\cos(\theta-\varphi/2)\\ \sin(\theta-\varphi/2)\end{bmatrix}, \tag{6.4-3}$$

where $\varphi_o = \frac{1}{2}(\varphi_+ + \varphi_-)$ and $\varphi = \varphi_- - \varphi_+ = 2\pi(n_- - n_+)d/\lambda_o$. This Jones vector represents a linearly polarized wave with the plane of polarization rotated by an angle $\varphi/2 = \pi(n_- - n_+)d/\lambda_o$, as provided in (6.4-1). ■

Optical activity occurs in materials with an intrinsically helical structure. Examples include selenium, tellurium, tellurium oxide (TeO_2), quartz (α-SiO_2), and cinnabar (HgS). Optically active liquids consist of so-called chiral molecules, which come in distinct left- and right-handed mirror-image forms. Many organic compounds, such as amino acids and sugars, exhibit optical activity. Almost all amino acids are levorotatory, whereas common sugars come in both forms: dextrose (d-glucose) and levulose (fructose) are dextrorotatory and levorotatory, respectively, as their names imply. The rotatory power and sense of rotation for solutions of such substances are therefore sensitive to both the concentration and structure of the solute. A saccharimeter is used to determine the optical activity of sugar solutions, from which the sugar concentration is calculated.



(*a*) Forward wave (*b*) Backward wave

Figure 6.4-1 (*a*) The rotation of the plane of polarization by an optically active medium results from the difference in the velocities for the two circular polarizations. In this illustration, the right circularly polarized wave (R) is faster than the left circularly polarized wave (L), i.e., $n_+ < n_-$, so that ρ is positive and the material is dextrorotatory. (*b*) If the wave in (*a*) is reflected after traversing the medium, the plane of polarization rotates in the opposite direction so that the wave retraces itself.

Material Equations

A time-varying magnetic flux density $\mathbf{B}$ applied to an optically active structure induces a circulating current, by virtue of its helical character, that sets up an electric dipole moment (and hence a polarization) proportional to $j\omega\mathbf{B} = -\nabla \times \mathbf{E}$. The optically active medium is therefore spatially dispersive; i.e., the relation between $\mathbf{D}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r})$ is not local. $\mathbf{D}(\mathbf{r})$ at position $\mathbf{r}$ is determined not only by $\mathbf{E}(\mathbf{r})$, but also by $\mathbf{E}(\mathbf{r}')$ at points $\mathbf{r}'$ in the immediate vicinity of $\mathbf{r}$, since it is dependent on the spatial derivatives contained in $\nabla \times \mathbf{E}(\mathbf{r})$. For a plane wave, we have $\mathbf{E}(\mathbf{r}) = \mathbf{E}\exp(-j\mathbf{k}\cdot\mathbf{r})$ and $\nabla \times \mathbf{E} = -j\mathbf{k} \times \mathbf{E}$, so that the electric permittivity tensor is dependent on the wavevector $\mathbf{k}$. Spatial dispersiveness is analogous to temporal dispersiveness, which has its origin in the noninstantaneous response of the medium (see Sec. 5.2). While the permittivity of a medium exhibiting temporal dispersion depends on the frequency ω, that of a medium exhibiting spatial dispersion depends on the wavevector $\mathbf{k}$.

An optically active medium is described by the $\mathbf{k}$-dependent material equation

$$\mathbf{D} = \epsilon\mathbf{E} + j\epsilon_o\,\xi\mathbf{k} \times \mathbf{E}, \tag{6.4-4}$$

where ξ is a pseudoscalar whose sign depends on the handedness of the coordinate system. This relation is a first-order approximation of the $\mathbf{k}$ dependence of the permittivity tensor, under appropriate symmetry conditions.[†] The first term represents the response of an isotropic dielectric medium whereas the second term accounts for the optical activity, as will be shown subsequently. This $\mathbf{D}$–$\mathbf{E}$ relation is often written in the form

$$\mathbf{D} = \epsilon\mathbf{E} + j\epsilon_o\mathbf{G} \times \mathbf{E}, \tag{6.4-5}$$

where $\mathbf{G} = \xi\mathbf{k}$ is a pseudovector known as the **gyration vector**. In such media the vector $\mathbf{D}$ is clearly not parallel to $\mathbf{E}$ since the vector $\mathbf{G} \times \mathbf{E}$ in (6.4-5) is perpendicular to $\mathbf{E}$.

Normal Modes of the Optically Active Medium

We proceed to show that the two normal modes of the medium described by (6.4-5) are circularly polarized waves, and we determine the velocities c_o/n_+ and c_o/n_- in terms of the constant $G = \xi k$.

[†] See, e.g., L. D. Landau, E. M. Lifshitz, and L. P. Pitaevskii, *Electrodynamics of Continuous Media*, Butterworth–Heinemann, 2nd English ed. 1984, reprinted with corrections 2004, Chapter 12.

We assume that the wave propagates in the z direction, so that $\mathbf{k} = (0, 0, k)$ and thus $\mathbf{G} = (0, 0, G)$. Equation (6.4-5) may then be written in matrix form as

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \epsilon_o \begin{bmatrix} n^2 & -jG & 0 \\ jG & n^2 & 0 \\ 0 & 0 & n^2 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}, \tag{6.4-6}$$

where $n^2 = \epsilon/\epsilon_o$. The diagonal elements in (6.4-6) correspond to propagation in an isotropic medium with refractive index n, whereas the off-diagonal elements, proportional to G, represent the optical activity.

To prove that the normal modes are circularly polarized, consider the two circularly polarized waves with electric-field vectors $\mathbf{E} = (E_0, \pm jE_0, 0)$. The $+$ and $-$ signs correspond to right and left circularly polarized waves, respectively. Substitution in (6.4-6) yields $\mathbf{D} = (D_0, \pm jD_0, 0)$, where $D_0 = \epsilon_o(n^2 \pm G)E_0$. It follows that $\mathbf{D} = \epsilon_o n_\pm^2 \mathbf{E}$, where

$$\boxed{n_\pm = \sqrt{n^2 \pm G}\,.} \tag{6.4-7}$$

Hence, for either of the two circularly polarized waves the vector $\mathbf{D}$ is parallel to the vector $\mathbf{E}$. Equation (6.3-11) is satisfied if the wavenumber $k = n_\pm k_o$. Thus, the right and left circularly polarized waves propagate without changing their state of polarization, with refractive indices n_+ and n_-, respectively. They are therefore the normal modes for this medium.

EXERCISE 6.4-1

Rotatory Power of an Optically Active Medium. Show that if $G \ll n$, the rotatory power of an optically active medium (rotation of the polarization plane per unit length) is approximately given by

$$\rho \approx -\frac{\pi G}{\lambda_o n}\,. \tag{6.4-8}$$

The rotatory power is strongly dependent on the wavelength. Since G is proportional to k, as indicated by (6.4-5), it is inversely proportional to the wavelength λ_o. Thus, the rotatory power in (6.4-8) is inversely proportional to λ_o^2. Moreover, the refractive index n is itself wavelength dependent. By way of example, the rotatory power ρ of quartz is ≈ 31 deg/mm at $\lambda_o = 500$ nm and ≈ 22 deg/mm at $\lambda_o = 600$ nm; for silver thiogallate ($AgGaS_2$), ρ is ≈ 700 deg/mm at 490 nm and ≈ 500 deg/mm at 500 nm.

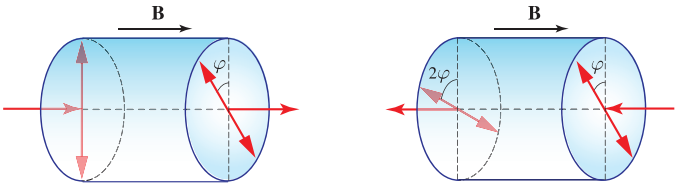
B. Magneto-Optics: The Faraday Effect

Many materials act as polarization rotators in the presence of a static magnetic field, a property known as the **Faraday effect**. The angle of rotation is then proportional to the thickness of the material, and the rotatory power ρ (rotation angle per unit length) is proportional to the component of the magnetic flux density B in the direction of the wave propagation,

$$\rho = VB, \tag{6.4-9}$$

where V is called the **Verdet constant**.

The sense of rotation is governed by the direction of the magnetic field: for $V > 0$, the rotation is in the direction of a right-handed screw pointing in the direction of the magnetic field [Fig. 6.4-2(a)]. In contrast to optical activity, however, the sense of rotation does not reverse with the reversal of the direction of propagation of the wave. Thus, when a wave travels through a Faraday rotator and then reflects back onto itself, traveling once more through the rotator in the opposite direction, it undergoes twice the rotation [Fig. 6.4-2(b)]. Materials that exhibit the Faraday effect include glasses as well



(a) Forward wave (b) Backward wave

Figure 6.4-2 (a) Polarization rotation in a medium exhibiting the Faraday effect. (b) The sense of rotation is invariant to the direction of travel of the wave.

as yttrium iron garnet (YIG), terbium gallium garnet (TGG), and terbium aluminum garnet (TbAlG). The Verdet constant of TbAlG is $V \approx -1.16$ min/Oe-cm at $\lambda_o = 500$ nm. Thin films of these ferrimagnetic materials are used to make compact devices.

Material Equations

In magneto-optic materials, the electric permittivity tensor $\boldsymbol{\epsilon}$ is altered by the application of a *static* magnetic field $\mathbf{H}$, so that $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(\mathbf{H})$. This effect originates from the interaction of the static magnetic field with the motion of the electrons in the material in response to an *optical* electric field $\mathbf{E}$. For the Faraday effect, in particular, the material equation is

$$\mathbf{D} = \epsilon \mathbf{E} + j\epsilon_o \mathbf{G} \times \mathbf{E} \tag{6.4-10}$$

with

$$\mathbf{G} = \gamma_B \mathbf{B}. \tag{6.4-11}$$

Here, $\mathbf{B} = \mu\mathbf{H}$ is the static magnetic flux density, and γ_B is a constant of the medium known as the **magnetogyration coefficient**.

Equation (6.4-10) is identical to (6.4-5) so that the vector $\mathbf{G} = \gamma_B \mathbf{B}$ in Faraday rotators plays the role of the gyration vector $\mathbf{G} = \xi \mathbf{k}$ in optically active media. For the Faraday effect, however, $\mathbf{G}$ does not depend on $\mathbf{k}$, so that reversing the direction of propagation does not reverse the sense of rotation of the plane of polarization. This property is useful for constructing optical isolators, as explained in Sec. 6.6D.

With this analogy, and using (6.4-8), we conclude that the rotatory power of the Faraday medium is $\rho \approx -\pi G/\lambda_o n = -\pi\gamma_B B/\lambda_o n$, from which the Verdet constant (rotatory power per unit magnetic flux density) is seen to be

$$\boxed{V \approx -\frac{\pi\gamma_B}{\lambda_o n}.} \tag{6.4-12}$$

The Verdet constant is clearly a function of the wavelength λ_o.

6.5 OPTICS OF LIQUID CRYSTALS

Liquid Crystals

A liquid crystal comprises a collection of elongated organic molecules that are typically cigar-shaped. The molecules lack positional order (like liquids) but possess orientational order (like crystals). There are three types (phases) of liquid crystals, as illustrated in Fig. 6.5-1:

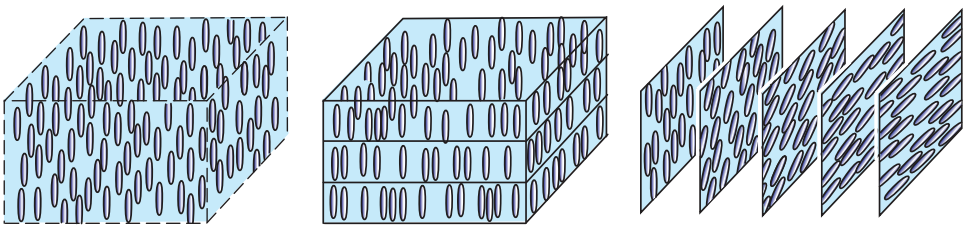


Figure 6.5-1 Molecular organizations of different types of liquid crystals.

- In **nematic liquid crystals** the orientations of the molecules tend to be the same but their positions are totally random.
- In **smectic liquid crystals** the orientations of the molecules are the same, but their centers are stacked in parallel layers within which they have random positions; they therefore have positional order only in one dimension.
- The **cholesteric liquid crystal** is a distorted form of its nematic cousin in which the orientations undergo helical rotation about an axis.

Liquid crystallinity, which was discovered in 1888, is a *fluid* state of matter; it is intermediate between liquid and solid. The molecules are able to change orientation when subjected to a force. When a thin layer of liquid crystal is placed between two parallel glass plates that are rubbed together, for example, the molecules orient themselves along the direction of rubbing.

Twisted nematic liquid crystals are nematic liquid crystals on which a twist (similar to the twist that exists naturally in the cholesteric phase) is externally imposed. This can be achieved, for example, by placing a thin layer of nematic liquid crystal between two glass plates that are polished in perpendicular directions, as schematized in Fig. 6.5-2. This section is devoted to a discussion of the optical properties of twisted nematic liquid crystals, which are widely used in photonics, e.g., for liquid-crystal displays. The electro-optic properties of twisted nematic liquid crystals, and their use as optical modulators and switches, are described in Sec. 21.3.

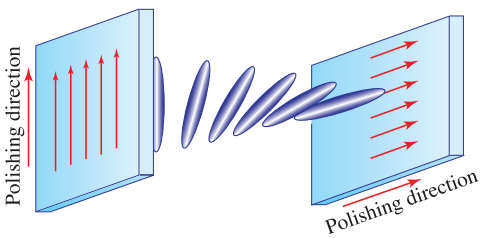


Figure 6.5-2 Molecular orientations of the twisted nematic liquid crystal.

Optical Properties of Twisted Nematic Liquid Crystals

The twisted nematic liquid crystal is an optically *inhomogeneous* and *anisotropic* medium that acts locally as a uniaxial crystal, with the optic axis parallel to the elongated direction. The optical properties are conveniently analyzed by considering the material to be divided into thin layers perpendicular to the axis of twist, each of which acts as a uniaxial crystal; the optic axis is taken to rotate gradually, in a helical fashion, along the axis of twist (Fig. 6.5-3). The cumulative effects of these layers on the transmitted wave is then calculated. We show that, under certain conditions, the twisted nematic liquid crystal acts as a polarization rotator in which the plane of polarization rotates in alignment with the molecular twist.

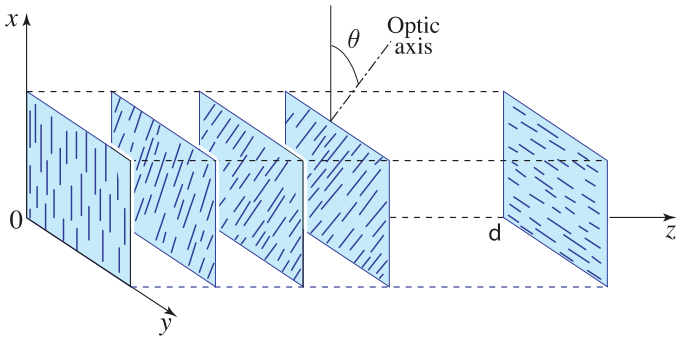


Figure 6.5-3 Propagation of light in a twisted nematic liquid crystal. In this diagram the angle of twist is 90°.

Consider the propagation of light along the axis of twist (the z axis) of a twisted nematic liquid crystal and assume that the twist angle θ varies linearly with z,

$$\theta = \alpha z, \tag{6.5-1}$$

where α is the twist coefficient (degrees per unit length). The optic axis is therefore parallel to the x–y plane and makes an angle θ with the x direction. The ordinary and extraordinary refractive indices are n_o and n_e, respectively (typically, $n_e > n_o$), and the phase-retardation coefficient (retardation per unit length) is

$$\beta = (n_e - n_o)k_o. \tag{6.5-2}$$

The liquid-crystal cell is completely characterized by the twist coefficient α and the retardation coefficient β.

In practice, $\beta \gg \alpha$ so that many cycles of phase retardation are introduced before the optic axis rotates appreciably. We show below that if this condition is satisfied, and the incident wave at $z = 0$ is linearly polarized in the x direction, then the wave maintains its linearly polarized state but the plane of polarization rotates in alignment with the molecular twist, so that the angle of rotation is $\theta = \alpha z$ and the total rotation in a crystal of length d is the angle of twist αd. The liquid-crystal cell then serves as a polarization rotator with rotatory power α. The polarization-rotation property of the twisted nematic liquid crystal is useful for making display devices, as explained in Sec. 21.3.

□ **Proof that the Twisted Nematic Liquid Crystal Acts as a Polarization Rotator.** We proceed to show that the twisted nematic liquid crystal acts as a polarization rotator if $\beta \gg \alpha$. We divide the overall width of the cell d into N incremental layers of equal widths $\Delta z = d/N$. The mth layer, located at the distance $z = z_m = m\Delta z$, $m = 1, 2, \ldots, N$, is a wave retarder whose slow axis (the optic axis) makes an angle $\theta_m = m\Delta\theta$ with the x axis, where $\Delta\theta = \alpha\Delta z$. It therefore has a Jones matrix [see (6.1-24)]

$$\mathbf{T}_m = \mathbf{R}(-\theta_m)\,\mathbf{T}_r\,\mathbf{R}(\theta_m), \tag{6.5-3}$$

where

$$\mathbf{T}_r = \begin{bmatrix} \exp(-jn_e k_o \Delta z) & 0 \\ 0 & \exp(-jn_o k_o \Delta z) \end{bmatrix} \tag{6.5-4}$$

is the Jones matrix of a wave retarder whose axis is along the x direction and $\mathbf{R}(\theta)$ is the coordinate rotation matrix in (6.1-22).

It is convenient to rewrite $\mathbf{T}_r$ in terms of the phase-retardation coefficient $\beta = (n_e - n_o)k_o$,

$$\mathbf{T}_r = \exp(-j\varphi\Delta z) \begin{bmatrix} \exp(-j\beta\Delta z/2) & 0 \\ 0 & \exp(j\beta\Delta z/2) \end{bmatrix}, \tag{6.5-5}$$

where $\varphi = (n_o + n_e)k_o/2$. Since multiplying the Jones vector by a constant phase factor does not affect the state of polarization, we simply ignore the prefactor $\exp(-j\varphi\Delta z)$ in (6.5-5).

The overall Jones matrix of the device is the product

$$\mathbf{T} = \prod_{m=N}^{1} \mathbf{T}_m = \prod_{m=N}^{1} \mathbf{R}(-\theta_m)\,\mathbf{T}_r\,\mathbf{R}(\theta_m). \tag{6.5-6}$$

Using (6.5-3) and noting that $\mathbf{R}(\theta_m)\,\mathbf{R}(-\theta_{m-1}) = \mathbf{R}(\theta_m - \theta_{m-1}) = \mathbf{R}(\Delta\theta)$, we obtain

$$\mathbf{T} = \mathbf{R}(-\theta_N)\,[\mathbf{T}_r\mathbf{R}(\Delta\theta)]^{N-1}\,\mathbf{T}_r\,\mathbf{R}(\theta_1). \tag{6.5-7}$$

Substituting from (6.5-5) and (6.1-22), we obtain

$$\mathbf{T}_r\,\mathbf{R}(\Delta\theta) = \begin{bmatrix} \exp(-j\beta\Delta z/2) & 0 \\ 0 & \exp(j\beta\Delta z/2) \end{bmatrix} \begin{bmatrix} \cos\alpha\Delta z & \sin\alpha\Delta z \\ -\sin\alpha\Delta z & \cos\alpha\Delta z \end{bmatrix}. \tag{6.5-8}$$

Using (6.5-7) and (6.5-8), the Jones matrix $\mathbf{T}$ of the device can, in principle, be determined in terms of the parameters α, β, and $d = N\Delta z$.

For $\alpha \ll \beta$, we may assume that the incremental rotation matrix $\mathbf{R}(\Delta\theta)$ is approximately the identity matrix, whereupon

$$\begin{aligned} \mathbf{T} &\approx \mathbf{R}(-\theta_N)\,[\mathbf{T}_r]^N\,\mathbf{R}(\theta_1) = \mathbf{R}(-\alpha N\Delta z) \begin{bmatrix} \exp(-j\beta\Delta z/2) & 0 \\ 0 & \exp(j\beta\Delta z/2) \end{bmatrix}^N \\ &= \mathbf{R}(-\alpha N\Delta z) \begin{bmatrix} \exp(-j\beta N\Delta z/2) & 0 \\ 0 & \exp(j\beta N\Delta z/2) \end{bmatrix}, \end{aligned} \tag{6.5-9}$$

so that

$$\mathbf{T} = \mathbf{R}(-\alpha d) \begin{bmatrix} \exp(-j\beta d/2) & 0 \\ 0 & \exp(j\beta d/2) \end{bmatrix}. \tag{6.5-10}$$

This Jones matrix represents a wave retarder of retardation βd with the slow axis along the x direction, followed by a polarization rotator with rotation angle αd. If the original wave is linearly polarized along the x direction, the wave retarder imparts only a phase shift; the device then simply rotates the polarization by an angle αd equal to the twist angle. A wave linearly polarized along the y direction is rotated by the same angle. ■

6.6 POLARIZATION DEVICES

This section offers a brief description of a number of devices that are used to modify the state of polarization of light. The basic principles underlying the operation of these devices have been set forth earlier in this chapter.

A. Polarizers

A linear polarizer is a device that transmits the component of the electric field that lies along the direction of its transmission axis while blocking the orthogonal component. The blocking action may be achieved by selective absorption, selective reflection from isotropic media, or selective reflection/refraction in anisotropic media.

Polarization by Selective Absorption (Dichroism)

The absorption of light by certain anisotropic media, called **dichroic materials**, depends on the direction of the incident electric field (Fig. 6.6-1). These materials generally have anisotropic molecular structures whose response is sensitive to the direction of the electric field. The most common dichroic material is **Polaroid H-sheet**, invented in 1938 and still in common use. It is fabricated from a sheet of iodine-impregnated polyvinyl alcohol that is heated and stretched in a particular direction. The analogous device in the infrared is the **wire-grid polarizer**, which comprises a planar configuration of closely spaced fine wires stretched in a single direction. The component of the incident electric field in the direction of the wires is absorbed whereas the component perpendicular to the wires passes through.

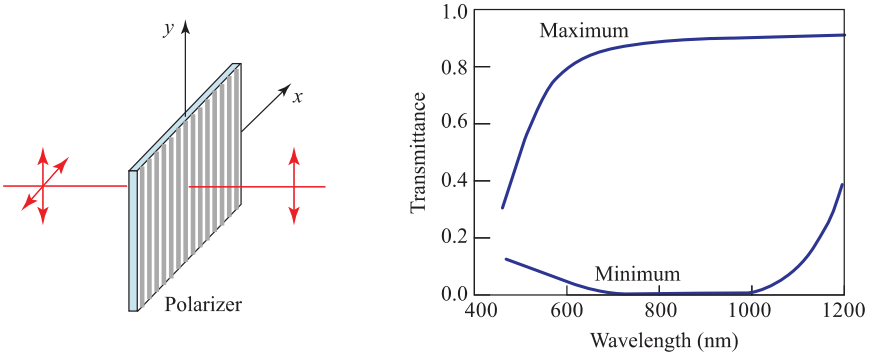


Figure 6.6-1 Power transmittances of a typical dichroic polarizer with the plane of polarization of the light aligned for maximum and minimum transmittance, as indicated.

Polarization by Selective Reflection

The reflectance of light at the boundary between two isotropic dielectric materials is dependent on its polarization, as discussed in Sec. 6.2. At the Brewster angle of incidence, in particular, the reflectance of TM-polarized light vanishes so that it is totally refracted (Fig. 6.2-4). At this angle, therefore, only TE-polarized light is reflected, so that the reflector serves as a polarizer, as depicted in Fig. 6.6-2.

Polarization by Selective Refraction (Polarizing Beamsplitters)

When light enters an anisotropic crystal, the ordinary and extraordinary waves refract at different angles and gradually separate from each other (see Sec. 6.3E and Fig. 6.3-15). This provides an effective means for obtaining polarized light from unpolarized

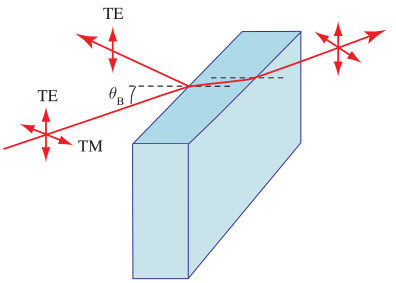
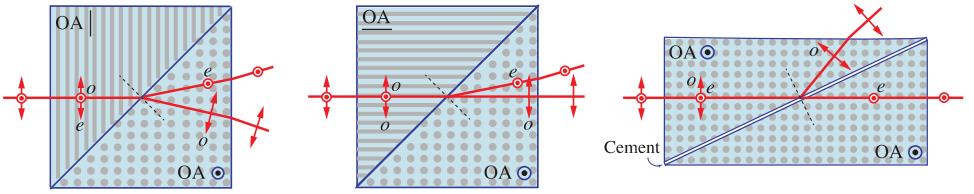


Figure 6.6-2 The Brewster-angle polarizer.

light, and it is commonly used. These devices usually consist of two cemented prisms comprising anisotropic (uniaxial) materials, often with different orientations, as illustrated by the examples in Fig. 6.6-3. These prisms therefore serve as **polarizing beamsplitters**.



(*a*) Wollaston prism (*b*) Rochon prism (*c*) Glan-Thompson prism

Figure 6.6-3 Examples of polarizing beamsplitters. The parallel (p) and orthogonal (s) polarization components of a beam are separated by refraction or reflection at the boundary between two uniaxial crystals whose optic axes (OA) have different orientations. In this illustration, the crystals are negative uniaxial ($n_o > n_e$), such as calcite. (*a*) In the Wollaston prism the p component is extraordinary in the first crystal and ordinary in the second, while the opposite is true for the s component, so that they undergo different angles of refraction. (*b*) In the Rochon prism, the p component is transmitted without refraction since it is ordinary in both crystals, while the s component is refracted since it is ordinary in the first crystal and extraordinary in the second. (*c*) The operation of the Glan-Thompson prism is based on total internal reflection at the boundary between the first crystal and the cement layer. This occurs for only the p polarization since it is an ordinary wave with the higher refractive index. The s component is transmitted. The Glan–Thompson device has the merit of providing a large angular separation between the emerging waves.

B. Wave Retarders

A wave retarder serves to convert a wave with one form of polarization into another form. It is characterized by its retardation Γ and its fast and slow axes (see Sec. 6.1B). The normal modes are linearly polarized waves polarized along the directions of the axes. The velocities of the two waves differ so that transmission through the retarder imparts a relative phase shift Γ to these modes.

Wave retarders are often constructed from anisotropic crystals in the form of plates. As explained in Sec. 6.3B, when light travels along a principal axis of a crystal (say the z axis), the normal modes are linearly polarized waves pointing along the two other principal axes (the x and y axes). These modes experience the principal refractive indices n_1 and n_2, and thus travel at velocities c_o/n_1 and c_o/n_2, respectively. If $n_1 < n_2$, the x axis is the fast axis. If the plate has thickness d, the phase retardation is $\Gamma = (n_2 - n_1)k_o d = 2\pi(n_2 - n_1)d/\lambda_o$. The retardation is thus directly proportional to the thickness d of the plate and inversely proportional to the wavelength λ_o (note, however, that $n_2 - n_1$ is itself wavelength dependent).

The refractive indices of a thin sheet of mica, for example, are 1.599 and 1.594 at $\lambda_o = 633$ nm, so that $\Gamma/d \approx 15.8\pi$ rad/mm. A sheet of thickness 63.3 μm yields $\Gamma \approx \pi$ and thus serves as a half-wave retarder.

Light Intensity Control via a Wave Retarder and Two Polarizers

Consider a wave retarder of retardation Γ placed between two crossed polarizers whose axes are oriented at 45° with respect to the axes of the retarder, as illustrated in Fig. 6.6-4. The power (or intensity) transmittance of this system is

$$\mathcal{T} = \sin^2(\Gamma/2), \tag{6.6-1}$$

which may be established by making use of Jones matrices or by examining the polarization ellipse of the retarded light as a function of Γ, and then determining the component that lies in the direction of the output polarizer, as illustrated in Fig. 6.6-4. If $\Gamma = 0$ no light is transmitted through the system since the polarizers are orthogonal. On the other hand, if $\Gamma = \pi$ all of the light is transmitted since the retarder then rotates the plane of polarization 90° whereupon it matches the transmission axis of the second polarizer.

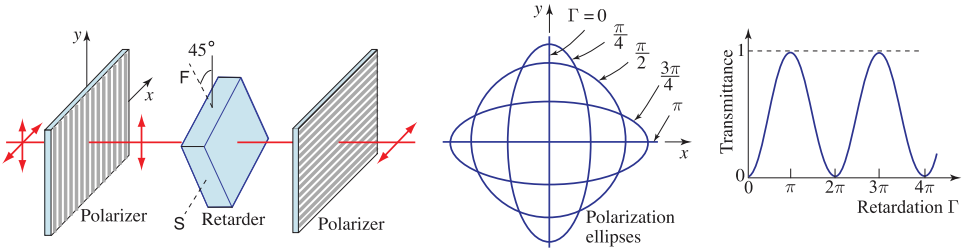


Figure 6.6-4 Controlling light intensity by means of a wave retarder with variable retardation Γ placed between two crossed polarizers.

The intensity of the transmitted light is thus readily controlled by altering the retardation Γ. This can be achieved, for example, by deliberately changing the indices n_1 and n_2 by application of an external DC electric field to the retarder. This is the basic principle that underlies the operation of electro-optic modulators, as discussed in Chapter 21.

Furthermore, since Γ depends on d, slight variations in the thickness of a sample can be monitored by examining the pattern of the transmitted light. Moreover, since Γ is wavelength dependent, the transmittance of the system is frequency sensitive. Though it can be used as a filter, the selectivity is not sharp. Other configurations using wave retarders and polarizers can be used to construct narrowband transmission filters.

C. Polarization Rotators

A polarization rotator serves to rotate the plane of polarization of linearly polarized light by a fixed angle, while maintaining its linearly polarized nature. Optically active media and materials exhibiting the Faraday effect act as polarization rotators, as discussed in Sec. 6.4. The twisted nematic liquid crystal also acts as a polarization rotator under certain conditions, as shown in Sec. 6.5.

If a polarization rotator is placed between two polarizers, the amount of light transmitted depends on the rotation angle. The intensity of the light can therefore be controlled (modulated) if the angle of rotation is externally changed (e.g., by varying the magnetic flux density applied to a Faraday rotator or by changing the molecular orientation of a liquid crystal by means of an applied electric field). Electro-optic modulation of light and liquid-crystal display devices are discussed in Chapter 21.

D. Nonreciprocal Polarization Devices

A device whose effect on the polarization state is invariant to reversal of the direction of propagation is said to be **reciprocal**. If a wave is transmitted through such a device in one direction and the emerging wave is retransmitted in the opposite direction, then it retraces the changes in the polarization state and arrives at the input in the very same initial polarization state. Devices that do not have this directional invariance are termed **nonreciprocal**. All of the polarization systems described in this chapter are reciprocal, with the exception of the Faraday rotator (Sec. 6.4B). A number of useful nonreciprocal polarization devices may be implemented by combining the Faraday rotator with other reciprocal polarization components (see Sec. 24.1C).

Optical Isolator

An **optical isolator** is a device that transmits light in only one direction, thereby acting as a "one-way valve." Optical isolators are useful for preventing reflected light from returning back to the source. Such feedback can have deleterious effects on the operation of certain devices, such as laser diodes.

An optical isolator is constructed by placing a Faraday rotator between two polarizers whose axes make a 45° angle with respect to each other. The magnetic flux density applied to the rotator is adjusted so that it rotates the polarization by 45° in the direction of a right-handed screw pointing in the z direction [Fig. 6.6-5(a)]. Light traveling through the system in the forward direction (from left to right) thus crosses polarizer A, rotates 45°, and is thence transmitted through polarizer B. Linearly polarized light with the polarization plane at 45° but traveling through the system in the backward direction [from right to left in Fig. 6.6-5(b)] successfully crosses polarizer B. However, on passing through the Faraday rotator, the plane of polarization rotates an additional 45° and is therefore blocked by polarizer A. Since the backward light might be generated by reflection of the forward wave from subsequent surfaces, the isolator serves to protect its source from reflected light.

Note that the Faraday rotator is a necessary component of the optical isolator. An optically active, or liquid-crystal, polarization rotator cannot be used in its place. In those *reciprocal* components, the sense of rotation is such that the polarization of the reflected wave retraces that of the incident wave so that the light would be transmitted back through the polarizers to the source.

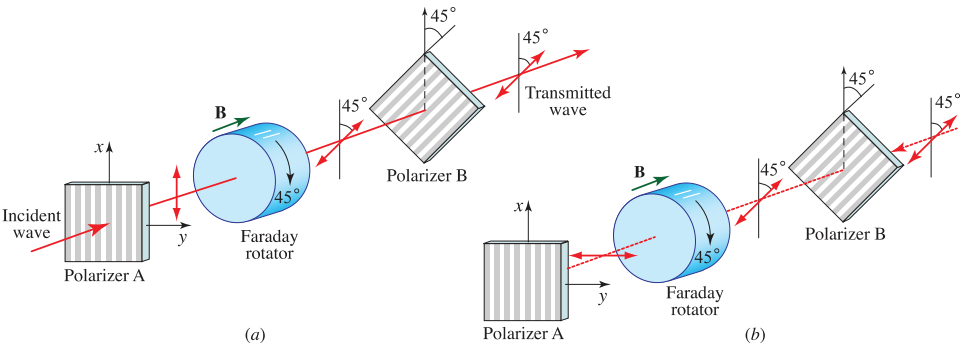


Figure 6.6-5 An optical isolator that makes use of a Faraday rotator transmits light in one direction. (a) A wave traveling in the forward direction is transmitted. (b) A wave traveling in the backward (or reverse) direction is blocked.

Faraday-rotator isolators constructed from yttrium iron garnet (YIG) or terbium gallium garnet (TGG) offer attenuations of the backward wave of up to 90 dB over a relatively wide wavelength range. Thin films of these materials placed in permanent magnetic fields are used to make compact optical isolators.

Nonreciprocal Polarization Rotation

The combination of a $45°$ Faraday rotator and a half-wave retarder is another useful nonreciprocal device. As illustrated in Fig. 6.6-6(a), the state of polarization of a forward-traveling linearly polarized wave, with its plane of polarization oriented at $22.5°$ with the fast axis of the retarder, maintains its state of polarization upon transmission through the device (since it undergoes $45°$ rotation by the Faraday rotator, followed by $-45°$ rotation by the retarder). However, for a wave traveling in the reverse direction, the plane of polarization is rotated by $45° + 45° = 90°$, as can be readily seen in Fig. 6.6-6(b). This device may therefore be used in combination with a polarizing beamsplitter to deflect the backward-traveling wave away from the source of the forward-traveling wave so that it can be accessed independently. A system of this kind can be useful for implementing nonreciprocal interconnects, such as **optical circulators**, as portrayed in Fig. 24.1-10(b).

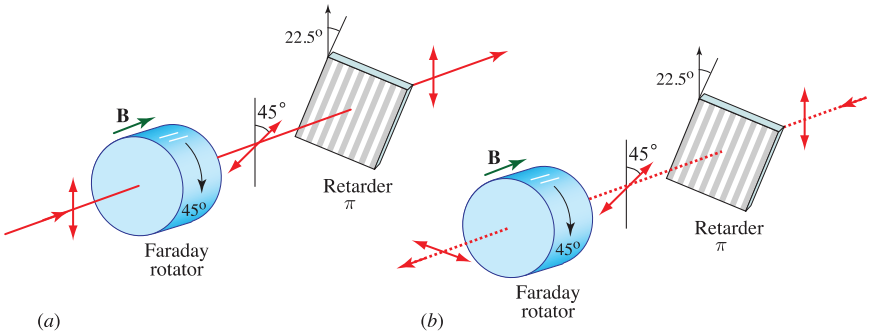


Figure 6.6-6 A Faraday rotator followed by a half-wave (π) retarder is a nonreciprocal device that: (a) maintains the polarization state of a linearly polarized forward-traveling wave, but (b) rotates the plane of polarization of the backward-traveling wave by $90°$.

READING LIST

Polarization

See also the reading list on general optics in Chapter 1 and the reading list in Chapter 5.

A. Kumar and A. Ghatak, *Polarization of Light with Applications in Optical Fibers*, SPIE Optical Engineering Press, 2011.

D. H. Goldstein, *Polarized Light*, CRC Press/Taylor & Francis, 3rd ed. 2010.

L. Nikolova and P. S. Ramanujam, *Polarization Holography*, Cambridge University Press, 2009.

R. Martínez-Herrero, P. M. Mejías, and G. Piquero, *Characterization of Partially Polarized Light Fields*, Springer-Verlag, 2009.

E. Collett, *Field Guide to Polarization*, SPIE Optical Engineering Press, 2005.

J. N. Damask, *Polarization Optics in Telecommunications*, Springer-Verlag, 2004.

S. Sugano and N. Kojima, eds., *Magneto-Optics*, Springer-Verlag, 2000.

C. Brosseau, *Fundamentals of Polarized Light: A Statistical Optics Approach*, Wiley, 1998.

S. Huard, *Polarization of Light*, Wiley, 1996.

E. Collett, *Polarized Light: Fundamentals and Applications*, CRC Press, 1993.

A. Yariv and P. Yeh, *Optical Waves in Crystals: Propagation and Control of Laser Radiation*, Wiley, 1985, paperback reprint 2002.

R. M. A. Azzam and N. M. Bashara, *Ellipsometry and Polarized Light*, North-Holland, 1977, reprinted 1989.

B. A. Robson, *The Theory of Polarization Phenomena*, Clarendon, 1974.

L. Velluz, M. Le Grand, and M. Grosjean, *Optical Circular Dichroism: Principles, Measurements, and Applications*, Academic Press, 1965.

W. A. Shurcliff and S. S. Ballard, *Polarized Light*, Van Nostrand, 1964.

W. A. Shurcliff, *Polarized Light: Production and Use*, Harvard University Press, 1962, reprinted 1966.

Crystals and Tensor Analysis

C. Malgrange, C. Ricolleau, and M. Schlenker, *Symmetry and Physical Properties of Crystals*, Springer-Verlag, 2014.

D. Fleisch, *A Student's Guide to Vectors and Tensors*, Cambridge University Press, 2011.

S. Haussühl, *Physical Properties of Crystals*, Wiley–VCH, 2007.

R. J. D. Tilley, *Crystals and Crystal Structures*, Wiley, 2006.

E. A. Wood, *Crystals and Light: An Introduction to Optical Crystallography*, Dover, 2nd ed. 1977.

J. F. Nye, *Physical Properties of Crystals: Their Representation by Tensors and Matrices*, Oxford University Press, 1985.

Liquid Crystals

See also the reading list on liquid crystal devices and displays in Chapter 21.

J. V. Selinger, *Introduction to the Theory of Soft Matter: From Ideal Gases to Liquid Crystals*, Springer-Verlag, 2016.

I.-C. Khoo, *Liquid Crystals*, Wiley, 2nd ed. 2007.

T. Scharf, *Polarized Light in Liquid Crystals and Polymers*, Wiley, 2006.

P. Oswald and P. Pieranski, *Nematic and Cholesteric Liquid Crystals: Concepts and Physical Properties Illustrated by Experiments*, CRC Press/Taylor & Francis, 2005.

P. J. Collings, *Liquid Crystals: Nature's Delicate Phase of Matter*, Princeton University Press, 2nd ed. 2002.

M. E. Lines and A. M. Glass, *Principles and Applications of Ferroelectrics and Related Materials*, Clarendon, 1977; Oxford University Press, paperback 2nd ed. 2001.

P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Oxford University Press, 2nd ed. 1993.

S. Chandrasekhar, *Liquid Crystals*, Cambridge University Press, 2nd ed. 1992.

L. M. Blinov, *Electro-Optical and Magneto-Optical Properties of Liquid Crystals*, Wiley, 1983.

Articles

L. A. Whitehead and M. A. Mossman, Reflections on Total Internal Reflection, *Optics & Photonics News*, vol. 20, no. 2, pp. 28–34, 2009.

M. Mansuripur, The Faraday Effect, *Optics & Photonics News*, vol. 10, no. 11, pp. 32–36, 1999.

B. H. Billings, ed., *Selected Papers on Applications of Polarized Light*, SPIE Optical Engineering Press (Milestone Series Volume 57), 1992.

S. D. Jacobs, ed., *Selected Papers on Liquid Crystals for Optics*, SPIE Optical Engineering Press (Milestone Series Volume 46), 1992.

B. H. Billings, ed., *Selected Papers on Polarization*, SPIE Optical Engineering Press (Milestone Series Volume 23), 1990.

A. Lakhtakia, ed., *Selected Papers on Natural Optical Activity*, SPIE Optical Engineering Press (Milestone Series Volume 15), 1990.

W. Swindell, ed., *Benchmark Papers in Optics: Polarized Light*, Dowden, Hutchinson & Ross, 1975.

Sir David Brewster's Scientific Work, *Nature*, vol. 25, pp. 157–159, 15 December 1881.

PROBLEMS

6.1-5 **Orthogonal Polarizations.** Show that if two elliptically polarized states are orthogonal, the major axes of their ellipses are perpendicular and the senses of rotation are opposite.

6.1-6 **Rotating a Polarization Rotator.** Show that the Jones matrix of a polarization rotator is invariant to rotation of the coordinate system.

6.1-7 **Jones Matrix of a Polarizer.** Show that the Jones matrix of a linear polarizer whose transmission axis makes an angle θ with the x axis is

$$\mathbf{T} = \begin{bmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{bmatrix}.$$

Derive this result using (6.1-18), (6.1-22), and (6.1-24).

6.1-8 **Half-Wave Retarder.** Consider linearly polarized light passed through a half-wave retarder. If the polarization plane makes an angle θ with the fast axis of the retarder, show that the transmitted light is linearly polarized at an angle $-\theta$, i.e., it is rotated by an angle 2θ. Why is the half-wave retarder not equivalent to a polarization rotator?

6.1-9 **Wave Retarders in Tandem.** Write the Jones matrices for:

(a) A $\pi/2$ wave retarder with the fast axis along the x direction.

(b) A π wave retarder with the fast axis at 45° to the x direction.

(c) A $\pi/2$ wave retarder with the fast axis along the y direction.

If these three retarders are cascaded (placed in tandem), with (c) following (b) following (a), show that the resulting device introduces a 90° rotation. What happens if the order of the three retarders is reversed?

6.1-10 **Reflection of Circularly Polarized Light.** Show that circularly polarized light changes handedness (right becomes left, and *vice versa*), upon reflection from a mirror.

6.1-11 **Anti-Glare Screen.** A self-luminous object is viewed through a glass window. An anti-glare screen is used to eliminate glare caused by the reflection of background light from the surfaces of the window. Show that such a screen may be constructed from a combination of a linear polarizer and a quarter-wave retarder whose axes are at 45° with respect to the transmission axis of the polarizer. Can the screen be regarded as an optical isolator?

6.2-2 **Derivation of Fresnel Equations.** Derive the reflection equation (6.2-6), which is used to derive the Fresnel equation (6.2-8) for TE polarization. How would you go about obtaining the reflection coefficient if the incident light took the form of a beam rather than a plane wave?

6.2-3 **Reflectance of Glass.** A plane wave is incident from air ($n = 1$) onto a glass plate ($n = 1.5$) at an angle of incidence of 45°. Determine the power reflectances of the TE and TM waves. What is the average reflectance for unpolarized light (light carrying TE and TM waves of equal intensities)?

6.2-4 **Refraction at the Brewster Angle.** Use the condition $n_1 \sec\theta_1 = n_2 \sec\theta_2$ and Snell's law, $n_1 \sin\theta_1 = n_2 \sin\theta_2$, to derive (6.2-12) for the Brewster angle. Also show that at the Brewster angle, $\theta_1 + \theta_2 = 90^\circ$, so that the directions of the reflected and refracted waves are orthogonal, and hence the electric field of the refracted TM wave is parallel to the direction of the reflected wave. The reflection of light may be regarded as a scattering process in which the refracted wave acts as a source of radiation generating the reflected wave. At the Brewster angle, this source oscillates in a direction parallel to the direction of propagation of the reflected wave, so that radiation cannot occur and no TM light is reflected.

6.2-5 **Retardation Associated with Total Internal Reflection.** Determine the phase retardation between the TE and TM waves that is introduced by total internal reflection at the boundary between glass ($n = 1.5$) and air ($n = 1$) at an angle of incidence $\theta = 1.2\,\theta_c$, where θ_c is the critical angle.

6.2-6 **Goos–Hänchen Shift.** Consider two TE plane waves undergoing total internal reflection at angles θ and $\theta + d\theta$, where $d\theta$ is an incremental angle. If the phase retardation introduced between the reflected waves is written in the form $d\varphi = \xi\, d\theta$, find an expression for the coefficient ξ. Sketch the interference patterns of the two incident waves and the two reflected waves and verify that they are shifted by a lateral distance proportional to ξ. When the incident wave is a beam (composed of many plane-wave components), the reflected beam is displaced laterally by a distance proportional to ξ. This is known as the Goos–Hänchen effect.

6.2-7 **Reflection from an Absorptive Medium.** Use Maxwell's equations and appropriate boundary conditions to show that the complex amplitude reflectance at the boundary between free space and a medium with refractive index n and absorption coefficient α, at normal incidence, is $\mathrm{r} = [(n - j\alpha c/2\omega) - 1]/[(n - j\alpha c/2\omega) + 1]$.

6.3-1 **Maximum Retardation in Quartz.** Quartz is a positive uniaxial crystal with $n_e = 1.553$ and $n_o = 1.544$. (a) Determine the retardation per mm at $\lambda_o = 633$ nm when the crystal is oriented such that retardation is maximized. (b) At what thickness(es) does the crystal act as a quarter-wave retarder?

6.3-2 **Maximum Extraordinary Effect.** Determine the direction of propagation in quartz ($n_e = 1.553$ and $n_o = 1.544$) at which the angle between the wavevector $\mathbf{k}$ and the Poynting vector $\mathbf{S}$ (which is also the direction of ray propagation) is maximum.

6.3-3 **Double Refraction.** An unpolarized plane wave is incident from free space onto a quartz crystal ($n_e = 1.553$ and $n_o = 1.544$) at an angle of incidence $30°$. The optic axis lies in the plane of incidence and is perpendicular to the direction of the incident wave before it enters the crystal. Determine the directions of the wavevectors and the rays of the two refracted components.

6.3-4 **Lateral Shift in Double Refraction.** What is the optimum geometry for maximizing the lateral shift between the refracted ordinary and extraordinary beams in a positive uniaxial crystal? Indicate all pertinent angles and directions.

6.3-5 **Transmission Through a $LiNbO_3$ Plate.** Examine the transmission of an unpolarized He–Ne laser beam ($\lambda_o = 633$ nm) normally incident on a $LiNbO_3$ plate ($n_e = 2.29$, $n_o = 2.20$) of thickness 1 cm, cut such that its optic axis makes an angle $45°$ with the normal to the plate. Determine the lateral shift at the output of the plate and the retardation between the ordinary and extraordinary beams.

*6.3-6 **Conical Refraction.** When the wavevector $\mathbf{k}$ points along an optic axis of a biaxial crystal an unusual situation occurs. The two sheets of the $\mathbf{k}$ surface meet and the surface can be approximated by a conical surface. Consider a ray normally incident on the surface of a biaxial crystal for which one of its optic axes is also normal to the surface. Show that multiple refraction occurs with the refracted rays forming a cone. This effect is known as **conical refraction**. What happens when the conical rays refract from the parallel surface of the crystal into air?

6.6-1 **Circular Dichroism.** Certain materials have different absorption coefficients for right and left circularly polarized light, a property known as **circular dichroism**. Determine the Jones matrix for a device that converts light with any state of polarization into right circularly polarized light.

6.6-2 **Polarization Rotation by a Sequence of Linear Polarizers.** A wave that is linearly polarized in the x direction is transmitted through a sequence of N linear polarizers whose transmission axes are inclined by angles $m\theta$ ($m = 1, 2, \ldots, N$; $\theta = \pi/2N$) with respect to the x axis. Show that the transmitted light is linearly polarized in the y direction but its amplitude is reduced by the factor $\cos^N\theta$. What happens in the limit as $N \to \infty$? *Hint:* Use Jones matrices and note that

$$\mathbf{R}[(m+1)\theta]\,\mathbf{R}(-m\theta) = \mathbf{R}(\theta),$$

where $\mathbf{R}(\theta)$ is the coordinate transformation matrix.

CHAPTER

7

PHOTONIC-CRYSTAL OPTICS

7.1 OPTICS OF DIELECTRIC LAYERED MEDIA 258
- A. Matrix Theory of Multilayer Optics
- B. Fabry–Perot Etalon
- C. Bragg Grating

7.2 ONE-DIMENSIONAL PHOTONIC CRYSTALS 277
- A. Bloch Modes
- B. Matrix Optics of Periodic Media
- C. Fourier Optics of Periodic Media
- D. Boundaries Between Periodic and Homogeneous Media

7.3 TWO- AND THREE-DIMENSIONAL PHOTONIC CRYSTALS 291
- A. Two-Dimensional Photonic Crystals
- *B. Three-Dimensional Photonic Crystals

Felix Bloch (1905–1983) developed a theory that describes electron waves in the periodic structure of solids.

Eli Yablonovitch (born 1946) coinvented the concept of the photonic bandgap; he made the first photonic-bandgap crystal.

Sajeev John (born 1957) invoked the notion of photon localization and coinvented the photonic-bandgap concept.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

The propagation of light in homogeneous media and its reflection and refraction at the boundaries between different media are a principal concern of optics, as described in the earlier chapters of this book. Photonic devices often comprise multiple layers of different materials arranged, for example, to suppress or enhance reflectance or to alter the spectral or the polarization characteristics of light. Multilayered and stratified media are also found in natural physical and biological systems and are responsible for the distinct colors of some insects and butterfly wings. Multilayered media can also be periodic, i.e., comprise identical dielectric structures replicated in a one-, two-, or three-dimensional periodic arrangement, as illustrated in Fig. 7.0-1. One-dimensional periodic structures include stacks of identical parallel planar multilayer segments. These are often used as gratings that reflect optical waves incident at certain angles, or as filters that selectively reflect waves of certain frequencies. Two-dimensional periodic structures include sets of parallel rods as well as sets of parallel cylindrical holes, such as those used to modify the characteristics of optical fibers known as holey fibers (see Sec. 10.4). Three-dimensional periodic structures comprise arrays of cubes, spheres, or holes of various shapes, organized in lattice structures much like those found in natural crystals. Photonic crystals are a special class of optical metamaterials, which are considered in Chapter 8.

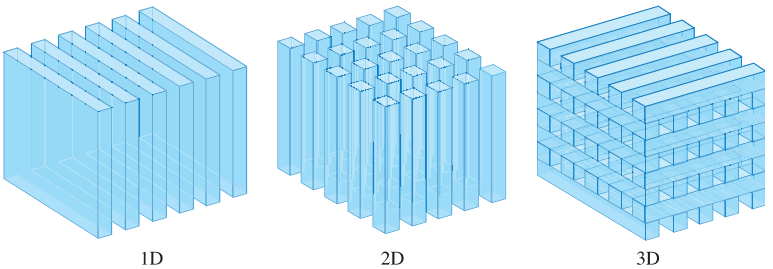


Figure 7.0-1 Periodic photonic structures in one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) configurations.

Optical waves, which are inherently periodic, interact with periodic media in a unique way, particularly when the scale of the periodicity is of the same order as that of the wavelength. For example, spectral bands emerge in which light waves cannot propagate through the medium without severe attenuation. Waves with frequencies lying within these forbidden bands, called **photonic bandgaps**, behave in a manner akin to total internal reflection, but are applicable for all directions. The dissolution of the transmitted wave is a result of destructive interference among the waves scattered by elements of the periodic structure in the forward direction. Remarkably, this effect extends over finite spectral bands, rather than occurring for just single frequencies.

This phenomenon is analogous to the electronic properties of crystalline solids such as semiconductors. In that case, the periodic wave associated with an electron travels in a periodic crystal lattice, and energy bandgaps often materialize. Because of this analogy, the photonic periodic structures have come to be called **photonic crystals**. Photonic crystals enjoy a whole raft of applications, including use as waveguides, fibers, resonators, lasers, filters, routers, switches, gates, and sensors; other applications are in the offing.

An electromagnetic-optics analysis is usually required to describe the optical properties of inhomogeneous media such as multilayered and periodic materials. For inhomogeneous dielectric media, as we know from Sec. 5.2B, the permittivity $\epsilon(\mathbf{r})$ is

spatially varying and the wave equation takes the general forms of (5.2-16) and (5.2-17). For a harmonic wave of angular frequency ω, this leads to generalized Helmholtz equations for the electric and the magnetic fields expressible as

$$\eta(\mathbf{r})\ \nabla \times (\nabla \times \mathbf{E}) = \frac{\omega^2}{c_o^2}\mathbf{E}, \tag{7.0-1}$$

$$\nabla \times [\eta(\mathbf{r})\ \nabla \times \mathbf{H}] = \frac{\omega^2}{c_o^2}\mathbf{H}, \tag{7.0-2}$$

Generalized Helmholtz Equations

where $\eta(\mathbf{r}) = \epsilon_o/\epsilon(\mathbf{r})$ is the electric impermeability (see Sec. 6.3A). One of these equations may be solved for either the electric or the magnetic field, and the other field may be directly determined by use of Maxwell's equations. Note that (7.0-1) and (7.0-2) are cast in the form of an eigenvalue problem: a differential operator applied on the field function equals a constant multiplied by the field function. The eigenvalues are ω^2/c_o^2 and the eigenfunctions provide the spatial distributions of the modes of the propagating field (see Appendix C). For reasons to be explained in Secs. 7.2C and 7.3, we work with the magnetic-field equation (7.0-2) rather than the electric-field equation (7.0-1).

For multilayered media, $\epsilon(\mathbf{r})$ is piecewise constant, i.e., it is uniform within any given layer but changes from one layer to another. Wave propagation can then be studied by using the known properties of optical waves in homogeneous media, together with the appropriate boundary conditions that dictate the laws of reflection and transmission.

Periodic dielectric media are characterized by periodic values of $\epsilon(\mathbf{r})$ and $\eta(\mathbf{r})$. This periodicity imposes certain conditions on the optical wave. For example, the propagation constant deviates from simply proportionality to the angular frequency ω, as for a homogeneous medium. While the modes of propagation in a homogeneous medium are plane waves of the form $\exp(-j\mathbf{k}\cdot\mathbf{r})$, the modes of the periodic medium, known as **Bloch modes**, are traveling waves modulated by standing waves.

This Chapter

Previous chapters have focused on the optics of thin optical components that are well separated, such as thin lenses, planar gratings, and image-bearing films across which the light travels. This chapter addresses the optics of *bulk* media comprising multiple dielectric layers and periodic 1D, 2D, and 3D photonic structures. Section 7.1, in which 1D layered media are considered, serves as a prelude to periodic media and photonic crystals. A matrix approach offers a systematic treatment of the multiple reflections that occur at the multiple boundaries of the medium. Section 7.2 introduces photonic crystals in their simplest form — 1D periodic structures. Matrix methods are adopted to determine the dispersion relation and the band structure. An alternate approach, based on a Fourier-series representation of the periodic functions associated with the medium and the wave, is also presented. These results are generalized in Sec. 7.3 to two- and three-dimensional photonic crystals.

Throughout this chapter, the various media are assumed to be isotropic, and therefore described by a scalar permittivity ϵ, although reflection and refraction at boundaries have inherent polarization-sensitive characteristics.

Photonic Crystals in Other Chapters

By virtue of their omnidirectional reflection property, photonic crystals can be used as "perfect" dielectric mirrors. A slab of homogeneous medium embedded in a pho-

tonic crystal may be used to guide light by multiple reflections from the boundaries. Applications to optical waveguides are described in Sec. 9.5. Similarly, light may be guided through an optical fiber with a homogeneous core embedded in a cladding of the same material, but with cylindrical holes drilled parallel to the fiber axis. Such "holey" fibers, described in Sec. 10.4, offer a number of salutary features not present in ordinary optical fibers. A cavity burrowed in a photonic crystal may function as an optical resonator since it has perfectly reflecting walls at frequencies within the photonic bandgap. Photonic-crystal microresonators and lasers will be described briefly in Secs. 11.4D and 18.5C, respectively.

7.1 OPTICS OF DIELECTRIC LAYERED MEDIA

A. Matrix Theory of Multilayer Optics

A plane wave normally incident on a layered medium undergoes reflections and transmissions at the layer boundaries, which in turn undergo their own reflections and transmissions in an unending process, as illustrated in Fig. 7.1-1(a). The complex amplitudes of the transmitted and reflected waves may be determined by use of the Fresnel equations at each boundary (see Sec. 6.2); the overall transmittance and reflectance of the medium can, in principle, be calculated by superposition of these individual waves. This technique was used in Sec. 2.5B to determine the transmittance of the Fabry–Perot interferometer.

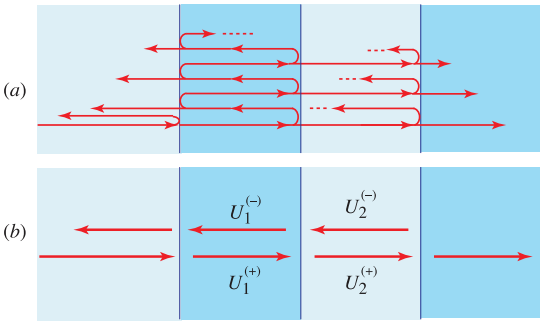


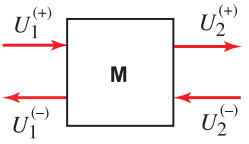
Figure 7.1-1 (a) Reflections of a single wave at the boundaries of a multilayered medium. (b) In each layer, the forward waves are lumped into a forward collected wave $U^{(+)}$ while the backward waves are lumped into a backward collected wave $U^{(-)}$.

When the number of layers is large, tracking the infinite number of "micro" reflections and transmissions can be tedious. An alternative "macro" approach is based on the recognition that within each layer there are two types of waves: forward waves traveling to the right, and backward waves traveling to the left. The sums of these waves add up to a single forward collected wave $U^{(+)}$ and a single backward collected wave $U^{(-)}$ at any point, as illustrated in Fig. 7.1-1(b). Determining the wave propagation in a layered medium is then equivalent to determining the amplitudes of this pair of waves everywhere. The complex amplitudes of the four waves on the two sides of a boundary may be related by imposing the appropriate boundary conditions, or by simply using the Fresnel equations of reflection and transmission.

Wave-Transfer Matrix

Tracking the complex amplitudes of the forward and backward waves through the boundaries of a multilayered medium is facilitated by use of matrix methods. Consider two arbitrary planes within a given optical system, denoted plane 1 and plane 2. The amplitudes of the forward and backward collected waves at plane 1, $U_1^{(+)}$ and $U_1^{(-)}$,

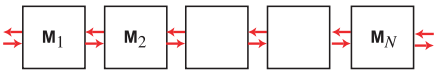
respectively, are represented by a column matrix of dimension 2, and similarly for plane 2. These two column matrices are related by the matrix equation



$$\begin{bmatrix} U_2^{(+)} \\ U_2^{(-)} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} U_1^{(+)} \\ U_1^{(-)} \end{bmatrix}. \tag{7.1-1}$$

The matrix **M**, whose elements are A, B, C, and D, is called the **wave-transfer matrix** (or transmission matrix). It depends on the optical properties of the layered medium between the two planes.

A multilayered medium is conveniently divided into a concatenation of basic elements described by known wave-transfer matrices, say $\mathbf{M}_1$, $\mathbf{M}_2$,..., $\mathbf{M}_N$. The amplitudes of the forward and backward collected waves at the two ends of the overall medium are then related by a single matrix that is the matrix product

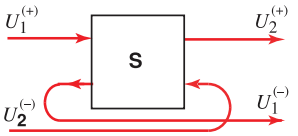


$$\mathbf{M} = \mathbf{M}_N \ldots \mathbf{M}_2 \mathbf{M}_1, \tag{7.1-2}$$

where the elements $1, 2, \ldots, N$ are numbered from left to right as shown in the figure. The wave-transfer matrix cascade formula provided in (7.1-2) is identical to the ray-transfer matrix cascade formula given in (1.4-10), and it proves equally useful.

Scattering Matrix

An alternative to the wave-transfer matrix that relates the four complex amplitudes $U_{1,2}^{(\pm)}$ at the two edges of a layered medium is the scattering matrix, or **S** matrix. It is often used to describe transmission lines, microwave circuits, and scattering systems. In this case, the outgoing waves are expressed in terms of the incoming waves,



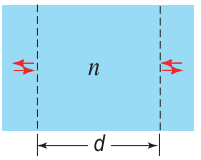
$$\begin{bmatrix} U_2^{(+)} \\ U_1^{(-)} \end{bmatrix} = \begin{bmatrix} \mathrm{t}_{12} & \mathrm{r}_{21} \\ \mathrm{r}_{12} & \mathrm{t}_{21} \end{bmatrix} \begin{bmatrix} U_1^{(+)} \\ U_2^{(-)} \end{bmatrix}, \tag{7.1-3}$$

where the elements of the **S** matrix are denoted t_{12}, r_{21}, r_{12}, and t_{21}. Unlike the wave-transfer matrix, these elements have direct physical significance. The quantities t_{12} and r_{12} are the forward amplitude transmittance and reflectance (i.e., the transmittance and reflectance of a wave incident from the left), respectively, while t_{21} and r_{21} are the amplitude transmittance and reflectance in the backward direction (i.e., a wave coming from the right), respectively. The subscript 12, for example, signifies that the light is incident from medium 1 into medium 2. This can be easily verified by noting that if there is no backward wave at plane 2, so that $U_2^{(-)} = 0$, we obtain $U_2^{(+)} = \mathrm{t}_{12} U_1^{(+)}$ and $U_1^{(-)} = \mathrm{r}_{12} U_1^{(+)}$. Similarly, if there is no forward wave at plane 1, so that $U_1^{(+)} = 0$, we obtain $U_2^{(+)} = \mathrm{r}_{21} U_2^{(-)}$ and $U_1^{(-)} = \mathrm{t}_{21} U_2^{(-)}$.

A distinct advantage of the **S**-matrix formalism is that its elements are directly related to the physical parameters of the system. On the other hand, a disadvantage is that the **S** matrix of a cascade of elements is not the product of the **S** matrices of the constituent elements. A useful systematic procedure for analyzing a cascaded system

therefore draws on both the wave-transfer and scattering matrix approaches: we use the handy multiplication formula of the **M** matrices and then convert to the **S** matrix to determine the overall transmittance and reflectance of the cascaded system.

EXAMPLE 7.1-1. ***Propagation Through a Homogeneous Medium.*** For a homogeneous layer of width d and refractive index n, the complex amplitudes of the collected waves at the planes indicated by the arrows are related by $U_2^{(+)} = e^{-j\varphi}U_1^{(+)}$ and $U_1^{(-)} = e^{-j\varphi}U_2^{(-)}$, where $\varphi = nk_o d$, so that in this case the wave-transfer matrix and the scattering matrix are:



$$\mathbf{M} = \begin{bmatrix} \exp(-j\varphi) & 0 \\ 0 & \exp(j\varphi) \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \exp(-j\varphi) & 0 \\ 0 & \exp(-j\varphi) \end{bmatrix}, \quad \varphi = nk_o d. \tag{7.1-4}$$

Relation between Scattering Matrix and Wave-Transfer Matrix

The elements of the **M** and **S** matrices are related by manipulating the defining equations (7.1-1) and (7.1-3), whereupon the following conversion equations emerge:

$$\mathbf{M} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \frac{1}{\mathrm{t}_{21}} \begin{bmatrix} \mathrm{t}_{12}\mathrm{t}_{21} - \mathrm{r}_{12}\mathrm{r}_{21} & \mathrm{r}_{21} \\ -\mathrm{r}_{12} & 1 \end{bmatrix}, \tag{7.1-5}$$

$$\mathbf{S} = \begin{bmatrix} \mathrm{t}_{12} & \mathrm{r}_{21} \\ \mathrm{r}_{12} & \mathrm{t}_{21} \end{bmatrix} = \frac{1}{D} \begin{bmatrix} AD - BC & B \\ -C & 1 \end{bmatrix}. \tag{7.1-6}$$

(7.1-5), (7.1-6) Matrix Conversion Relations

These equations are not valid in the limiting cases when $\mathrm{t}_{21} = 0$ or $D = 0$.

Summary

Matrix wave optics offers a systematic procedure for determining the amplitude transmittance and reflectance of a stack of dielectric layers with prescribed thicknesses and refractive indices:

- The stack is divided into a cascade of elements encompassing boundaries with homogeneous layers between them.
- The **M** matrix is determined for each element. This may be achieved by using the Fresnel formulas for transmission and reflection to determine its **S** matrix, and then using the conversion relation (7.1-5) to calculate the corresponding **M** matrix.
- The **M** matrix for the full stack of elements is obtained by simply multiplying the **M** matrices for the individual elements, in accordance with the wave-transfer matrix formula provided in (7.1-2).
- Finally, the **S** matrix for the full stack is determined by conversion from the overall **M** matrix via (7.1-6). The elements of the **S** matrix then directly yield the amplitude transmittance and reflectance for the full stack of dielectric layers.

Two Cascaded Systems: Airy Formulas

Matrix methods may be used to derive explicit expressions for elements of the scattering matrix of a composite system in terms of elements of the scattering matrices of the constituent systems. Consider a wave transmitted through a system described by an **S** matrix with elements t_{12}, r_{21}, r_{12}, and t_{21}, followed by another system with **S** matrix elements t_{23}, r_{32}, r_{23}, and t_{32}. By multiplying the two associated **M** matrices, and then converting the result to an **S** matrix, the following formulas for the overall forward transmittance and reflectance can be derived:

$$\mathrm{t}_{13} = \frac{\mathrm{t}_{12}\mathrm{t}_{23}}{1 - \mathrm{r}_{21}\mathrm{r}_{23}}, \qquad \mathrm{r}_{13} = \mathrm{r}_{12} + \frac{\mathrm{t}_{12}\mathrm{t}_{21}\mathrm{r}_{23}}{1 - \mathrm{r}_{21}\mathrm{r}_{23}}. \tag{7.1-7}$$

If the two cascaded systems are mediated by propagation through a homogeneous medium, as illustrated in Fig. 7.1-2, then by use of the wave-transfer matrix in (7.1-4), with the phase $\varphi = nk_o d$, where d is the propagation distance and n is the refractive index of the medium, the following formulas for the overall transmittance and reflectance, known as the **Airy formulas**, may be derived:

$$\mathrm{t}_{13} = \frac{\mathrm{t}_{12}\mathrm{t}_{23}\exp(-j\varphi)}{1 - \mathrm{r}_{21}\mathrm{r}_{23}\exp(-j2\varphi)}, \qquad \mathrm{r}_{13} = \mathrm{r}_{12} + \frac{\mathrm{t}_{12}\mathrm{t}_{21}\mathrm{r}_{23}\exp(-j2\varphi)}{1 - \mathrm{r}_{21}\mathrm{r}_{23}\exp(-j2\varphi)}. \tag{7.1-8}$$

Airy Formulas

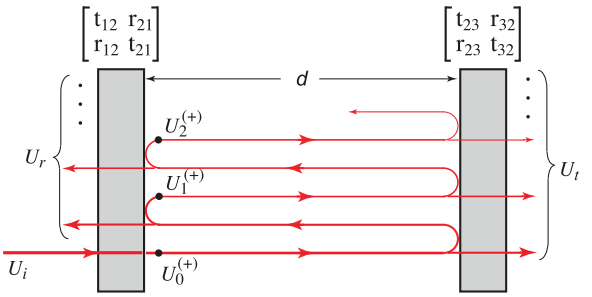


Figure 7.1-2 Transmission of a plane wave through a cascade of two separated systems.

The Airy formulas may also be derived by tracking the multiple transmissions and reflections experienced by an incident wave between the two systems and adding up their amplitudes, as portrayed in Fig. 7.1-2. A plane wave of complex amplitude U_i incident on the first system produces an initial internal wave of amplitude $U_0^{(+)} = \mathrm{t}_{12}U_i$, which reflects back and forth between the two subsystems producing additional internal waves $U_1^{(+)}, U_2^{(+)}, \ldots$, all traveling in the forward direction. The amplitude of the overall transmitted wave U_t is related to the total internal amplitude $U^{(+)} = U_0^{(+)} + U_1^{(+)} + \ldots$ by $U_t = \mathrm{t}_{23}\exp(-j\phi)\,U^{(+)}$, where $\phi = nk_o d$. The overall amplitude transmittance is therefore $\mathrm{t}_{13} = U_t/U_i = \mathrm{t}_{12}\mathrm{t}_{23}\exp(-j\phi)(U^{(+)}/U_0^{(+)})$. Since $U^{(+)} = U_0^{(+)}(1 + h + h^2 + \cdots) = U_0^{(+)}/(1-h)$, where $h = \mathrm{r}_{21}\mathrm{r}_{23}\exp(-j2\phi)$ is the round-trip multiplication factor, the overall amplitude transmittance t_{13} yields the Airy formula in (7.1-8).

Conservation Relations for Lossless Media

If the medium between planes 1 and 2 is nonlossy, then the incoming and outgoing optical powers must be equal. Furthermore, if the media at the input and output planes have the same impedance and refractive index, then these powers are represented by

the squared magnitudes of the complex amplitudes $\left|U_{1,2}^{(\pm)}\right|^2$. In this case, conservation of power dictates that $\left|U_1^{(+)}\right|^2 + \left|U_2^{(-)}\right|^2 = \left|U_2^{(+)}\right|^2 + \left|U_1^{(-)}\right|^2$ for any combination of incoming amplitudes. By choosing the incoming amplitudes $U_1^{(+)}$ and $U_2^{(-)}$ to be (1,0), (0,1), and (1,1), the conservation formula above yields three equations that relate the elements of the **S** matrix. These equations can be used to prove the following formulas:

$$|\mathrm{t}_{12}| = |\mathrm{t}_{21}| \equiv |\mathrm{t}|, \quad |\mathrm{r}_{12}| = |\mathrm{r}_{21}| \equiv |\mathrm{r}|, \quad |\mathrm{t}|^2 + |\mathrm{r}|^2 = 1, \tag{7.1-9}$$

$$\mathrm{t}_{12}/\mathrm{t}_{21}^* = -\mathrm{r}_{12}/\mathrm{r}_{21}^*. \tag{7.1-10}$$

Equations (7.1-9) relate the magnitudes of the elements of the **S** matrix for lossless media whose input and output planes see the same refractive index, whereas (7.1-10) relates their arguments.

The formulas in (7.1-9) and (7.1-10) translate to the following relations among the elements of the **M** matrix:

$$|D| = |A|, \quad |C| = |B|, \quad |A|^2 - |B|^2 = 1, \tag{7.1-11}$$

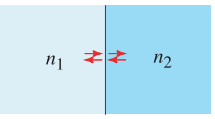
$$\det \mathbf{M} = C/B^* = A/D^* = \mathrm{t}_{12}/\mathrm{t}_{21}, \quad |\det \mathbf{M}| = 1. \tag{7.1-12}$$

These results can be derived by substituting the conservation relations for lossless media, (7.1-9) and (7.1-10), into the conversion relations between the wave-transfer and scattering matrices, (7.1-5) and (7.1-6).

EXAMPLE 7.1-2. ***Single Dielectric Boundary.*** Consider a system comprising a single boundary. In accordance with the Fresnel equations (Sec. 6.2), the transmittance and reflectance at a boundary between two media of refractive indices n_1 and n_2 are governed by the **S** matrix

$$\mathbf{S} = \begin{bmatrix} \mathrm{t}_{12} & \mathrm{r}_{21} \\ \mathrm{r}_{12} & \mathrm{t}_{21} \end{bmatrix} = \frac{1}{n_1 + n_2} \begin{bmatrix} 2n_1 & n_2 - n_1 \\ n_1 - n_2 & 2n_2 \end{bmatrix}. \tag{7.1-13}$$

Substituting (7.1-13) into (7.1-5) yields the **M** matrix



$$\mathbf{M} = \frac{1}{2n_2} \begin{bmatrix} n_2 + n_1 & n_2 - n_1 \\ n_2 - n_1 & n_2 + n_1 \end{bmatrix}. \tag{7.1-14}$$

Lossless Symmetric Systems

For lossless systems with reciprocal symmetry, namely systems whose transmission/reflection in the forward and backward directions are identical, we have $\mathrm{t}_{21} = \mathrm{t}_{12} \equiv \mathrm{t}$ and $\mathrm{r}_{21} = \mathrm{r}_{12} \equiv \mathrm{r}$. In this case, (7.1-9) and (7.1-10) yield

$$|\mathrm{t}|^2 + |\mathrm{r}|^2 = 1, \quad \mathrm{t}/\mathrm{r} = -(\mathrm{t}/\mathrm{r})^*, \quad \arg\{\mathrm{r}\} - \arg\{\mathrm{t}\} = \pm\pi/2\,, \tag{7.1-15}$$

indicating that the phases associated with transmission and reflection differ by $\pi/2$. Under these conditions, the elements of the **M** matrix satisfy the following relations:

$$A = D^*, \quad B = C^*, \quad |A|^2 - |B|^2 = 1, \quad \det \mathbf{M} = 1. \tag{7.1-16}$$

The **S** and **M** matrices then take the simple form

$$\mathbf{S} = \begin{bmatrix} \mathrm{t} & \mathrm{r} \\ \mathrm{r} & \mathrm{t} \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} 1/\mathrm{t}^* & \mathrm{r}/\mathrm{t} \\ \mathrm{r}^*/\mathrm{t}^* & 1/\mathrm{t} \end{bmatrix}, \tag{7.1-17}$$

Lossless Symmetric System

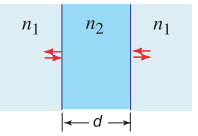
and the system is described by two complex numbers t and r, related by (7.1-15).

Only relative phases are significant in such systems, so we may assume without loss of generality that $\arg\{\mathrm{t}\} = 0$. It then follows from (7.1-15) that $\arg\{\mathrm{r}\} = \pm\pi/2$ so that $\mathrm{r} = \pm j|\mathrm{r}|$, whereupon the matrices in (7.1-17) take the simpler forms

$$\mathbf{S} = \begin{bmatrix} |\mathrm{t}| & j|\mathrm{r}| \\ j|\mathrm{r}| & |\mathrm{t}| \end{bmatrix}, \quad \mathbf{M} = \frac{1}{|\mathrm{t}|}\begin{bmatrix} 1 & j|\mathrm{r}| \\ -j|\mathrm{r}| & 1 \end{bmatrix}, \quad |\mathrm{t}|^2 + |\mathrm{r}|^2 = 1.. \tag{7.1-18}$$

These equations are commonly used to describe lossless symmetric systems such as beamsplitters (e.g., cube and pellicle beamsplitters) and integrated-optic couplers. Moreover, if the system is balanced, i.e., if $|r| = |t| = \frac{1}{\sqrt{2}}$, we have $\mathbf{S} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & j \\ j & 1 \end{bmatrix}$.

EXAMPLE 7.1-3. ***Dielectric Slab.*** Consider a lossless symmetric system comprising a cascade of three subsystems: a boundary between media of refractive indices n_1 and n_2, followed by travel through a medium of index n_2, followed in turn by a boundary between media with indices n_2 and n_1. By virtue of the results provided in (7.1-2) and Example 7.1-2, the overall **M** matrix is then a product of the three constituent **M** matrices, with the matrix multiplication taking place in reverse order:



$$\mathbf{M} = \frac{1}{4n_1n_2}\begin{bmatrix} n_1+n_2 & n_1-n_2 \\ n_1-n_2 & n_1+n_2 \end{bmatrix}\begin{bmatrix} e^{-j\varphi} & 0 \\ 0 & e^{j\varphi} \end{bmatrix}\begin{bmatrix} n_2+n_1 & n_2-n_1 \\ n_2-n_1 & n_2+n_1 \end{bmatrix}. \tag{7.1-19}$$

Here $\varphi = n_2 k_o d$ where d is the width of the slab. The elements of this matrix **M**, which are given by

$$A = D^* = \frac{1}{\mathrm{t}^*} = \frac{1}{4n_1n_2}\left[(n_1+n_2)^2\, e^{-j\varphi} - (n_2-n_1)^2\, e^{j\varphi}\right], \tag{7.1-20}$$

$$B = C^* = \frac{\mathrm{r}}{\mathrm{t}} = -j\frac{(n_2^2-n_1^2)}{4n_1n_2}\sin\varphi, \tag{7.1-21}$$

satisfy the properties of a lossless symmetric system, as described by (7.1-16). Expressions for t and r are determined directly from (7.1-20) and (7.1-21):

$$\mathrm{t} = \frac{4n_1n_2\exp(-j\varphi)}{(n_1+n_2)^2 - (n_1-n_2)^2\exp(-j2\varphi)}, \quad \mathrm{r} = -j\left[\frac{n_2^2-n_1^2}{4n_1n_2}\sin\varphi\right]\mathrm{t}. \tag{7.1-22}$$

The intensity transmittance $|\mathrm{t}|^2$, and the intensity reflectance $|\mathrm{r}|^2 = 1-|\mathrm{t}|^2$, are periodic functions of the phase φ, with period π. The magnitude of the phase difference between r and t is always maintained at $\pi/2$, regardless of the value of φ, but its sign changes as $\sin\varphi$ switches its sign every π interval. It is worthy of note that the expressions provided in (7.1-20) and (7.1-21) can also be directly derived by regarding the system as a combination of two boundaries mediated by propagation through a distance in a medium, and using the Airy formula (7.1-8) with $\mathrm{t}_{12} = \mathrm{t}_{32} = 2n_1/(n_1+n_2)$, $\mathrm{t}_{21} = \mathrm{t}_{23} = 2n_2/(n_1+n_2)$, and $\mathrm{r}_{12} = \mathrm{r}_{32} = -\mathrm{r}_{21} = -\mathrm{r}_{23} = (n_1-n_2)/(n_1+n_2)$.

EXERCISE 7.1-1

Quarter-Wave Film as an Antireflection Coating. Specially designed thin dielectric films are often used to reduce or eliminate reflection at the boundary between two media of different refractive indices. Consider a thin film of refractive index n_2 and thickness d sandwiched between media of refractive indices n_1 and n_3. Derive an expression for the B element of the **M** matrix for this multilayer medium. Show that light incident from medium 1 has zero reflectance if $d = \lambda/4$ and $n_2 = \sqrt{n_1 n_3}$, where $\lambda = \lambda_o/n_2$.

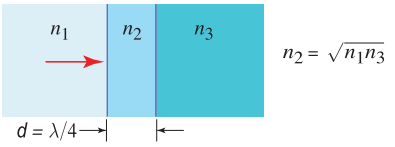


Figure 7.1-3 Antireflection coating.

Off-Axis Waves in Layered Media

When an oblique wave is incident on a layered medium, the transmitted and reflected waves, along with their reflections and transmissions in turn, bounce back and forth between the layers, as illustrated by its real part as shown in Fig. 7.1-4(a). The laws of reflection and refraction ensure that, within the same layer, all of the forward waves are parallel, and all of the backward waves are parallel. Moreover, within any given layer the forward and backward waves travel at the same angle, when measured from the $+z$ and $-z$ directions, respectively.

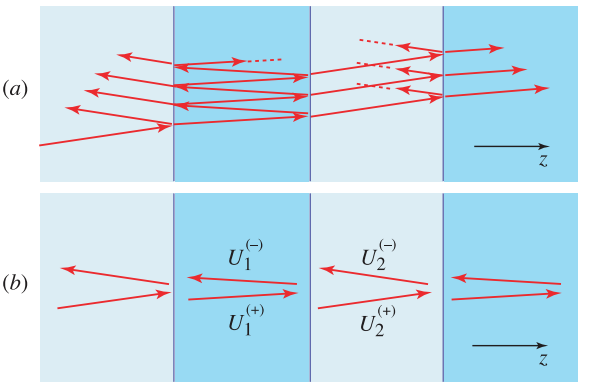


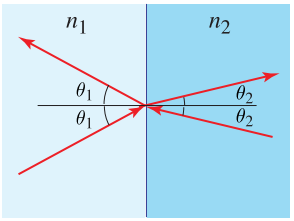
Figure 7.1-4 (a) Reflections of a single incident oblique wave at the boundaries of a multilayered medium. (b) In each layer, the forward waves are lumped into a collected forward wave while the backward waves are lumped into a collected backward wave.

The "macro" approach that was used earlier for normally incident waves is similarly applicable for oblique waves. The distinction is that the Fresnel transmittances and reflectances at a boundary, t_{12}, r_{21}, r_{12}, and t_{21}, are angle-dependent as well as polarization-dependent (see Sec. 6.2).

The simplest example is propagation a distance d through a homogeneous medium of refractive index n, at an angle θ measured from the z axis. The wave-transfer matrix **M** is then given by (7.1-4), where the phase is now $\varphi = nk_o d\cos\theta$. Two other examples are presented below.

EXAMPLE 7.1-4. ***Single Boundary: Oblique TE Wave.*** A wave transmitted through a planar boundary between media of refractive indices n_1 and n_2 at angles θ_1 and θ_2, satisfying Snell's

law ($n_1 \sin\theta_1 = n_2 \sin\theta_2$), is described by an **S** matrix determined from the Fresnel equations (6.2-8) and (6.2-9), and its corresponding **M** matrix:



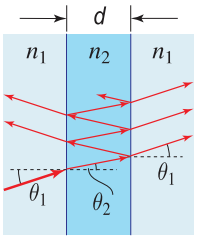
$$\mathbf{S} = \begin{bmatrix} \mathrm{t}_{12} & \mathrm{r}_{21} \\ \mathrm{r}_{12} & \mathrm{t}_{21} \end{bmatrix} = \frac{1}{\tilde{n}_1 + \tilde{n}_2} \begin{bmatrix} 2a_{12}\tilde{n}_1 & \tilde{n}_2 - \tilde{n}_1 \\ \tilde{n}_1 - \tilde{n}_2 & 2a_{21}\tilde{n}_2 \end{bmatrix}, \tag{7.1-23}$$

$$\mathbf{M} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \frac{1}{2a_{21}\tilde{n}_2} \begin{bmatrix} \tilde{n}_1 + \tilde{n}_2 & \tilde{n}_2 - \tilde{n}_1 \\ \tilde{n}_2 - \tilde{n}_1 & \tilde{n}_1 + \tilde{n}_2 \end{bmatrix}. \tag{7.1-24}$$

These expressions are applicable for both TE and TM polarized waves with the following definitions:

$$\begin{aligned} &\text{TE:} \quad \tilde{n}_1 = n_1 \cos\theta_1, \quad \tilde{n}_2 = n_2 \cos\theta_2, \quad a_{12} = a_{21} = 1, \\ &\text{TM:} \quad \tilde{n}_1 = n_1 \sec\theta_1, \quad \tilde{n}_2 = n_2 \sec\theta_2, \quad a_{12} = \cos\theta_1 / \cos\theta_2 = 1/a_{21}. \end{aligned}$$

EXAMPLE 7.1-5. ***Dielectric Slab: Off-Axis Wave.*** We now consider an oblique wave traveling through the system described in Example 7.1-3: a slab of thickness d and refractive index n_2 in a medium of refractive index n_1. The wave-transfer matrix for an oblique wave is a generalization of the on-axis result:



$$\mathbf{M} = \frac{1}{4\tilde{n}_1\tilde{n}_2} \begin{bmatrix} \tilde{n}_1 + \tilde{n}_2 & \tilde{n}_1 - \tilde{n}_2 \\ \tilde{n}_1 - \tilde{n}_2 & \tilde{n}_1 + \tilde{n}_2 \end{bmatrix} \begin{bmatrix} e^{-j\tilde{\varphi}} & 0 \\ 0 & e^{j\tilde{\varphi}} \end{bmatrix} \begin{bmatrix} \tilde{n}_2 + \tilde{n}_1 & \tilde{n}_2 - \tilde{n}_1 \\ \tilde{n}_2 - \tilde{n}_1 & \tilde{n}_2 + \tilde{n}_1 \end{bmatrix}, \tag{7.1-25}$$

where $\tilde{\varphi} = n_2 k_o d \cos\theta_2$, and, as in Example 7.1-4, $\tilde{n}_1 = n_1 \cos\theta_1$ and $\tilde{n}_2 = n_2 \cos\theta_2$ for the TE polarization, and $\tilde{n}_1 = n_1 \sec\theta_1$ and $\tilde{n}_2 = n_2 \sec\theta_2$ for the TM polarization.

The expression for the matrix **M** in (7.1-25) is identical to that provided in (7.1-19), which describes the on-axis system, except that the parameters n_1, n_2, and φ are replaced by the angle- and polarization-dependent parameters $\tilde{n}_1$ and $\tilde{n}_2$, and by the angle-dependent parameter $\tilde{\varphi}$, respectively. Note that the factors a_{12} and a_{21}, which appear in (7.1-24) at each boundary, cancel out since $a_{12}a_{21} = 1$. With these substitutions, the expressions developed in (7.1-22) for the on-axis transmittance and reflectance in Example 7.1-3 generalize to the following off-axis, polarization-dependent formulas:

$$\mathrm{t} = \frac{4\tilde{n}_1\tilde{n}_2 \exp(-j\tilde{\varphi})}{(\tilde{n}_1 + \tilde{n}_2)^2 - (\tilde{n}_1 - \tilde{n}_2)^2 \exp(-j2\tilde{\varphi})}, \quad \mathrm{r} = -j\left[\frac{\tilde{n}_2^2 - \tilde{n}_1^2}{4\tilde{n}_1\tilde{n}_2} \sin\tilde{\varphi}\right] \mathrm{t}. \tag{7.1-26}$$

B. Fabry–Perot Etalon

The Fabry–Perot etalon was introduced in Sec. 2.5B; it is an interferometer comprising two parallel and highly reflective mirrors that transmit light only at a set of specific, uniformly spaced frequencies, which depend on the optical pathlength between the mirrors. It is used both as a filter and as a spectrum analyzer, and is controlled by varying the pathlength, e.g., by moving one of the mirrors with respect to the other. It is also used as an optical resonator, as discussed in Sec. 11.1. In this section, we examine this multilayer device using the matrix methods developed in this chapter.

Mirror Fabry–Perot Etalon

Consider two lossless partially reflective mirrors with amplitude transmittances t_1 and t_2, and amplitude reflectances r_1 and r_2, separated by a distance d filled with a medium of refractive index n. The overall system is described by the matrix product

$$\mathbf{M} = \begin{bmatrix} 1/\mathrm{t}_1^* & \mathrm{r}_1/\mathrm{t}_1 \\ \mathrm{r}_1^*/\mathrm{t}_1^* & 1/\mathrm{t}_1 \end{bmatrix} \begin{bmatrix} \exp(-j\varphi) & 0 \\ 0 & \exp(j\varphi) \end{bmatrix} \begin{bmatrix} 1/\mathrm{t}_2^* & \mathrm{r}_2/\mathrm{t}_2 \\ \mathrm{r}_2^*/\mathrm{t}_2^* & 1/\mathrm{t}_2 \end{bmatrix}, \tag{7.1-27}$$

where $\varphi = nk_o d$. Since the system is lossless and symmetric, $\mathbf{M}$ takes the simplified form provided in (7.1-17) and the amplitude transmittance t is therefore the inverse of the D element of $\mathbf{M}$, so that

$$\mathrm{t} = \frac{\mathrm{t}_1 \mathrm{t}_2 \exp(-j\varphi)}{1 - \mathrm{r}_1 \mathrm{r}_2 \exp(-j2\varphi)}. \tag{7.1-28}$$

This relation may also be derived by direct use of the Airy formula (7.1-8).

As a result, the intensity transmittance of the etalon is

$$\mathcal{T} = |\mathrm{t}|^2 = \frac{|\mathrm{t}_1 \mathrm{t}_2|^2}{|1 - \mathrm{r}_1 \mathrm{r}_2 \exp(-j2\varphi)|^2}. \tag{7.1-29}$$

This expression is similar to (2.5-16) for the intensity of an infinite number of waves with equal phase differences, and with amplitudes that decrease at a geometric rate, as described in Sec. 2.5B. Assuming that $\arg\{\mathrm{r}_1 \mathrm{r}_2\} = 0$, (7.1-29) can be written in the form[†]

$$\mathcal{T} = \frac{\mathcal{T}_{\max}}{1 + (2\mathcal{F}/\pi)^2 \sin^2 \varphi}, \tag{7.1-30}$$

where

$$\mathcal{T}_{\max} = \frac{|\mathrm{t}_1 \mathrm{t}_2|^2}{(1 - |\mathrm{r}_1 \mathrm{r}_2|)^2} = \frac{(1 - |\mathrm{r}_1|^2)(1 - |\mathrm{r}_2|^2)}{(1 - |\mathrm{r}_1 \mathrm{r}_2|)^2} \tag{7.1-31}$$

and

$$\boxed{\mathcal{F} = \frac{\pi \sqrt{|\mathrm{r}_1 \mathrm{r}_2|}}{1 - |\mathrm{r}_1 \mathrm{r}_2|}}. \tag{7.1-32}$$

Finesse

The parameter $\mathcal{F}$, called the **finesse**, is a monotonic increasing function of the reflectance product $\mathrm{r}_1 \mathrm{r}_2$, and is a measure of the quality of the etalon. For example, if $\mathrm{r}_1 \mathrm{r}_2 = 0.99$, then $\mathcal{F} \approx 313$.

As described in Sec. 2.5B, the transmittance $\mathcal{T}$ is a periodic function of φ, now with period π. It reaches its maximum value of $\mathcal{T}_{\max}$, which equals unity if $|\mathrm{r}_1| = |\mathrm{r}_2|$, when φ is an integer multiple of π. When the finesse $\mathcal{F}$ is large (i.e., when $|\mathrm{r}_1 \mathrm{r}_2| \approx 1$), $\mathcal{T}$ becomes a sharply peaked function of φ of approximate width $\pi/\mathcal{F}$. Thus, the higher the finesse $\mathcal{F}$, the sharper the peaks of the transmittance as a function of the phase φ.

[†] Equation (7.1-30), in which $\varphi = nk_o d$, reproduces (2.5-18), in which $\varphi = 2nk_o d$.

The phase $\varphi = nk_o d = (\omega/c)d$ is proportional to the frequency, so that the condition $\varphi = \pi$ corresponds to $\omega = \omega_F$, or $\nu = \nu_F$, where

$$\nu_F = \frac{c}{2d}, \quad \omega_F = \frac{\pi c}{d} \tag{7.1-33}$$

Free Spectral Range

is called the **free spectral range**. It follows that the transmittance as a function of frequency, $\mathcal{T}(\nu)$, is a periodic function of period ν_F,

$$\mathcal{T}(\nu) = \frac{\mathcal{T}_{\max}}{1 + (2\mathcal{F}/\pi)^2 \sin^2(\pi\nu/\nu_F)}, \tag{7.1-34}$$

Transmittance (Fabry–Perot Etalon)

as illustrated in Fig. 7.1-5. It reaches its peak value of $\mathcal{T}_{\max}$ at the resonance frequencies $\nu_q = q\nu_F$, where q is an integer. When the finesse $\mathcal{F} \gg 1$, $\mathcal{T}(\nu)$ drops sharply as the frequency deviates slightly from ν_q, so that $\mathcal{T}(\nu)$ takes the form of a comb-like function. The spectral width of each of these high-transmittance lines is

$$\delta\nu = \frac{\nu_F}{\mathcal{F}}, \tag{7.1-35}$$

i.e., is a factor of $\mathcal{F}$ smaller than the spacing between the resonance frequencies.

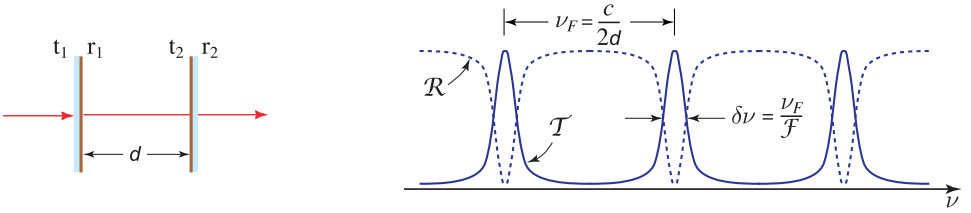


Figure 7.1-5 Intensity transmittance and reflectance, $\mathcal{T}$ and $\mathcal{R} = 1 - \mathcal{T}$, of the Fabry–Perot etalon as a function of frequency ν.

The Fabry–Perot etalon may be used as a sharply tuned optical filter or a spectrum analyzer. Because of the periodic nature of the spectral response, however, the spectral width of the measured light must be narrower than the free spectral range $\nu_F = c/2d$ in order to avoid ambiguity. The filter is tuned (i.e., the resonance frequencies are shifted) by adjusting the distance d between the mirrors. A slight change in mirror spacing Δd shifts the resonance frequency $\nu_q = qc/2d$ by a relatively large amount $\Delta\nu_q = -(qc/2d^2)\Delta d = -\nu_q \Delta d/d$. Although the frequency spacing ν_F also changes, it is by the far smaller amount $-\nu_F \Delta d/d$. As an example, a mirror separation of $d = 1.5$ cm leads to a free spectral range $\nu_F = 10$ GHz when $n = 1$. For a typical optical frequency of $\nu = 10^{14}$ Hz, corresponding to $q = 10^4$, a change of d by a factor of 10^{-4} ($\Delta d = 1.5$ μm) translates the peak frequency by $\Delta\nu_q = 10$ GHz, whereas the free spectral range is altered by only 1 MHz, becoming 9.999 GHz.

Applications of the Fabry–Perot etalon as a resonator are described in Sec. 11.1.

Off-Axis Transmittance of the Fabry–Perot Etalon

For an oblique wave traveling at an angle θ with the axis of a mirror etalon, the amplitude transmittance is given by (7.1-28) with the phase φ replaced by $\widetilde{\varphi} = nk_o d\cos\theta$. It follows that the intensity transmittance in (7.1-34) is generalized to

$$\mathcal{T}(\nu) = \frac{\mathcal{T}_{\max}}{1 + (2\mathcal{F}/\pi)^2 \sin^2(\pi\cos\theta\,\nu/\nu_F)} \tag{7.1-36}$$

in the off-axis case.

Maximum transmittance occurs at frequencies for which

$$\boxed{\nu = q\,\nu_F \sec\theta\,,} \qquad q = 1, 2, \ldots, \qquad \nu_F = c/2d. \tag{7.1-37}$$

Resonance Condition

If the finesse of the etalon is large, transmission occurs at these frequencies and is almost completely blocked at all other frequencies. The plot of this relation provided in Fig. 7.1-6(c) shows that at each angle θ only a set of discrete frequencies are transmitted. Likewise, a wave at frequency ν is transmitted at only a set of angles, so that a cone of incident broad-spectrum (white) light creates a set of concentric rings spread like a rainbow, as illustrated in Fig. 7.1-6(b). For incident light with a spectral width smaller than the free spectral range ν_F, each frequency component corresponds to one and only one angle, so that the etalon can be used as a spectrum analyzer.

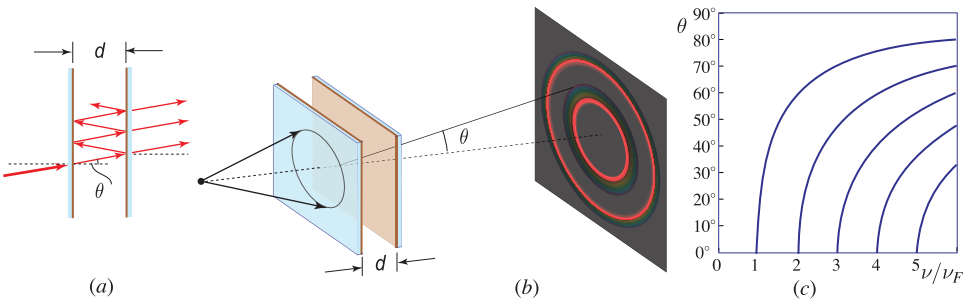


Figure 7.1-6 (a) An off-axis wave transmitted through a mirror Fabry–Perot etalon. (b) White light from a point source transmitted through the etalon creates a set of concentric rings of different frequencies (colors). (c) Frequencies and angles that satisfy the condition of peak transmittance, as set forth in (7.1-37).

EXAMPLE 7.1-6. ***The Dielectric-Slab Beamsplitter.*** The transmittance and reflectance of a dielectric slab of width d and refractive index n_2 surrounded by a medium of refractive index n_1 are given in (7.1-26). If the wave is incident at an angle θ_1, then for TE polarization we have $\widetilde{n}_1 = n_1\cos\theta_1$, $\widetilde{n}_2 = n_2\cos\theta_2$, and $\widetilde{\varphi} = n_2 k_o d\cos\theta_2$, as in Example 7.1-4, with $\sin\theta_2 = (n_1/n_2)\sin\theta_1$. This expression reproduces (7.1-28) for the Fabry–Perot etalon if we substitute $\mathrm{t}_1\mathrm{t}_2 = 4\widetilde{n}_1\widetilde{n}_2/(\widetilde{n}_1 + \widetilde{n}_2)$ and $\mathrm{r}_1\mathrm{r}_2 = (\widetilde{n}_1 - \widetilde{n}_2)^2/(\widetilde{n}_1 + \widetilde{n}_2)^2$. It follows that the expressions for the intensity transmittance of the mirror etalon, (7.1-30) and (7.1-34), are also applicable for the dielectric slab. Using (7.1-32), the finesse of the slab is thus given by

$$\mathcal{F} = \frac{\pi}{4}\,\frac{|\widetilde{n}_2^2 - \widetilde{n}_1^2|}{\widetilde{n}_1\widetilde{n}_2}\,. \tag{7.1-38}$$

Large values of $\mathcal{F}$ are typically not obtained in slab etalons. As an example, for $n_1 = 1.5$ (the refractive index of SiO_2) and $n_2 = 3.5$ (the refractive index of Si), and an angle of incidence $\theta_1 = 45°$, the finesse $\mathcal{F} = 1.89$. As illustrated in Fig. 7.1-7, the dependence of $\mathcal{T}$ and $\mathcal{R}$ on the phase $\widetilde{\varphi}$, which is proportional to the frequency, does not exhibit the sharp peaks observed in etalons with highly reflective mirrors (see, e.g., the illustration in Fig. 7.1-5). Higher values of $\mathcal{F}$ are obtained by coating the surfaces of the slab to enhance internal reflection.

The dielectric slab may be used as a beamsplitter. If the slab width $d = 1$ mm, for example, the range between two consecutive transmittance peaks $\widetilde{\varphi} = n_2 k_o d \cos\theta_2 = \pi$ corresponds to a frequency of ≈ 45 GHz. As illustrated in Fig. 7.1-7, the transmittance and reflectance near the center of this interval are reasonably flat. Since beamsplitters are routinely used in interferometers, it is important to be cognizant of the relation between the phases of the reflected and transmitted waves. As illustrated in the figure, the relative phase between the reflected and transmitted waves is always $\pm\pi/2$, with the sign changing at points of peak transmittance.

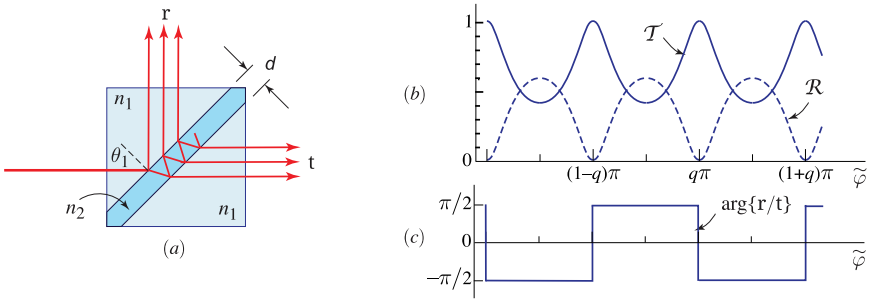


Figure 7.1-7 (a) A dielectric slab used as a beamsplitter. (b) Dependence of the intensity transmittance and reflectance ($\mathcal{T}$ and $\mathcal{R}$, respectively) on the phase $\widetilde{\varphi} = n_2 k_o d \cos\theta_2$ for $n_1 = 1.5$ (refractive index of SiO_2), $n_2 = 3.5$ (refractive index of Si), and an angle of incidence $\theta_1 = 45°$. Transmission peaks and reflectance minima occur at $\widetilde{\varphi} = q\pi$, where q is an integer. (c) The relative phase between the reflected and transmitted waves is $\pm\pi/2$; the sign switches at $\widetilde{\varphi} = q\pi$.

C. Bragg Grating

The Bragg grating was introduced in Exercise 2.5-3 as a set of uniformly spaced parallel partially reflective planar mirrors. Such a structure has angular and frequency selectivity that is useful in many applications. In this section, we generalize the definition of the Bragg grating to include a set of N uniformly spaced identical multilayer segments, and develop a theory for light reflection based on matrix wave optics. Devices fabricated according to this prescription include **distributed Bragg reflectors (DBRs)** and **fiber Bragg gratings (FBGs)**, which are often used in resonators and lasers.

Simplified Theory

The reflectance of the Bragg grating was determined in Exercise 2.5-3 under two assumptions: (1) the mirrors are weakly reflective so that the incident wave is not depleted as it propagates; and (2) secondary reflections (i.e., reflections of the reflected waves) are negligible. In this approximation, the reflectance $\mathcal{R}_N$ of an N-mirror grating is related to the reflectance $\mathcal{R}$ of a single mirror by the relation[†]

$$\mathcal{R}_N = \frac{\sin^2 N\varphi}{\sin^2\varphi}\,\mathcal{R}. \tag{7.1-39}$$

[†] In Exercise 2.5-3, the quantity φ denotes the phase between successive phasors whereas here that phase is denoted 2φ since it represents a round trip.

As described in Sec. 2.5B, the factor $\sin^2 N\varphi/\sin^2\varphi$ represents the intensity of the sum of N phasors of unit amplitude and phase difference 2φ. This function has a peak value of N^2 when the Bragg condition is satisfied, i.e., when 2φ equals $q2\pi$, where $q = 0, 1, 2, \ldots$. It drops away from these values sharply, with a width that is inversely proportional to N. In this simplified model, the intensity of the total reflected wave is, at most, a factor of N^2 greater than the intensity of the wave reflected from a single segment.

For a Bragg grating comprising partially reflective mirrors separated from each other by a distance Λ and a round-trip phase $2\varphi = 2k\Lambda\cos\theta$, where θ is the angle of incidence. Therefore, maximum reflection occurs when $2k\Lambda\cos\theta = 2q\pi$ or

$$\cos\theta = q\frac{\lambda}{2\Lambda} = q\frac{\omega_{\mathcal{B}}}{\omega} = q\frac{\nu_{\mathcal{B}}}{\nu}, \tag{7.1-40}$$

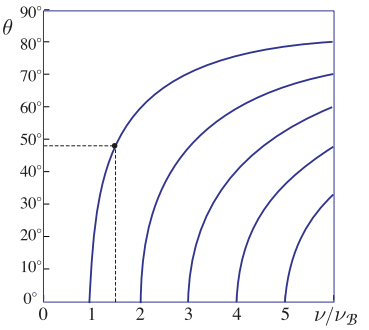
Bragg Condition

where

$$\nu_{\mathcal{B}} = \frac{c}{2\Lambda}, \quad \omega_{\mathcal{B}} = \frac{\pi c}{\Lambda}, \tag{7.1-41}$$

Bragg Frequency

is the Bragg frequency.



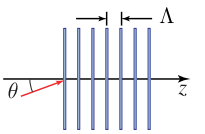


Figure 7.1-8 Locus of frequencies ν and angles θ at which the Bragg condition is satisfied. For example, if $\nu = 1.5\,\nu_{\mathcal{B}}$ (dot-dash line), we have $\theta = 48.2°$. This corresponds to a Bragg angle $\theta_{\mathcal{B}} = 41.8°$ (measured from the plane of the grating.)

At normal incidence ($\theta = 0°$), peak reflectance occurs at frequencies that are integer multiples of the Bragg frequency, i.e., $\nu = q\nu_{\mathcal{B}}$. At frequencies such that $\nu < \nu_{\mathcal{B}}$, the Bragg condition cannot be satisfied at any angle. At frequencies $\nu_{\mathcal{B}} < \nu < 2\nu_{\mathcal{B}}$, the Bragg condition is satisfied at one angle $\theta = \cos^{-1}(\lambda/2\Lambda) = \cos^{-1}(\nu_{\mathcal{B}}/\nu)$. The complement of this angle, $\theta_{\mathcal{B}} = \pi/2 - \theta$, is the Bragg angle (see (2.5-13) and Fig. 2.5-8),

$$\theta_{\mathcal{B}} = \sin^{-1}(\lambda/2\Lambda). \tag{7.1-42}$$

Bragg Angle

At frequencies $\nu \geq 2\nu_{\mathcal{B}}$, the Bragg condition is satisfied at more than one angle. Figure 7.1-8 illustrates the spectral and angular dependence of reflections from a Bragg grating, based on the simplified theory.

Matrix Theory

We now use the matrix approach introduced in the previous section to develop an exact theory of Bragg reflection that includes multiple transmissions and reflections, as well as depletion of the incident wave. It turns out that the collaborative effects of the reflections, and the reflections of reflections, can lead to enhancement of the total reflected wave, and a phenomenon whereby total reflection occurs not only at single frequencies that are multiples of $\nu_B/\cos\theta$, but over extended spectral bands surrounding these frequencies!

Consider a grating comprising a stack of N identical generic segments (Fig. 7.1-9), each described by a unimodular wave-transfer matrix $\mathbf{M}_o$ satisfying the conservation relations for a lossless, symmetrical system, so that

$$\mathbf{M}_o = \begin{bmatrix} 1/\mathrm{t}^* & \mathrm{r}/\mathrm{t} \\ \mathrm{r}^*/\mathrm{t}^* & 1/\mathrm{t} \end{bmatrix}, \tag{7.1-43}$$

where t and r are complex amplitude transmittance and reflectance satisfying the conditions set forth in (7.1-15), and $\mathcal{T} = |\mathrm{t}|^2$ and $\mathcal{R} = |\mathrm{r}|^2$ are the corresponding intensity transmittance and reflectance.

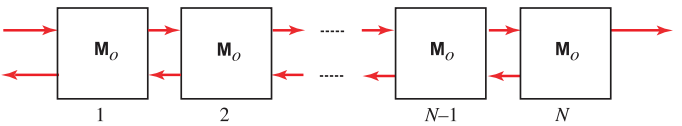


Figure 7.1-9 Bragg grating made of N segments, each of which is described by a matrix $\mathbf{M}_o$.

In accordance with (7.1-2), the wave-transfer matrix $\mathbf{M}$ for the N segments is simply the product $\mathbf{M} = \mathbf{M}_o^N$. Since $\mathbf{M}_o$ is unimodular, i.e., $\det \mathbf{M}_o = 1$, it satisfies the property

$$\mathbf{M}_o^N = \Psi_N \mathbf{M}_o - \Psi_{N-1}\mathbf{I}, \tag{7.1-44}$$

where

$$\boxed{\Psi_N = \frac{\sin N\Phi}{\sin \Phi},} \tag{7.1-45}$$

$$\boxed{\cos\Phi = \mathrm{Re}\{1/\mathrm{t}\},} \tag{7.1-46}$$

and $\mathbf{I}$ is the identity matrix. Equation (7.1-44) may be proved by induction (i.e., demonstrating that this relation is valid for N segments if it is valid for $N-1$ segments; this may be done by direct substitution with the help of trigonometric identities).

Since the N-segment system is also lossless and symmetric, its matrix may be written in the form

$$\mathbf{M}_o^N = \begin{bmatrix} 1/\mathrm{t}_N^* & \mathrm{r}_N/\mathrm{t}_N \\ \mathrm{r}_N^*/\mathrm{t}_N^* & 1/\mathrm{t}_N \end{bmatrix}, \tag{7.1-47}$$

where t_N and r_N are the N-segment amplitude transmittance and reflectance, respectively. Substituting from (7.1-43) and (7.1-47) into (7.1-44), and comparing the diagonal and off-diagonal elements of the matrices on both sides of the equation, leads to

$$\frac{1}{\mathrm{t}_N} = \Psi_N \frac{1}{\mathrm{t}} - \Psi_{N-1} \tag{7.1-48}$$

$$\frac{\mathrm{r}_N}{\mathrm{t}_N} = \Psi_N \frac{\mathrm{r}}{\mathrm{t}}. \tag{7.1-49}$$

These two equations define t_N and r_N in terms of t and r.

The intensity transmittance $\mathcal{T}_N = |\mathrm{t}_N|^2$ is obtained by taking the absolute-squared value of (7.1-49) and using the relation $\mathcal{R} = 1 - \mathcal{T}$,

$$\mathcal{T}_N = \frac{\mathcal{T}}{\mathcal{T} + \Psi_N^2(1 - \mathcal{T})}. \tag{7.1-50}$$

It follows that the power reflectance $\mathcal{R}_N = 1 - \mathcal{T}_N$ is given by

$$\boxed{\mathcal{R}_N = \frac{\Psi_N^2 \mathcal{R}}{1 - \mathcal{R} + \Psi_N^2 \mathcal{R}}.} \tag{7.1-51}$$

Bragg-Grating Reflectance

Summary

The reflectance $\mathcal{R}_N$ of a medium comprising N identical segments is related to the single-segment reflectance $\mathcal{R}$ by the nonlinear relation (7.1-51), which contains a factor Ψ_N that results from the interference effects associated with collective reflections from the N segments of the grating. Defined by (7.1-45), Ψ_N depends on the number of segments N and on an additional parameter Φ that is related to the single-segment complex amplitude transmittance t via (7.1-46).

The dependence of $\mathcal{R}_N$ on $\mathcal{R}$, described by (7.1-51), takes simpler forms in certain limits. If the single-segment reflectance is very small, i.e., $\mathcal{R} \ll 1$, and if Ψ_N^2 is not too large so that $\Psi_N^2 \mathcal{R} \ll 1$, then (7.1-51) may be approximated by:

$$\mathcal{R}_N \approx \Psi_N^2 \mathcal{R} = \frac{\sin^2 N\Phi}{\sin^2 \Phi} \mathcal{R}. \tag{7.1-52}$$

This relation is now similar in form to the approximate relation (7.1-39), with Φ playing the role of the phase φ.

In the opposite limit for which $\Psi_N^2 \gg 1$, the reflectance $\mathcal{R}_N \approx \Psi_N^2 \mathcal{R}/(1 + \Psi_N^2 \mathcal{R})$. This nonlinear relation between $\mathcal{R}_N$ and $\mathcal{R}$ exhibits saturation and is typical of systems with feedback, which in this case results from multiple internal reflections at the segment boundaries. Ultimately, if $\Psi_N^2 \mathcal{R} \gg 1$, then $\mathcal{R}_N$ approaches its maximum value of unity, so that the N-segment device acts as perfect mirror even though the single segment is only partially reflective. A large interference factor Ψ_N accelerates the rise of $\mathcal{R}_N$ to unity as $\mathcal{R}$ increases.

The interference factor Ψ_N, which depends on $\Phi = \cos^{-1}(\mathrm{Re}\{1/\mathrm{t}\})$ via (7.1-45), has two distinct regimes: (1) a normal regime for which Φ is real and the grating exhibits partial reflection/transmission (including zero reflection, or total transmission);

and (2) an anomalous regime for which Φ is complex and Ψ_N can be extremely large, corresponding to total reflection.

Partial- and Zero-Reflection Regime

This regime is defined by the condition $|\mathrm{Re}\{1/\mathrm{t}\}| \leqslant 1$, which ensures that $\Phi = \cos^{-1}(\mathrm{Re}\{1/\mathrm{t}\})$ is real. In this case, $\mathcal{R}_N$ depends on $\mathcal{R}$ and Ψ_N in accordance with (7.1-45) and (7.1-51). Maximum reflectance occurs when Ψ_N has its maximum value of N. In this case, $\mathcal{R}_N = N^2\mathcal{R}/(1 - \mathcal{R} + N^2\mathcal{R})$. Therefore, $\mathcal{R}_N$ cannot exactly equal unity unless $\mathcal{R} = 1$, exactly. For example, for $N = 10$, if $\mathcal{R} = 0.5$, then the maximum value of $\mathcal{R}_N \approx 0.99$.

Zero reflectance, or **total transmittance**, is possible, even if the reflectance R of the individual segment is substantial. This occurs when $\Psi_N = 0$, i.e., when $\sin N\Phi = 0$, or $\Phi = q\pi/N$ for $q = 0, 1, \ldots, N-1$. The N frequencies at which this complete transparency occurs are resonance frequencies of the grating. The phenomenon represents some form of tunneling through the individually reflective segments.

Total-Reflection Regime

In this regime, $|\mathrm{Re}\{1/\mathrm{t}\}| = |\cos\Phi| > 1$ so that Φ is a complex variable $\Phi = \Phi_\mathrm{R} + j\Phi_\mathrm{I}$. Using the identity $\cos(\Phi_\mathrm{R} + j\Phi_\mathrm{I}) = \cos\Phi_\mathrm{R}\cosh\Phi_\mathrm{I} - j\sin\Phi_\mathrm{R}\sinh\Phi_\mathrm{I}$, and equating the real and imaginary parts of both sides of (7.1-46), we obtain $\sin\Phi_\mathrm{R} = 0$ so that $\Phi_\mathrm{R} = m\pi$ and $\cos\Phi_\mathrm{R} = +1$, or -1, when m is an even or odd integer, respectively, which results in

$$\cosh\Phi_\mathrm{I} = |\mathrm{Re}\{1/\mathrm{t}\}|. \qquad \text{(7.1-53)}$$
Total-Reflection Regime

The factor $\Psi_N = \sin N\Phi/\sin\Phi$ then becomes

$$\Psi_N = \pm\frac{\sinh N\Phi_\mathrm{I}}{\sinh\Phi_\mathrm{I}}, \qquad \text{(7.1-54)}$$
Total-Reflection Regime

where the $\pm$ sign is the sign of the factor $\cos(Nm\pi)/\cos(m\pi)$. Since $\sinh(\cdot)$ increases exponentially with N for large N, $|\Psi_N|$ can be much greater than N. In this case, in accordance with (7.1-51), the reflectance $\mathcal{R}_N \approx 1$ and the grating acts as a total reflector. The forward waves become evanescent and do not penetrate the multisegment medium, much as occurs with total internal reflection.

Because Φ depends on t, which depends on the frequency ν, the two regimes correspond to distinct spectral bands, as illustrated in the following examples. The spectral bands associated with the total-reflection regime are called **stop bands** since they represent bands within which light transmission is almost completely blocked. The other regime corresponds to **passbands**. Total transmission (zero reflection) occurs at specific resonance frequencies within the passbands.

EXAMPLE 7.1-7. ***Stack of Partially Reflective Mirrors.*** Consider a grating made of a stack of N identical partially reflective mirrors (beamsplitters) that are mutually separated by a distance Λ and embedded in a homogeneous medium of refractive index n, as illustrated in Fig. 7.1-10(a). A single segment comprises a distance Λ in a homogeneous medium, followed by a partially reflective mirror of amplitude transmittance t and amplitude reflectance r.

The wave-transfer matrix $\mathbf{M}_o$ for this segment is determined by multiplying the matrix in (7.1-18) by the matrix in (7.1-4):

$$\mathbf{M}_o = \frac{1}{|\mathrm{t}|}\begin{bmatrix} e^{-j\varphi} & j|\mathrm{r}|e^{j\varphi} \\ -j|\mathrm{r}|e^{-j\varphi} & e^{j\varphi} \end{bmatrix}, \quad \varphi = nk_o\Lambda = \pi\nu/\nu_{\mathcal{B}}, \tag{7.1-55}$$

where $\nu_{\mathcal{B}} = c/2\Lambda$ is the Bragg frequency. This provides $\mathrm{t} = |\mathrm{t}|e^{j\varphi}$, and therefore Φ via

$$\cos\Phi = \frac{1}{|\mathrm{t}|}\cos\varphi \qquad \text{for} \quad |\cos\varphi| \le |\mathrm{t}|, \tag{7.1-56}$$

$$\cosh\Phi_{\mathrm{I}} = \frac{1}{|\mathrm{t}|}|\cos\varphi| \quad \text{for} \quad |\cos\varphi| > |\mathrm{t}|. \tag{7.1-57}$$

The relationships between Φ and φ, and between Φ_{I} and φ, are nonlinear and unusual, as illustrated in Fig. 7.1-10(*b*). The corresponding dependence of the power reflectance $\mathcal{R}_N$ on φ is shown in Fig. 7.1-10(*c*). In the normal regime (indicated by the shaded regions), Φ is real and the reflectance exhibits multiple peaks with zeros between. None of the peaks approaches unity, despite the fact that Ψ_N reaches a maximum value of $N = 10$.

The situation is quite different in the total-reflection regime (unshaded regions), where Φ is complex. The factor Ψ_N reaches a value ≈ 3000 at the center of the band ($\varphi = \pi$) when $|\mathrm{t}|^2 = 0.5$. These regions represent ranges of φ where total reflection occurs ($\mathcal{R}_N \approx 1$). Since φ is proportional to the frequency ν, Fig. 7.1-10(*c*) is actually a display of the spectral reflectance, and the unshaded regions correspond to the stop bands.

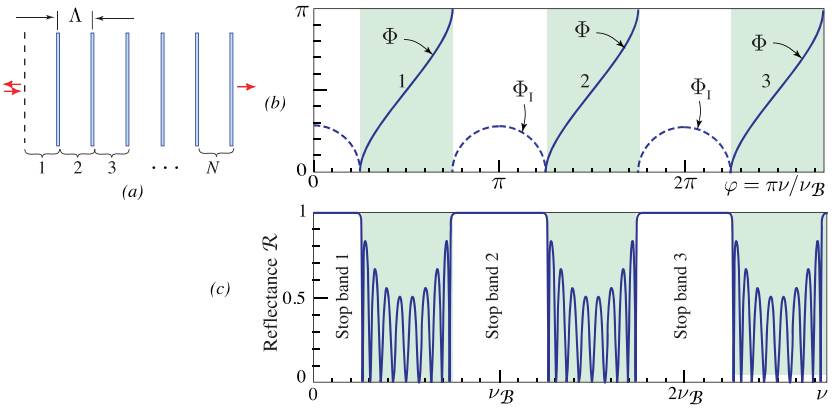


Figure 7.1-10 (*a*) Bragg grating comprising $N = 10$ identical mirrors, each with a power reflectance $|\mathrm{r}|^2 = 0.5$. (*b*) Dependence of Φ on the inter-mirror phase delay $\varphi = nk_o\Lambda$. Within the shaded regions, Φ is complex and its imaginary part Φ_{I} is represented by the dashed curves. (*c*) Reflectance $\mathcal{R}$ as a function of frequency (in units of the Bragg frequency $\nu_{\mathcal{B}} = c/2\Lambda$) . Within the stop bands, the reflectance is approximately unity.

EXAMPLE 7.1-8. ***Dielectric Bragg Grating.*** A grating is made of N identical dielectric layers of refractive index n_2, each of width d_2, buried in a medium of refractive index n_1 and separated by a distance d_1, as illustrated in Fig. 7.1-11. This multisegment system is a stack of N identical double layers, each of the type described in Example 7.1-3. The $A = 1/\mathrm{t}^*$ element of the wave-transfer matrix $\mathbf{M}_o$ is given by (7.1-20), from which

$$\mathrm{Re}\left\{\frac{1}{\mathrm{t}}\right\} = \frac{(n_1+n_2)^2}{4n_1n_2}\cos(\varphi_1+\varphi_2) - \frac{(n_2-n_1)^2}{4n_1n_2}\cos(\varphi_1-\varphi_2), \tag{7.1-58}$$

where $\varphi_1 = n_1k_od_1$ and $\varphi_2 = n_2k_od_2$ are the phases introduced by the two layers of a segment. This result can be used in conjunction with (7.1-45), (7.1-46), (7.1-51), (7.1-53), and (7.1-54) to determine the reflectance of the grating.

The spectral dependence of the reflectance can be computed as a function of ν by noting that $\varphi_1 + \varphi_2 = k_o(n_1d_1 + n_2d_2) = \pi\nu/\nu_{\mathcal{B}}$, where $\nu_{\mathcal{B}} = (c_o/\bar{n})/2\Lambda$, and $\bar{n} = (n_1d_1 + n_2d_2)/\Lambda$ is

the average refractive index. The Bragg frequency ν_B is the frequency at which the single-segment round-trip phase $2k_o(n_1d_1 + n_2d_2) = 2\pi$. The phase difference $\varphi_1 - \varphi_2 = \zeta\pi\nu/\nu_B$, with $\zeta = (n_1d_1 - n_2d_2)/(n_1d_1 + n_2d_2)$, is also proportional to the frequency. Figure 7.1-11(*b*) provides an example of the spectral reflectance as a function of ν.

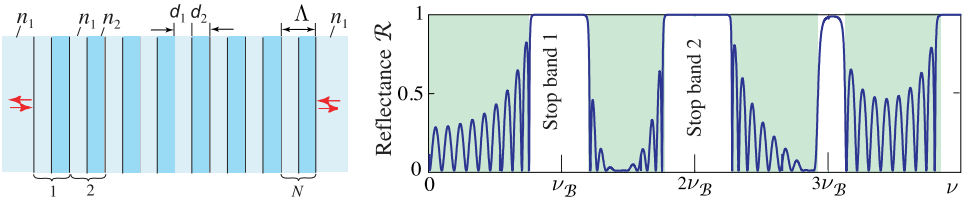


Figure 7.1-11 Power reflectance as a function of frequency for a dielectric Bragg grating comprising $N = 10$ segments, each of which has two layers of thickness $d_1 = d_2$ and refractive indices $n_1 = 1.5$ and $n_2 = 3.5$. The grating is placed in a medium with matching refractive index n_1. The reflectance is approximately unity within the stop bands centered about multiples of $\nu_B = c/2\Lambda$, where $c = c_o/\bar{n}$ and $\bar{n}$ is the mean refractive index.

EXAMPLE 7.1-9. ***Dielectric Bragg Grating: Oblique Incidence.*** The results in Example 7.1-8 may be generalized to oblique waves with angle of incidence θ_1 in medium 1, corresponding to angle θ_2 in layer 2, where $n_1 \sin\theta_1 = n_2 \sin\theta_2$. In this case, (7.1-58) becomes

$$\text{Re}\left\{\frac{1}{\text{t}}\right\} = \frac{(\widetilde{n}_1 + \widetilde{n}_2)^2}{4\widetilde{n}_1\widetilde{n}_2}\cos(\widetilde{\varphi}_1 + \widetilde{\varphi}_2) - \frac{(\widetilde{n}_2 - \widetilde{n}_1)^2}{4\widetilde{n}_1\widetilde{n}_2}\cos(\widetilde{\varphi}_1 - \widetilde{\varphi}_2), \tag{7.1-59}$$

where $\widetilde{\varphi}_1 = n_1k_od_1\cos\theta_1$ and $\widetilde{\varphi}_2 = n_2k_od_2\cos\theta_2$; $\widetilde{n}_1 = n_1\cos\theta_1$ and $\widetilde{n}_2 = n_2\cos\theta_2$ for TE polarization; and $\widetilde{n}_1 = n_1\sec\theta_1$ and $\widetilde{n}_2 = n_2\sec\theta_2$ for TM polarization. This relation may be used to compute the spectral reflectance at any angle of incidence. Figure 7.1-12 illustrates the dependence of the power reflectance R_N on frequency and the angle of incidence for both TE and TM polarization for a high- contrast grating. The range of angles over which unity reflectance obtains increases with increasing refractive-index contrast ratio n_2/n_1.

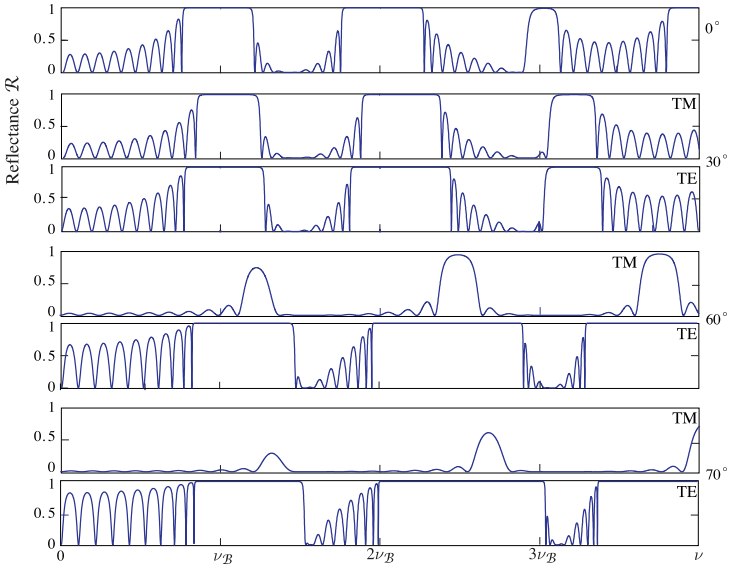


Figure 7.1-12 Spectral dependence of the reflectance $\mathcal{R}$ for the 10-segment dielectric Bragg grating displayed in Fig. 7.1-11, at various angles of incidence θ_1 and for TE and TM polarizations.

Bragg Grating in an Unmatched Medium

In the previous analysis, the Bragg grating was assumed to be made of N identical segments. If each segment is made of multiple dielectric layers, this requires that grating be placed in a matched medium, i.e., a medium with a refractive index equal to that of the front layer, so that the incident light undergoes no additional reflection at the front boundary, and reflects at the back boundary as if it were entering another layer of the grating. The device described in Example 7.1-8 meets this condition.

In most applications, the grating is placed in an unmatched medium, such as air, and boundary effects must be accounted for. This may be accomplished by writing the wave-transfer matrix $\mathbf{M}$ of the composite system, including all boundaries, and finding the corresponding scattering matrix $\mathbf{S}$ by use of the conversion relation. The reflectance of the composite system may be readily determined from $\mathbf{S}$.

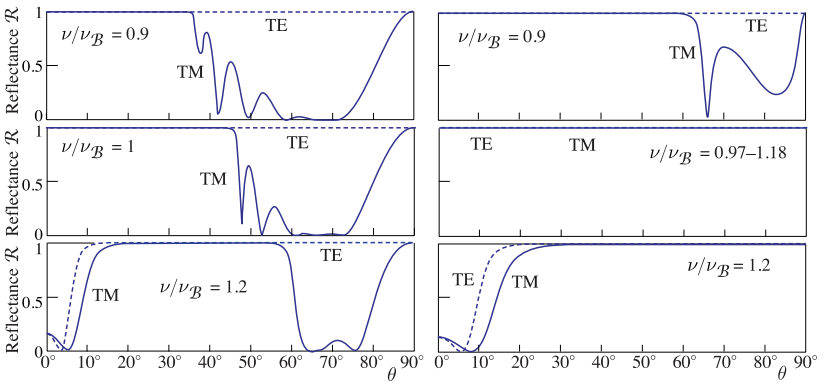
If $\mathbf{M}_o^N$ is the wave-transfer matrix of an N-segment grating in a medium matched to the front layer, then the overall wave transfer function takes the form

$$\mathbf{M} = \mathbf{M}_e \mathbf{M}_o^{N-1} \mathbf{M}_i, \tag{7.1-60}$$

where $\mathbf{M}_i$ is the wave-transfer matrix of the entrance boundary, and $\mathbf{M}_e$ is the wave-transfer matrix of the Nth segment with a boundary into the unmatched medium.

EXAMPLE 7.1-10. ***Reflectance of a Dielectric Bragg Grating in an Unmatched Medium.*** An N-segment Bragg grating is made of alternating layers of refractive indices n_1 and n_2, and widths d_1 and d_2, placed in a medium of refractive index n_0. We wish to determine the reflectance for a wave incident at an angle θ_0 in the external medium, corresponding to angles θ_1 and θ_2 in the first and second layer of each segment, as determined by Snell's law ($n_1 \sin\theta_1 = n_2 \sin\theta_2$).

In this case, (7.1-60) may be used with the following wave-transfer matrices: (1) $\mathbf{M}_i$ represents a boundary between media of refractive indices n_0 and n_1, as described in Example 7.1-4; (2) $\mathbf{M}_o$ represents a single segment of the grating, as described in Example 7.1-5; (3) $\mathbf{M}_e$ represents propagation a distance d_1 in a medium with refractive index n_1 followed by a slab of width d_2 and refractive index n_2, with boundary into a medium of refractive index n_0. Once the $\mathbf{M}$ matrix is determined, we use the conversion relation (7.1-6) to determine the corresponding scattering matrix $\mathbf{S}$. The overall reflectance is the element r_{12} in (7.1-4).



(*a*) Grating in matched medium (*b*) Grating in air

Figure 7.1-13 Power reflectance as a function of the angle of incidence θ at fixed frequencies for the grating described in Fig. 7.1-11. (*a*) Grating is placed in a matched medium ($n = n_1$). (*b*) Grating is placed in air ($n = 1$). In air, the grating has unity reflectance at all angles, for both TE and TM polarizations, at frequencies in the band $0.97\nu_{\mathcal{B}}$ to $1.18\,\nu_{\mathcal{B}}$.

7.2 ONE-DIMENSIONAL PHOTONIC CRYSTALS

One-dimensional (1D) photonic crystals are dielectric structures whose optical properties vary periodically in one direction, called the axis of periodicity, and are constant in the orthogonal directions. These structures exhibit unique optical properties, particularly when the period is of the same order of magnitude as the wavelength. If the axis of periodicity is taken to be the z direction, then optical parameters such as the permittivity $\epsilon(z)$ and the impermeability $\eta(z) = \epsilon_o/\epsilon(z)$ are periodic functions of z, satisfying

$$\eta(z + \Lambda) = \eta(z), \tag{7.2-1}$$

for all z, where Λ is the period. Wave propagation in such periodic media may be studied by solving the generalized Helmholtz equations (7.0-2), for periodic $\eta(z)$.

For an on-axis wave traveling along the z axis and polarized in the x direction, the electric and the magnetic field components E_x and H_y are functions of z, independent of x and y, so that (7.0-2) becomes

$$-\frac{d}{dz}\left[\eta(z)\frac{d}{dz}\right]H_y = \frac{\omega^2}{c_o^2}\,H_y. \tag{7.2-2}$$

For an off-axis wave, i.e., a wave traveling in an arbitrary direction in the x–z plane, the generalized Helmholtz equation has a more complex form. For example, for a TM-polarized off-axis wave, the magnetic field points in the y direction and (7.0-2) gives:

$$\left\{-\frac{\partial}{\partial z}\left[\eta(z)\frac{\partial}{\partial z}\right] + \eta(z)\frac{\partial^2}{\partial x^2}\right\}H_y = \frac{\omega^2}{c_o^2}\,H_y. \tag{7.2-3}$$

Note that (7.2-2) and (7.2-3) are cast in the form of an eigenvalue problem from which the modes $H_y(x, z)$ can be determined.

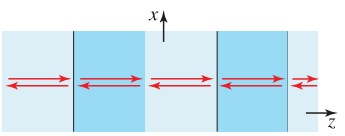
Before embarking on finding solutions to these eigenvalue problems, we first examine the conditions imposed on the propagating modes by the translational symmetry associated with the periodicity.

A. Bloch Modes

Consider first a *homogeneous medium*, which is invariant to an arbitrary translation of the coordinate system. For this medium, an optical mode is a wave that is unaltered by such a translation; it changes only by a multiplicative constant of unity magnitude (a phase factor). The plane wave $\exp(-jkz)$ is such a mode since, upon translation by a distance d, it becomes $\exp[-jk(z + d)] = \exp(-jkd)\exp(-jkz)$. The phase factor $\exp(-jkd)$ is the eigenvalue of the translation operation, as discussed in Appendix C.

On-Axis Bloch Modes

Consider now a *1D periodic medium*, which is invariant to translation by the distance Λ along the axis of periodicity. Its optical modes are waves that maintain their form upon such translation, changing only by a phase factor. As explained in Appendix C, these modes must have the form



$$U(z) = p_K(z)\exp(-jKz), \tag{7.2-4}$$

Bloch Mode

where U represents any of the field components E_x, E_y, H_x, or H_y; K is a constant, and $p_K(z)$ is a periodic function of period Λ. This form satisfies the condition that a translation Λ alters the wave by only a phase factor $\exp(-jK\Lambda)$ since the periodic function is unaltered by such translation. This optical wave is known as a **Bloch mode**, and the parameter K, which specifies the mode and its associated periodic function $p_K(z)$, is called the **Bloch wavenumber**.

The Bloch mode is thus a plane wave $\exp(-jKz)$ with propagation constant K, modulated by a periodic function $p_K(z)$, which has the character of a standing wave, as illustrated by its real part displayed in Fig. 7.2-1(a). Since a periodic function of period Λ can be expanded in a Fourier series as a superposition of harmonic functions of the form $\exp(-jmgz)$, $m = 0, \pm1, \pm2, \ldots$, with

$$\boxed{g = 2\pi/\Lambda,} \tag{7.2-5}$$

it follows that the Bloch wave is a superposition of plane waves of multiple spatial frequencies $K + mg$. The fundamental spatial frequency g of the periodic structure and its harmonics mg, added to the Bloch wavenumber K, constitute the spatial spectrum of the Bloch wave, as shown in Fig. 7.2-1(b). The spatial frequency shift introduced by the periodic medium is analogous to the temporal frequency (Doppler) shift introduced by reflection from a moving object.

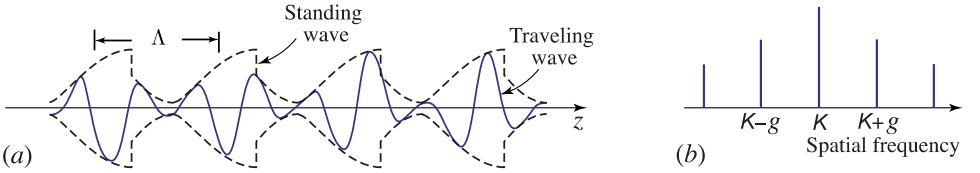
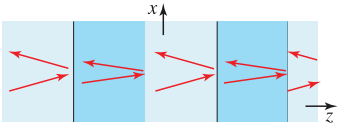


Figure 7.2-1 (a) The Bloch mode. (b) Spatial spectrum of the Bloch mode.

Two modes with Bloch wavenumbers K and $K' = K + g$ are equivalent since they correspond to the same phase factor, $\exp(-jK'\Lambda) = \exp(-jK\Lambda)\exp(-j2\pi) = \exp(-jK\Lambda)$. This is also evident since the factor $\exp(-jg\,z)$ is itself periodic and can be lumped with the periodic function $p_K(z)$. Therefore, for a complete specification of all modes, we need only consider values of K in a spatial-frequency interval of width $g = 2\pi/\Lambda$. The interval $[-g/2, g/2] = [-\pi/\Lambda, \pi/\Lambda]$, known as the first **Brillouin zone**, is a commonly used construct.

Off-Axis Bloch Modes

Off-axis optical modes traveling at some angles in the x–z plane assume the Bloch form



$$U(x, y, z) = p_K(z)\exp(-jKz)\exp(-jk_x x). \tag{7.2-6}$$

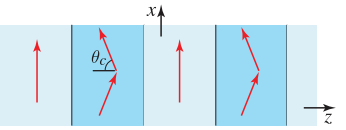
Off-Axis Bloch Mode

The uniformity of the medium in the x direction constrains the x dependence of the optical mode to the harmonic form $\exp(-jk_x x)$, posing no other restriction on the transverse component k_x of the wavevector. At a location where the refractive index

is n, $k_x = nk_o \sin\theta$, where θ is the inclination angle of the wave with respect to the z axis. As the wave travels through the various layers of the inhomogeneous medium, this angle changes, but in view of Snell's law, $n\sin\theta$ and k_x are unaltered.

Normal-to-Axis Bloch Modes

When the angle of incidence in the densest medium is greater than the critical angle, the modes do not travel along the axis of periodicity (the z direction). Rather, they are normal-to-axis modes traveling along the lateral x direction that take the Bloch form (7.2-6) with $K = 0$,



$$U(x,y,z) = p_0(z)\exp(-jk_x x), \qquad \text{(7.2-7)}$$

Normal-to-Axis Bloch Mode

where $p_0(z)$ is a periodic function representing a standing wave along the axis of periodicity.

Eigenvalue Problem, Dispersion Relation, and Photonic Bandgaps

Now that we have established the mathematical form of the modes, as imposed by the translational symmetry of the periodic medium, the next step is to solve the eigenvalue problem described by the generalized Helmholtz equation. For a mode with a Bloch wavenumber K, the eigenvalues ω^2/c_o^2 provide a discrete set of frequencies ω. These values are used to construct the ω-K dispersion relation. The eigenfunctions help us determine the Bloch periodic functions $p_K(z)$ for each of the values of ω associated with each K.

The ω-K relation is a periodic multivalued function of K with period g, the fundamental spatial frequency of the periodic structure; it is often plotted over the Brillouin zone $[-g/2 < k \leq g/2]$, as illustrated schematically in Fig. 7.2-2(a). When visualized as a monotonically increasing function of k, it appears as a continuous function with discrete jumps at values of K equal to integer multiples of $g/2$. These discontinuities correspond to the photonic bandgaps, which are spectral bands not crossed by the dispersion lines, so that no propagating modes exist.

The origin of the discontinuities in the dispersion relation lies in the special symmetry that emerges when $k = g/2$, i.e., when the period of the medium equals exactly half the period of the traveling wave. Consider the two modes with $k = \pm g/2$ and Bloch periodic functions $p_K(z) = p_{\pm g/2}(z)$. Since these modes travel with the same wavenumber, but in opposite directions, i.e. see inverted versions of the medium, $p_{-g/2}(z) = p_{g/2}(-z)$. But these two modes are in fact one and the same, because their Bloch wavenumbers differ by g. It therefore follows that at the edge of a Brillouin zone, there are two Bloch periodic functions that are inverted versions of one another. Since the medium is inhomogeneous or piecewise homogeneous within a unit cell, these two distinct functions interact with the medium differently, and therefore have two distinct eigenvalues, i.e., distinct values of ω. This explains the discontinuity that emerges as the continuous ω-K line crosses the boundary of the Brillouin zone. A similar argument explains the discontinuities that occur when K equals other integer multiples of $g/2$.

The variational principle (see Appendix C) is helpful in pointing out certain features of these eigenfunctions. Based on this principle, the eigenfunctions of a Hermitian operator are orthogonal distributions that minimize the variational energy. The variational energy associated with the linear operator $\mathcal{L}$ in the eigenvalue equation (7.0-2) is $E_v = \frac{1}{2}(\mathbf{H}, \mathcal{L}\mathbf{H})/(\mathbf{H}, \mathbf{H})$. By use of Maxwell's equations, it can be shown that $(\mathbf{H}, \mathcal{L}\mathbf{H}) = (\mathbf{H}, \nabla \times [\eta(\mathbf{r})\ \nabla \times \mathbf{H}]) = \int |\mathbf{D}(\mathbf{r})|^2/\epsilon(\mathbf{r})d\mathbf{r}$, so that minimization of E_v is achieved

by distributions for which higher displacement fields $\mathbf{D}(\mathbf{r})$ are located at positions of lower $1/\epsilon(\mathbf{r})$, i.e. higher refractive index. For example, if the periodic medium is made of two alternating dielectric layers, as illustrated in Fig. 7.2-2(b), then at a discontinuity the eigenfunction of the lower frequency concentrates its displacement field in the layer with the greater refractive index, whereas the eigenfunction of the higher frequency has an inverted distribution for which the displacement field is concentrated in the layer with the lower refractive index.

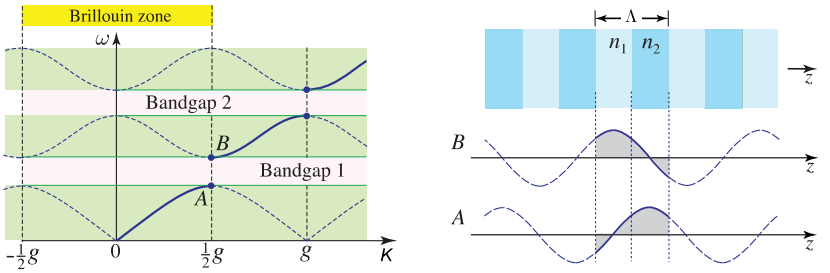


Figure 7.2-2 (a) The dispersion relation is a multivalued periodic function with period $g = 2\pi/\Lambda$ and discontinuities at k equals integer multiples of $g/2$. (b) Bloch functions at the points A and B at the edge of the Brillouin zone for an alternating dielectric-layer periodic medium with $n_2 > n_1$.

The challenging problem now is the solution of the eigenvalue problem associated with the Helmholtz equation. There are two approaches:

- The first approach is based on expanding the periodic function $\eta(z)$ of the medium and the periodic function $p_K(z)$ of the Bloch mode in Fourier series and converting the Helmholtz differential equation into a set of algebraic equation cast in the form of a matrix eigenvalue problem, which are solved numerically. This approach is called the **Fourier Optics** approach.
- The second approach is applicable to layered (piecewise homogeneous) media with planar boundaries. Instead of solving the Helmholtz equation, we make direct use of the laws of propagation and reflection/refraction at boundaries, which are known consequences of Maxwell's equations. We then use the matrix methods developed for multilayer media in Sec. 7.1A and applied to Bragg gratings in Sec. 7.1C. This **Matrix Optics** approach leads to a 2×2 matrix eigenvalue problem from which the dispersion relation and the Bloch modes are determined.

The matrix-optics approach is discussed next, and the Fourier-optics approach is examined in Sec. 7.2C.

B. Matrix Optics of Periodic Media

A one-dimensional periodic medium comprises identical segments, called **unit cells**, that are repeated periodically along one direction (the z axis) and separated by the period Λ (Fig. 7.2-3). Each unit cell contains a succession of lossless dielectric layers or partially reflective mirrors in some order, forming a symmetric system represented by a generic unimodular wave-transfer matrix

$$\mathbf{M}_o = \begin{bmatrix} 1/\mathrm{t}^* & \mathrm{r}/\mathrm{t} \\ \mathrm{r}^*/\mathrm{t}^* & 1/\mathrm{t} \end{bmatrix}, \tag{7.2-8}$$

where t and r are complex amplitude transmittance and reflectance satisfying the conditions set forth in (7.1-17), and $\mathcal{T} = |\mathrm{t}|^2$ and $\mathcal{R} = |\mathrm{r}|^2$ are the corresponding intensity

transmittance and reflectance. The medium is a Bragg grating, like that described in Sec. 7.1C, with an infinite number of segments. A wave traveling through the medium undergoes multiple transmissions and reflections that add up to one forward and one backward wave at every plane. We now use the matrix method developed in Sec. 7.1A to determine the Bloch modes.

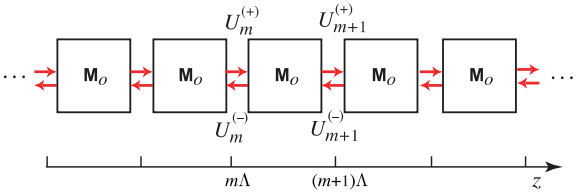


Figure 7.2-3 Wave-transfer matrix representation of a periodic medium.

Let $\{U_m^{(\pm)}\}$ be the complex amplitudes of the forward and backward waves at the initial position $z = m\Lambda$ of unit cell m. Knowing these amplitudes, the amplitudes elsewhere within the cell can be determined by straightforward application of the appropriate wave-transfer matrices, as described in Sec. 7.1. We therefore direct our attention to the dynamics of the amplitudes $\{U_m^{(\pm)}\}$ as they vary from one cell to the next. These dynamics are described by the recurrence relations

$$\begin{bmatrix} U_{m+1}^{(+)} \\ U_{m+1}^{(-)} \end{bmatrix} = \mathbf{M}_o \begin{bmatrix} U_m^{(+)} \\ U_m^{(-)} \end{bmatrix}, \tag{7.2-9}$$

which can be used to determine the amplitudes at a particular cell if the amplitudes at the previous cell are known.

Eigenvalue Problem and Bloch Modes

By definition, the modes of the periodic medium are self-reproducing waves, for which

$$\begin{bmatrix} U_{m+1}^{(+)} \\ U_{m+1}^{(-)} \end{bmatrix} = e^{-j\Phi} \begin{bmatrix} U_m^{(+)} \\ U_m^{(-)} \end{bmatrix}, \quad m = 1, 2, \ldots; \tag{7.2-10}$$

after transmission through a distance Λ (in this case a unit cell), the magnitudes of the forward and backward waves remain unchanged and the phases are altered by a common shift Φ, called the **Bloch phase**. The corresponding Bloch wavenumber is $K = \Phi/\Lambda$, so that

$$\boxed{\Phi = K\Lambda.} \tag{7.2-11}$$

Bloch Phase

Determination of the complex amplitudes $U_m^{(\pm)}$ and the phase $\Phi = K\Lambda$ that satisfy the self-reproduction condition (7.2-10) can be cast as an eigenvalue problem. This is accomplished by using (7.2-9) with $m = 0$ to write (7.2-10) in the form

$$\mathbf{M}_o \begin{bmatrix} U_0^{(+)} \\ U_0^{(-)} \end{bmatrix} = e^{-j\Phi} \begin{bmatrix} U_0^{(+)} \\ U_0^{(-)} \end{bmatrix}. \tag{7.2-12}$$

This is an eigenvalue problem for the 2×2 unit-cell matrix $\mathbf{M}_o$. The factor $e^{-j\Phi}$ is the eigenvalue and the vector with components $U_0^{(+)}$ and $U_0^{(-)}$ is the eigenvector.

The eigenvalues are determined by equating the determinant of the matrix $\mathbf{M}_o - e^{-j\Phi}\mathbf{I}$ to zero. Noting that $|\mathrm{t}|^2+|\mathrm{r}|^2 = 1$, the solution to the ensuing quadratic equation yields $e^{-j\Phi} = \frac{1}{2}(1/\mathrm{t} + 1/\mathrm{t}^*) \pm j\{1 - [\frac{1}{2}(1/\mathrm{t} + 1/\mathrm{t}^*)]^2\}^{1/2}$, from which

$$\boxed{\cos\Phi = \mathrm{Re}\left\{\frac{1}{\mathrm{t}}\right\}.} \tag{7.2-13}$$

Equation (7.2-13) is identical to (7.1-46) for the Bragg grating. This is gratifying inasmuch as the periodic medium at hand is an extended Bragg grating with an infinite number of segments.

Since $\mathbf{M}_o$ is a 2×2 matrix, it has two eigenvalues. Hence, only two of the multiple solutions of (7.2-13) are independent. Since the $\cos^{-1}(\cdot)$ function is even, the two solutions within the interval $[-\pi, \pi]$ have equal magnitudes and opposite signs. They correspond to Bloch modes traveling in the forward and backward directions. Other solutions obtained by adding multiples of 2π are not independent since they are irrelevant to the phase factor $e^{-j\Phi}$.

The associated eigenvectors of $\mathbf{M}_o$ are therefore

$$\begin{bmatrix} U_0^{(+)} \\ U_0^{(-)} \end{bmatrix} \propto \begin{bmatrix} \mathrm{r}/\mathrm{t} \\ e^{-j\Phi} - 1/\mathrm{t}^* \end{bmatrix}, \tag{7.2-14}$$

as can be ascertained by operating on the right-hand side of (7.2-14) with the $\mathbf{M}_o$ matrix; the result is again the right-hand side of (7.2-14) to within a constant.

The periodic function $p_K(z)$ associated with the Bloch wave can be determined by propagating the amplitudes $U_0^{(+)}$ and $U_0^{(-)}$ through the unit cell. For example, if the initial layer in the unit cell is a homogeneous medium of refractive index n_1 and width d_1, then the wave at distance z into this layer is

$$p_K(z)\, e^{-jKz} = U_0^{(+)} e^{-jn_1k_0z} + U_0^{(-)} e^{jn_1k_0z}, \quad 0 < z < d_1. \tag{7.2-15}$$

Using (7.2-14) and (7.2-11), (7.2-15) then provides

$$p_K(z) \propto [-\mathrm{r}e^{-jn_1k_0z} + (e^{-jK\Lambda} - 1)\, e^{jn_1k_0z}]\, e^{jKz}, \quad 0 < z < d_1. \tag{7.2-16}$$

The waves in (7.2-16) may be propagated further into the subsequent layers within the cell by using the appropriate $\mathbf{M}$ matrices.

Dispersion Relation and Photonic Band Structure

The **dispersion relation** is an equation relating the Bloch wavenumber K and the angular frequency ω. Equation (7.2-13), which provides the eigenvalues $\exp(-j\Phi)$ of the unit-cell matrix, is the progenitor of the dispersion relation for the 1D periodic medium. The phase $\Phi = K\Lambda$ is proportional to K, and $\mathrm{t} = \mathrm{t}(\omega)$ is related to ω through the phase delay associated with propagation through the unit cell, so that (7.2-13), written in the form

$$\boxed{\cos\left(2\pi\frac{K}{g}\right) = \mathrm{Re}\left\{\frac{1}{\mathrm{t}(\omega)}\right\},} \tag{7.2-17}$$

Dispersion Relation

is the $\omega - K$ dispersion relation. Here, $g = 2\pi/\Lambda$ is the fundamental spatial frequency of the periodic medium.

The function $\cos(2\pi K/g)$ is a periodic function of K of period $g = 2\pi/\Lambda$, so that for a given ω, (7.2-17) has multiple solutions. However, solutions separated by the period g are not independent since they lead to identical Bloch waves. It is therefore common to limit the domain of the dispersion relation to a period with values of K in the interval $[-g/2, g/2]$ or $[-\pi/\Lambda, \pi/\Lambda]$, which is the Brillouin zone. This corresponds precisely to limiting the phase Φ to the interval $[-\pi, \pi]$. Also, since $\cos(2\pi K/g)$ is an even function of K, at each value ω there are two independent values of K of equal magnitudes and opposite signs within the Brillouin zone. They correspond to independent Bloch modes traveling in the forward and backward directions.

The dispersion relation exhibits multiple spectral bands classified into two regimes:

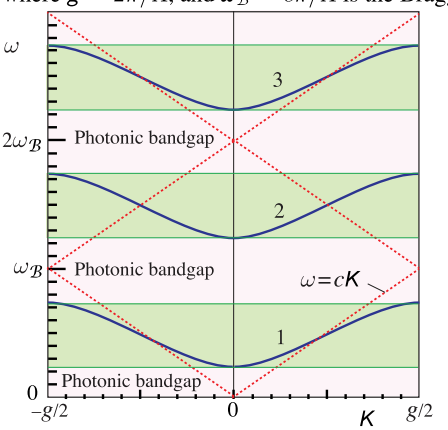
- **Propagation regime**. Spectral bands within which K is real correspond to propagating modes. Defined by the condition $|\mathrm{Re}\{1/\mathrm{t}(\omega)\}| \leq 1$, these bands are numbered, 1, 2, ..., starting with the lowest-frequency band.
- **Photonic-bandgap regime**. Spectral bands within which K is complex correspond to evanescent waves that are rapidly attenuated. Defined by the condition $|\mathrm{Re}\{1/\mathrm{t}(\omega)\}| > 1$, these bands correspond to the stop bands of the diffraction grating discussed in Sec. 7.1C. They are also called **photonic bandgaps** (**PBG**) or **forbidden gaps** since propagating modes do not exist.

The dispersion relation is often plotted with K measured in units of $g = 2\pi/\Lambda$, the fundamental spatial frequency of the periodic structure, whereas ω is measured in units of the Bragg frequency $\omega_{\mathrm{B}} = \pi c/\Lambda$, where $c = c_o/\bar{n}$ and $\bar{n}$ is the average refractive index of the periodic medium. The ratio $\omega_{\mathrm{B}}/(g/2) = c$, which is the slope of the dispersion relation $\omega = cK$ for propagation in a homogeneous medium with the average refractive index.

EXAMPLE 7.2-1. ***Periodic Stack of Partially Reflective Mirrors.*** The dispersion relation for a wave traveling along the axis of a periodic stack of identical partially reflective lossless mirrors with power reflectance $|\mathrm{r}|^2$ and intensity transmittance $|\mathrm{t}|^2 = 1 - |\mathrm{r}|^2$, separated by a distance Λ, is determined directly from Example 7.1-7. Using the results obtained there, namely $\mathrm{t} = |\mathrm{t}|e^{j\varphi}$ with $\varphi = nk_o\Lambda = (\omega/c)\Lambda$, in conjunction with (7.2-13), provides the dispersion relation

$$\cos\left(2\pi\frac{K}{g}\right) = \frac{1}{|\mathrm{t}|}\cos\left(\pi\frac{\omega}{\omega_{\mathrm{B}}}\right), \tag{7.2-18}$$

where $g = 2\pi/\Lambda$, and $\omega_{\mathrm{B}} = c\pi/\Lambda$ is the Bragg frequency. This result is plotted in Fig. 7.2-4.



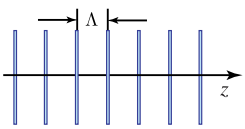


Figure 7.2-4 Dispersion diagram of a periodic set of mirrors, each with intensity transmittance $|\mathrm{t}|^2 = 0.5$, separated by a distance Λ. Here, $\omega_{\mathrm{B}} = \pi c/\Lambda$ and $g = 2\pi/\Lambda$. The dotted straight lines represent propagation in a homogeneous medium for which $\omega/K = \omega_{\mathrm{B}}/(g/2) = c$.

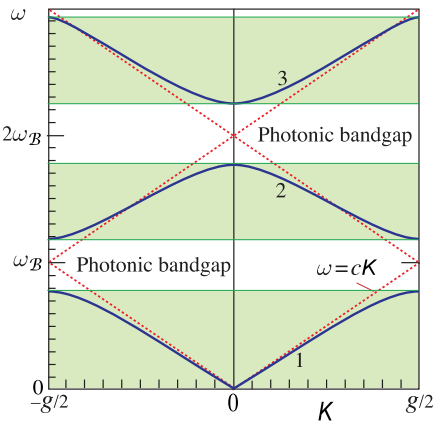
The photonic bandgaps, which correspond to frequency regions where (7.2-18) does not admit a real solution, are centered at $\omega_{\mathcal{B}}$, $2\omega_{\mathcal{B}}$, ... These frequency regions do not permit propagating modes; rather, they correspond to the stop bands that exhibit unity reflectance in Fig. 7.1-10. In this system, the onset of the lowest photonic bandgap is at $\omega = 0$.

EXAMPLE 7.2-2. ***Alternating Dielectric Layers.*** A periodic medium comprises alternating dielectric layers of refractive indices n_1 and n_2, with corresponding widths d_1 and d_2, and period $\Lambda = d_1 + d_2$. This system is the dielectric Bragg grating described in Example 7.1-8 with $N = \infty$. For a wave traveling along the axis of periodicity, $\mathrm{Re}\{1/\mathrm{t}\} = \mathrm{Re}\{A\}$ is given by (7.1-58). Using the relations $\varphi_1 + \varphi_2 = k_o(n_1 d_1 + n_2 d_2) = \pi\omega/\omega_{\mathcal{B}}$ and $\varphi_1 - \varphi_2 = \zeta\pi\omega/\omega_{\mathcal{B}}$, where $\omega_{\mathcal{B}} = (c_o/\bar{n})(\pi/\Lambda)$ is the Bragg frequency, $\bar{n} = (n_1 d_1 + n_2 d_2)/\Lambda$ is the average refractive index and $\zeta = (n_1 d_1 - n_2 d_2)/(n_1 d_1 + n_2 d_2)$, (7.2-13) provides the dispersion relation

$$\cos\left(2\pi\frac{K}{g}\right) = \frac{1}{\mathrm{t}_{12}\mathrm{t}_{21}}\left[\cos\left(\pi\frac{\omega}{\omega_{\mathcal{B}}}\right) - |\mathrm{r}_{12}|^2\cos\left(\pi\zeta\frac{\omega}{\omega_{\mathcal{B}}}\right)\right], \tag{7.2-19}$$

where $\mathrm{t}_{12}\mathrm{t}_{21} = 4n_1n_2/(n_1+n_2)^2$ and $|\mathrm{r}_{12}|^2 = (n_2-n_1)^2/(n_1+n_2)^2$.

An example of this dispersion relation is plotted in Fig. 7.2-5 for dielectric materials with $n_1 = 1.5$ and $n_2 = 3.5$, and $d_1 = d_2$. As with the periodic stack of partially reflective mirrors considered in Example 7.2-1, the photonic bandgaps are centered at the frequencies $\omega_{\mathcal{B}}$ and its multiples, and occur at either the center of the Brillouin zone ($K = 0$) or at its edge ($K = g/2$). In this case, however, the frequency region surrounding $\omega = 0$ admits propagating modes instead of a forbidden gap. Dielectric materials with lower contrast have bandgaps of smaller width, but the bandgaps exist no matter how small the contrast.



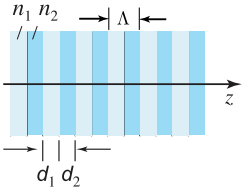


Figure 7.2-5 Dispersion diagram of an alternating-layer periodic dielectric medium with $n_1 = 1.5$ and $n_2 = 3.5$, and $d_1 = d_2$. Here, $\omega_{\mathcal{B}} = \pi c_o/\Lambda\bar{n}$ and $g = 2\pi/\Lambda$. The dotted straight lines represent propagation in a homogeneous medium of mean refractive index $\bar{n}$, so that $\omega/K = \omega_{\mathcal{B}}/(g/2) = c_o/\bar{n} = c$.

Phase and Group Velocities

The propagation constant K corresponds to a phase velocity ω/K and an effective refractive index $n_{\mathrm{eff}} = c_o K/\omega$. The group velocity $v = d\omega/dK$, which governs pulse propagation in the medium, is associated with an effective group index $N_{\mathrm{eff}} = c_o dK/d\omega$ (see Sec.5.7). These indices can be determined at any point on the ω-K dispersion curve by finding the slope $d\omega/dK$, and the ratio ω/K, i.e., the slope of a line joining the point with the origin. Figure 7.2-6 is a schematic illustration of a dispersion relation of an alternating-layer dielectric periodic medium, together with the effective index and group index, at frequencies extending over two photonic bands with a photonic bandgap in-between.

At low frequencies within the first photonic band, n_{eff} is approximately equal to the average refractive index $\bar{n}$. This is expected since at long wavelengths the material

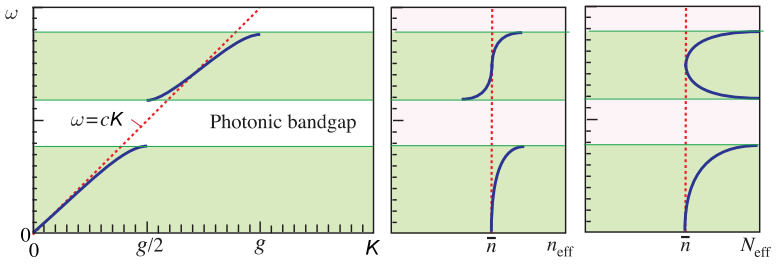


Figure 7.2-6 Frequency dependence of the effective refractive index n_{eff}, which determines the phase velocity, and the effective group index N_{eff}, which determines the group velocity.

behaves as a homogeneous medium with the average refractive index. With increase of frequency, n_{eff} increases above $\bar{n}$, reaching its highest value at the band edge. At the bottom of the second band, n_{eff} is smaller than $\bar{n}$ but increases at higher frequencies, approaching $\bar{n}$ in the middle of the band.

This drop of n_{eff} from a value above average just below the bandgap to a value below average just above the bandgap is attributed to the significantly different spatial distributions of the corresponding Bloch modes, which are orthogonal. The mode at the top of the lower band, has greater energy in the dielectric layers with the higher refractive index, so that its effective index is greater than the average. For the mode at the bottom of the upper band, greater energy is localized in the layers with the lower refractive index, and the effective index is therefore lower than the average.

The frequency dependence of the effective group index follows a different pattern, as shown in Fig. 7.2-6. As the edges of the bandgap are approached, from below or above, this index increases substantially, so that the group velocity is much smaller, i.e., optical pulses are very slow near the edges of the bandgap.

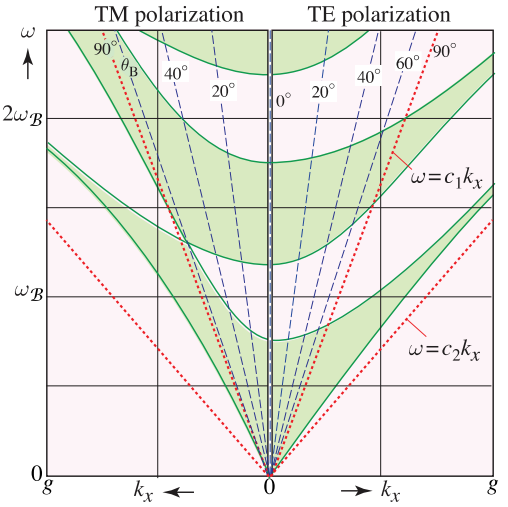
Off-Axis Dispersion Relation and Band Structure

The dispersion relation for off-axis waves may be determined by using the same equation, $\cos(K\Lambda) = \text{Re}\{1/\text{t}(\omega)\}$, where $\text{Re}\{1/\text{t}(\omega)\}$ now depends on the angles of incidence within the layers of each segment and on the state of polarization (TE or TM). For example, for a periodic medium made of alternating dielectric layers, $\text{Re}\{1/\text{t}(\omega)\}$ takes the more general form in (7.1-59).

Since the same transverse component k_x of the wavevector determines the angles of incidence at the two layers ($k_x = n_1 k_o \sin\theta_1 = n_2 k_o \sin\theta_2$), it is more convenient to express the dispersion relation as a function of k_x, in the form of a three-dimensional surface $\omega = \omega(K, k_x)$. Every value of k_x yields a dispersion diagram with bands and bandgaps similar to those of Fig. 7.2-5.

A simpler representation of the $\omega(K, k_x)$ 3D surface is the **projected dispersion diagram**, which displays in a two-dimensional plot of the edges of the bands and bandgaps at each value of k_x, for both TE and TM polarizations, as illustrated in Fig. 7.2-7. This figure is constructed by determining the ranges of angular frequencies over which photonic bands and bandgaps exist in the dispersion diagram for a particular value of k_x, and then projecting these onto corresponding vertical lines at that value of k_x in the projected dispersion diagram. The loci of all such vertical lines for the bands at different values of k_x correspond to the shaded (green) areas displayed in Fig. 7.2-7; the unshaded (white) areas represent the bandgaps.

In this diagram, each angle of incidence is represented by a straight line passing through the origin. For example, the incidence angle θ_1 in layer 1 corresponds to the line $k_x = (\omega/c_1)\sin\theta_1$, i.e., $\omega = (c_1/\sin\theta_1)k_x$, where $c_1 = c_o/n_1$. The line $\omega = c_1 k_x$, called the **light line** corresponds to $\theta_1 = 90°$. Similar lines may be drawn for



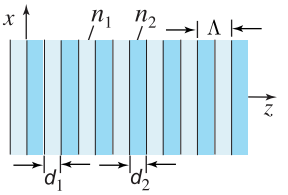


Figure 7.2-7 Projected dispersion diagram for an alternating-layer periodic dielectric medium with $n_1 = 1.5$, $n_2 = 3.5$, and $d_1 = d_2 = \Lambda/2$. Here, $\omega_{\mathcal{B}} = \pi c_o/\Lambda\bar{n}$ and $g = 2\pi/\Lambda$. Photonic bands are shaded (green). The dashed lines represent fixed angles of incidence θ_1 in layer 1, including the Brewster angle $\theta_{\mathrm{B}} = 66.8°$. Points within the region bounded by the light lines $\omega = c_1 k_x$ and $\omega = c_2 k_x$ represent normal-to-axis waves.

the incidence angles in medium 2; Fig. 7.2-7 shows only the light line $\omega = c_2 k_x$, assuming that $n_2 > n_1$, i.e., $c_2 < c_1$. Points in the region bounded by the two light lines represent normal-to-axis modes, which travel in the lateral direction by undergoing total reflection in the denser medium (medium 2).

The question arises as to whether there exists a frequency range over which propagation is forbidden at all angles of incidence θ_1 and θ_2 and for both polarizations. This could occur if the forbidden gaps at all values of k_x between the lines $k_x = 0$ and $k_x = \omega/c_2$, and for both polarizations, were to align in such a way as to create a common or **complete photonic bandgap**. This is clearly not the case in the example in Fig. 7.2-7. It turns out that this is not possible; complete photonic bandgaps cannot exist within 1D periodic structures. However, they can occur in 2D and 3D periodic structures, as we shall see in Sec. 7.3.

Indeed, there is one special case in which a photonic bandgap cannot occur at all, and that is an oblique TM wave propagating at the Brewster angle $\theta_{\mathrm{B}} = \tan^{-1}(n_2/n_1)$ in layer 1. As shown in Fig. 7.2-7, the line at the Brewster angle does not pass through a gap. This is not surprising since at this angle, the reflectance of a unit cell is zero, and the forward and backward waves are uncoupled so that the collective effect that leads to total reflection is absent.

C. Fourier Optics of Periodic Media

The matrix analysis of periodic media presented in the previous section is applicable to layered (i.e., piecewise homogeneous) media. A more general approach, applicable for arbitrary periodic media, including continuous media, is based on a Fourier-series representation of periodic functions and conversion of the Helmholtz equation into a set of algebraic equations whose solution provides the dispersion relation and the Bloch modes. This approach can also be generalized to 2D and 3D periodic media, as will be shown in Sec. 7.3.

A wave traveling along the axis of a 1D periodic medium (the z axis) and polarized in the x direction is described by the generalized Helmholtz equation (7.2-2). Since $\eta(z)$ is periodic with period Λ, it can be expanded in a Fourier series,

$$\eta(z) = \sum_{\ell=-\infty}^{\infty} \eta_\ell \exp(-j\ell g z), \tag{7.2-20}$$

where $g = 2\pi/\Lambda$ is the spatial frequency (rad/mm) of the periodic structure and η_ℓ is the Fourier coefficient representing the ℓth harmonic. The impermeability $\eta(z)$ is real, so that $\eta_{-\ell} = \eta_\ell^*$.

The periodic portion of the Bloch wave $p_K(z)$ in (7.2-4) may also be expanded in a Fourier series,

$$p_K(z) = \sum_{m=-\infty}^{\infty} C_m \exp(-jmgz), \tag{7.2-21}$$

whereupon the Bloch wave representation of the magnetic field may be written as

$$H_y(z) = \sum_{m=-\infty}^{\infty} C_m \exp[-j(K + mg)z]. \tag{7.2-22}$$

For brevity, the dependence of the Fourier coefficients $\{C_m\}$ on the Bloch wavenumber K is suppressed. Substituting these expansions into the Helmholtz equation (7.2-2) and equating harmonic terms of the same spatial frequency, we obtain

$$\sum_{\ell=-\infty}^{\infty} F_{m\ell} C_\ell = \frac{\omega^2}{c_o^2} C_m, \quad F_{m\ell} = (K + mg)(K + \ell g)\,\eta_{m-\ell}, \tag{7.2-23}$$

where $m = 0, \pm1, \pm2, \ldots$.

The differential equation (7.2-2) has now been converted into a set of linear equations (7.2-23) for the unknown Fourier coefficients $\{C_m\}$. These equations may be cast in the form of a matrix eigenvalue problem. For each K, the eigenvalues ω^2/c_o^2 correspond to multiple values of ω, from which the ω-K dispersion relation may be constructed. The eigenvectors are sets of $\{C_m\}$ coefficients, which determine the periodic function $p_K(z)$ of the Bloch mode for each K.

Posed as an eigenvalue problem for a matrix $\mathbf{F}$ with elements $F_{m\ell}$, this set of coupled equations may be solved using standard numerical techniques. Since $\eta_{m-\ell} = \eta_{\ell-m}^*$, the matrix $\mathbf{F}$ is Hermitian, i.e., $F_{m\ell} = F_{\ell m}^*$. Note that if we were to use the electric-field Helmholtz equation instead of the magnetic-field Helmholtz equation (7.2-2), we would obtain another matrix representation of the eigenvalue problem, but the matrix would be non-Hermitian, and therefore more difficult to solve. This is the rationale for working with the Helmholtz equation for the magnetic field.[†]

Approximate Solution of the Eigenvalue Problem

In (7.2-23), the harmonics of the optical wave are coupled via the harmonics of the periodic medium. An optical-wave harmonic of spatial frequency $K + \ell g$ mixes with a medium harmonic of spatial frequency $(m - \ell)g$ and contributes to the optical-wave harmonic of spatial frequency $(K + \ell g) + (m - \ell)g = K + mg$.

The conditions under which strong coupling emerges can be determined by separating out the mth term in (7.2-23), which leads to

$$C_m = \sum_{\ell \neq m} \frac{\eta_{m-\ell}}{\eta_o} \frac{(K + \ell g)(K + mg)}{(\bar{n}\omega/c_o)^2 - (K + mg)^2} C_\ell, \quad m = 0, \pm1, \pm2, \ldots, \tag{7.2-24}$$

[†] It can be shown that the differential operator in the generalized magnetic-field Helmholtz equation (7.0-2) is Hermitian whereas that in the electric-field Helmholtz equation (7.0-2) is non-Hermitian.

where $\bar{n} = 1/\sqrt{\eta_0}$ is an average refractive index of the medium. Strong coupling between the mth harmonic of the wave and other harmonics exists if the denominator in (7.2-24) is small, i.e.,

$$\omega\bar{n}/c_o \approx |K + mg|. \tag{7.2-25}$$

This equation represents a resonance condition for interaction between the harmonics. It can also be regarded as a **phase-matching condition**.

Figure 7.2-8 is a plot of (7.2-25) as an equality. For each value of m, the ω-K relation is a V-shaped curve. The intersection points of these curves represent common values of ω and K at which (7.2-25) is simultaneously satisfied for two harmonics. The intersections between the $m = 0$ curve (dashed) and the curves for $m = -1$, $m = -2$, ..., are marked by filled circles; they correspond to the lowest-order bandgaps 1, 2, ..., respectively. At each intersection point, K is an integer multiple of $\frac{1}{2}g$, and ω is an integer multiple of the Bragg frequency $\omega_{\mathcal{B}} = (c_o/\bar{n})g/2$ or $\omega_{\mathcal{B}}/2\pi = \nu_{\mathcal{B}} = (c_o/\bar{n})/2\Lambda$. This corresponds to the Bragg wavelength $\lambda_{\mathcal{B}} = 2\Lambda$ in the medium, and therefore to total reflection. Unmarked intersections in Fig. 7.2-8 are not independent since each of these has the same ω as a marked intersection, and a value of K differing by a reciprocal lattice constant g.

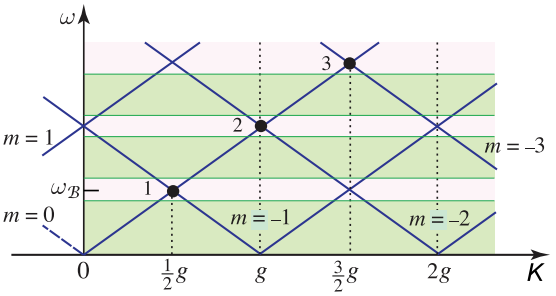


Figure 7.2-8 Plot of (7.2-25), as an equality, for various values of m. The $m = 0$ curve is indicated as dashed. Strong coupling between the harmonics of the optical wave and those of the medium occurs at the intersection points 1, 2,..., which correspond to the lowest-order bandgaps.

The lowest-order bandgap occurs at the intersection of the $m = 0$ and $m = -1$ curves (point 1 in Fig. 7.2-8). In this case, only the coefficients C_0 and C_{-1} are strongly coupled, so that (7.2-24) yields two coupled equations

$$C_0 = \frac{\eta_1}{\eta_0} \frac{(K-g)K}{\omega^2\bar{n}^2/c_o^2 - K^2} C_{-1}\,, \tag{7.2-26}$$

$$C_{-1} = \frac{\eta_1^*}{\eta_0} \frac{K(K-g)}{\omega^2\bar{n}^2/c_o^2 - (K-g)^2} C_0\,, \tag{7.2-27}$$

where $\eta_{-1} = \eta_1^*$. These equations are self-consistent if

$$\frac{|\eta_1|^2}{\eta_0^2} K^2(K-g)^2 = \left[\omega^2\frac{\bar{n}^2}{c_o^2} - K^2\right]\left[\omega^2\frac{\bar{n}^2}{c_o^2} - (K-g)^2\right]. \tag{7.2-28}$$

(7.2-28) Dispersion Relation

A plot of this equation (Fig. 7.2-9) yields the ω-K dispersion relation near the edge of the bandgap, where the equation is valid. For $K = \frac{1}{2}g$, (7.2-28) yields two frequencies,

$$\omega_\pm = \omega_{\mathcal{B}}\sqrt{1 \pm |\eta_1|/\eta_0}\,, \tag{7.2-29}$$

representing the edges of the first photonic bandgap. The center of the bandgap is at the Bragg frequency $\omega_{\mathcal{B}} = (c_o/\bar{n})(g/2) = (\pi/\Lambda)(c_o/\bar{n})$. The ratio of the gap width to the midgap frequency, which is called the **gap-midgap ratio**, grows with increasing impermeability contrast ratio $|\eta_1|/\eta_0$.

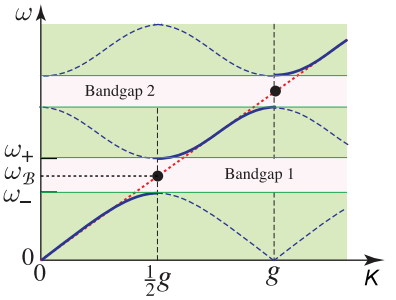


Figure 7.2-9 Dispersion relation in the vicinity of photonic bandgaps.

A similar procedure can be followed to determine the spectral width of higher-order bandgaps. The width of the mth bandgap is determined by a formula identical to (7.2-29), but the ratio $|\eta_m|/\eta_0$ replaces $|\eta_1|/\eta_0$, so that higher bandgaps are governed by higher spatial harmonics of the periodic function $\eta(z)$.

Off-Axis Waves

The dispersion relation for off-axis waves may be determined by use of the same Fourier expansion technique. For a TM-polarized off- axis wave traveling in an arbitrary direction in the x–z plane, the Helmholtz equation is given by (7.2-3). The Bloch wave is a generalization of (7.2-22) obtained via (7.2-6),

$$H_y(z) = \sum_{m=-\infty}^{\infty} C_m \exp\left[-j(K + mg)z\right] \exp(-jk_x x). \tag{7.2-30}$$

Carrying out calculations similar to the on-axis case, leads to the following set of algebraic equations for the C_m coefficients:

$$\sum_{\ell=-\infty}^{\infty} F_{m\ell}\, C_\ell = \frac{\omega^2}{c_o^2}\, C_m, \qquad F_{m\ell} = \left[(K + \ell g)(K + mg) + k_x^2\right] \eta_{m-\ell}. \tag{7.2-31}$$

Equation (7.2-31) is a generalization of (7.2-23) for the off-axis wave. The dispersion relation may be determined by solving this matrix eigenvalue problem for the set of frequencies ω associated with each pair of values of K and k_x.

D. Boundaries Between Periodic and Homogeneous Media

The study of light wave propagation in periodic media has so far been limited to determining the dispersion relation and its associated band structure, as well as estimating the phase and group velocity of such waves. By definition, the periodic medium extend indefinitely in all directions. The next step is to examine reflection and transmission at boundaries between periodic and homogeneous media. We first examine reflection from a single boundary and subsequently consider a slab of periodic medium embedded in a homogeneous medium. Other configurations made of homogeneous structures such slabs or holes embedded in extended periodic media are described in Sec. 10.4 and Sec. 11.4D.

Omnidirectional Reflection at a Single Boundary

We examine the reflection and transmission of an optical wave at the boundary between a semi-infinite homogeneous medium and a semi-infinite one-dimensional periodic medium, as portrayed in Fig. 7.2-10. We demonstrate that, under certain conditions and within a specified range of angular frequencies, the periodic medium acts as a perfect mirror, totally reflecting waves incident from any direction and with any polarization!

Wave transmission and reflection at the boundary between two media is governed by the phase-matching condition. At the boundary between two homogeneous media, for example, the transverse components of the wavevector k_x must be the same on both sides of the boundary. Since $k_x = k \sin\theta = (\omega/c_o) n \sin\theta$, this condition means that the product $n \sin\theta$ is invariant. This is the origin of Snell's law of refraction, as explained in Sec. 2.4A.

Similarly, for a wave incident from a homogeneous medium into a one-dimensional periodic medium, k_x must remain the same. Thus, if the incident wave has angular frequency ω and angle of incidence θ, we have $k_x = (\omega/c_o) n \sin\theta$, where n is the refractive index of the homogeneous medium. Knowing k_x and ω, we can use the dispersion relation $\omega = \omega(K, k_x)$ for the periodic medium at the appropriate polarization to determine the Bloch wavenumber K. If the angular frequency ω lies within a forbidden gap at this value of k_x, the incident wave will not propagate into the periodic medium but will, instead, be totally reflected. This process is repeated for all frequencies, angles of incidence, and polarizations of the incident wave.

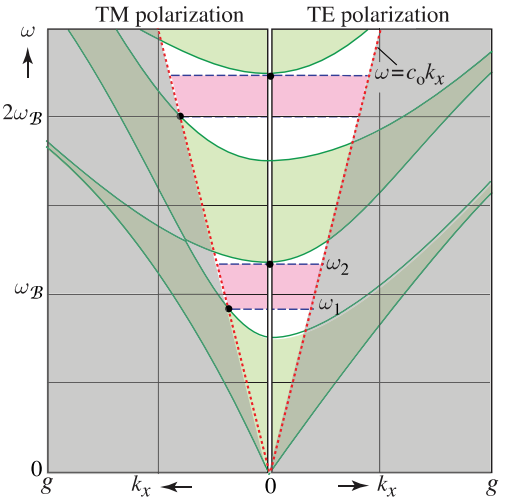
We now consider the possibility that the boundary acts as an omnidirectional reflector (a perfect mirror). For this purpose, we use the projected dispersion diagram, which displays the bandgaps for each value of k_x, as illustrated in the example provided in Fig. 7.2-10. On the same diagram, we delineate by a red dashed line the ω-k_x region that can be accessed by waves entering from the homogeneous medium. This region is defined by the equation $k_x = (\omega/c_o) n \sin\theta$, which dictates that $k_x < (\omega/c_o) n$ or $\omega > (c_o/n) k_x$; it is thus bounded by the line $\omega = (c_o/n) k_x$, or $\omega/\omega_{\mathcal{B}} = (\bar{n}/n)[k_x/(g/2)]$, known as the light line. This line corresponds to an angle $\theta = 90^\circ$ in the surrounding medium.

Figure 7.2-10 reproduces Fig. 7.2-7 with the light lines added, and the permissible ω-k_x regions within the light lines highlighted. Waves incident from the homogeneous medium at all angles, and both polarizations, are represented by points within this region; points outside this region are not accessible by waves incident from the homogeneous medium regardless of their angle of incidence or polarization. The spectral band bounded by the angular frequencies ω_1 and ω_2, as defined in Fig. 7.2-10, is of particular interest inasmuch as all ω-k_x points lying in this band are within a photonic bandgap. In this spectral band, therefore, no incident wave, regardless of its angle or polarization, can be matched with a propagating wave in the periodic medium — the boundary then acts as a perfect omnidirectional reflector. Also illustrated in Fig. 7.2-10 is a second spectral band, at higher angular frequencies, that behaves in the same way.

Slab of Periodic Material in a Homogeneous Medium

A slab of 1D periodic material embedded in a homogeneous medium is nothing but a 1D Bragg grating with a finite number of segments. Reflection and transmission from the Bragg grating has already been examined in Sec. 7.1C.

One would expect that a Bragg grating with a large, but finite, number of segments N captures the basic properties of a periodic medium made of the same unit cell. This is in fact the case since the passbands and stop bands of the grating are mathematically identical to the photonic bands and bandgaps of the extended periodic medium. However, the spectral transmittance and reflectance of the Bragg grating, which exhibit oscillatory properties sensitive to the size of the grating and the presence of its boundaries, do not have their counterparts in the extended periodic structure.



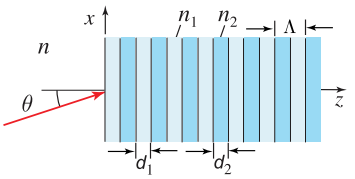


Figure 7.2-10 Projected dispersion diagram for an alternating-layer dielectric medium with $n_1 = 1.5$, $n_2 = 3.5$, and $d_1 = d_2 = \Lambda/2$. The dotted lines (red) are light lines for a homogeneous medium with refractive index $n = 1$. In the spectral band between ω_1 and ω_2, the medium acts as a perfect omnidirectional reflector for all polarizations. A similar band is shown at higher angular frequencies.

Likewise, the phase and group velocities and the associated effective refractive indices determined from the dispersion relation in the extended periodic medium do not have direct counterparts in the finite-size Bragg grating. Nevertheless, such parameters can be defined for a grating by determining the complex amplitude transmittance $\mathrm{t}(\omega)$ and matching it with an effective homogeneous medium of the same total thickness d such that $\arg\{\mathrm{t}_N\} = (\omega/c_o)n_{\text{eff}}d$. An effective group index $N_{\text{eff}} = n_{\text{eff}} + \omega dn_{\text{eff}}/d\omega$ is then determined [see (5.7-2)]. The dependence of these effective indices on frequency is different from that shown in Fig. 7.2-6 for the extended periodic medium in that it exhibits oscillations within the passbands. However, for sufficiently large N, say greater than 100, these oscillations are washed out and the effective indices become nearly the same as those of the extended periodic medium.

Another configuration of interest is a slab of homogeneous medium embedded in a periodic medium. In this configuration, the light may be trapped in the slab by omnidirectional reflection from the surrounding periodic medium, so that the slab becomes an optical waveguide. This configuration is discussed in Sec. 9.5.

7.3 TWO- AND THREE-DIMENSIONAL PHOTONIC CRYSTALS

The concepts introduced in Sec. 7.2 for the study of optical-wave propagation in 1D periodic media can be readily generalized to 2D and 3D structures. These include Bloch waves as the modes of the periodic medium and ω-K dispersion relations with photonic bands and bandgaps. In contrast to 1D structures, 2D photonic crystals have complete 2D photonic bandgaps, i.e., common bandgaps for waves of both polarization traveling in any direction in the plane of periodicity. However, complete 3D photonic bandgaps, i.e., common bandgaps for all directions and polarizations, can be achieved only in 3D photonic crystals. The mathematical treatment of 2D and 3D periodic media is more elaborate and the visualization of the dispersion diagrams is more difficult because of the additional degrees of freedom involved, but the concepts are essentially the same as those encountered for 1D periodic media. This section begins with a simple treatment of 2D structures followed by a more detailed 3D treatment.

A. Two-Dimensional Photonic Crystals

2D Periodic Structures

Consider a 2D periodic structure such as a set of identical parallel rods, tubes, or veins embedded in a homogeneous host medium [Fig. 7.3-1(a)] and organized at the points of a rectangular lattice, as illustrated in Fig. 7.3-1(b). The impermeability $\eta(x,y) = \epsilon_o/\epsilon(x,y)$ is periodic in the transverse directions, x and y, and uniform in the axial direction z. If a_1 and a_2 are the periods in the x and y directions, then $\eta(x,y)$ satisfies the translational symmetry relation

$$\eta(x + m_1 a_1, y + m_2 a_2) = \eta(x,y), \tag{7.3-1}$$

for all integers m_1 and m_2. This periodic function is represented as a 2D Fourier series,

$$\eta(x,y) = \sum_{\ell_1=-\infty}^{\infty} \sum_{\ell_2=-\infty}^{\infty} \eta_{\ell_1,\ell_2} \exp\left(-j\ell_1 g_1 x\right) \exp\left(-j\ell_2 g_2 x\right), \tag{7.3-2}$$

where $g_1 = 2\pi/a_1$ and $g_2 = 2\pi/a_2$ are fundamental spatial frequencies (radians/mm) in the x and y directions, and $\ell_1 g_1$ and $\ell_2 g_2$ are their harmonics. The coefficients η_{ℓ_1,ℓ_2} depend on the actual profile of the periodic function, e.g., the size of the rods.

The 2D Fourier transform of the periodic function is composed of points (delta functions) on a rectilinear lattice, as shown in Fig. 7.3-1(c). This Fourier-domain lattice is known to solid-state physicists as the **reciprocal lattice**.

What are the optical modes of a medium with such symmetry? The answer is a simple generalization of the 1D case given in (7.2-4). For waves traveling in a direction parallel to the x–y plane, the modes are 2D Bloch waves,

$$U(x,y) = p_{K_x,K_y}(x,y) \exp(-jK_x x) \exp(-jK_y y), \tag{7.3-3}$$

where $p_{K_x,K_y}(x,y)$ is a periodic function with the same periods as the medium. The wave is specified by a pair of Bloch wavenumbers (K_x, K_y). Another wave with Bloch wavenumbers $(K_x + g_1, K_y + g_2)$ is not a new mode. As shown in Fig. 7.3-1(c) a complete set of modes in the Fourier plane has Bloch wavenumbers located at points in a rectangle defined by $[-g_1/2 < K_x \leqslant g_1/2]$ and $[-g_2/2 < K_y \leqslant g_2/2]$, which is the first Brillouin zone.

Other symmetries may be used to reduce the set of independent Bloch wavevectors within the Brillouin zone. When all symmetries are included, the result is an area called the **irreducible Brillouin zone**. For example, the rotational symmetry inherent in the square lattice results in an irreducible Brillouin zone in the form of a triangle, as shown in Fig 7.3-1(d).

2D Skew-Periodic Structures

An example of another class of 2D periodic structures is a set of parallel cylindrical holes placed at the points of a triangular lattice, as illustrated in Fig. 7.3-2(a). Since the lattice points are skewed (not aligned with x and y axis), we use two primitive *vectors* $\mathbf{a}_1$ and $\mathbf{a}_2$ [Fig. 7.3-2(b)] to generate the lattice via the lattice vector $\mathbf{R} = m_1\mathbf{a}_1 + m_2\mathbf{a}_2$, where m_1 and m_2 are integers. We also define a position vector $\mathbf{r}_{\mathrm{T}} = (x,y)$ so that the periodic function $\epsilon(\mathbf{r}_{\mathrm{T}}) \equiv \epsilon(x,y)$ satisfies the translational symmetry relation $\epsilon(\mathbf{r}_{\mathrm{T}} + \mathbf{R}) = \epsilon(\mathbf{r}_{\mathrm{T}})$ (the subscript "T" indicates transverse).

The 2D Fourier series of such a function is a set of points on a reciprocal lattice defined by the vectors $\mathbf{g}_1$ and $\mathbf{g}_2$, which are orthogonal to $\mathbf{a}_1$ and $\mathbf{a}_2$, respectively, and have magnitudes $g_1 = 2\pi/a_1 \sin\theta$ and $g_2 = 2\pi/a_2 \sin\theta$, where θ is the angle between $\mathbf{a}_1$ and $\mathbf{a}_2$. The 2D reciprocal lattice is also a triangular lattice generated by

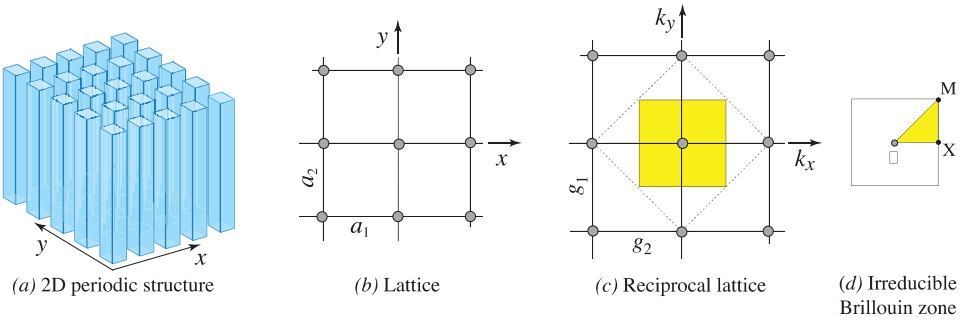


Figure 7.3-1 (a) A 2D periodic structure comprising parallel rods. (b) The rectangular lattice at which the rods are placed. (c) The 2D Fourier transform of the lattice points is another set of points forming a reciprocal lattice with periods $g_1 = 2\pi/a_1$ and $g_2 = 2\pi/a_2$. The shaded (yellow) area is the Brillouin zone. (d) For a square lattice ($a_1 = a_2 = a$), the irreducible Brillouin zone is the triangle ΓMX.

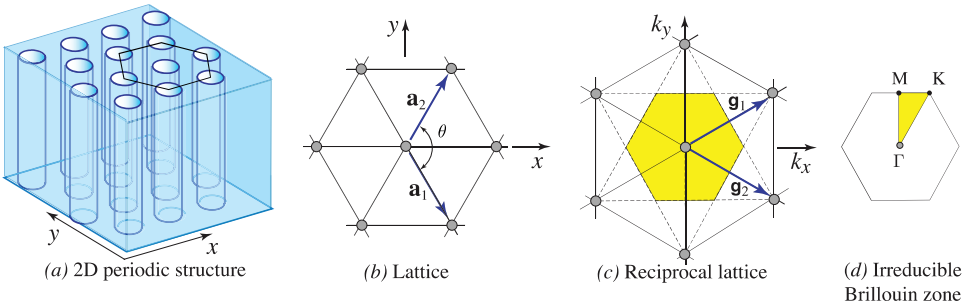


Figure 7.3-2 (a) A 2D periodic structure comprising parallel cylindrical holes. (b) The triangular lattice at which the holes are placed. In this diagram the magnitudes $a_1 = a_2 = a$ and $\theta = 120°$. (c) Reciprocal lattice; the shaded (yellow) area is the Brillouin zone, a hexagon. (d) The irreducible Brillouin zone is the triangle ΓMK.

the vector $\mathbf{G} = \ell_1\mathbf{g}_1 + \ell_2\mathbf{g}_2$, where ℓ_1 and ℓ_2 are integers, as illustrated in Fig. 7.3-2(c).

For waves traveling in a direction parallel to the x–y plane, the Bloch modes are

$$U(\mathbf{r}_\mathrm{T}) = p_\mathbf{K}(\mathbf{r}_\mathrm{T}) \exp(-j\mathbf{K}_\mathrm{T} \cdot \mathbf{r}_\mathrm{T}), \tag{7.3-4}$$

where $\mathbf{K}_\mathrm{T} = (K_x, K_y)$ is the Bloch wavevector and $p_{\mathbf{K}_\mathrm{T}}(\mathbf{r}_\mathrm{T})$ is a 2D periodic function on the same lattice. Two Bloch modes with Bloch wavevectors $\mathbf{K}_\mathrm{T}$ and $\mathbf{K}_\mathrm{T} + \mathbf{G}$ are equivalent. To cover a complete set of Bloch wavevectors, we therefore need only consider vectors within the Brillouin zone shown in Fig. 7.3-2(c).

The dispersion relation can be determined by ensuring that the Bloch wave in (7.3-3) or (7.3-4) satisfies the generalized Helmholtz equation. The calculations are facilitated by use of a Fourier series approach, as was done in the 1D case and as will be described (in a more general form) in the 3D case.

EXAMPLE 7.3-1. ***Cylindrical Holes on a Triangular Lattice.*** A 2D photonic crystal comprises a homogeneous medium ($n = 3.6$) with air-filled cylindrical holes of radius $0.48\,a$ organized at the points of a triangular lattice with lattice constant a. The calculated dispersion relation, shown in Fig. 7.3-3, for TE and TM waves traveling in the plane of periodicity ($k_z = 0$) exhibits a complete

2D photonic bandgap at frequencies near the angular frequency $\omega_0 = \pi c_o/a$.[†] As in the 1D case, the gap can be made wider by use of materials with greater refractive-index contrast. Indeed, most geometries exhibit photonic bandgaps if the constituent materials have sufficiently high contrast.

This photonic-crystal structure finds use as a "holey" optical fiber, which has a number of salutary properties (see Sec. 10.4).

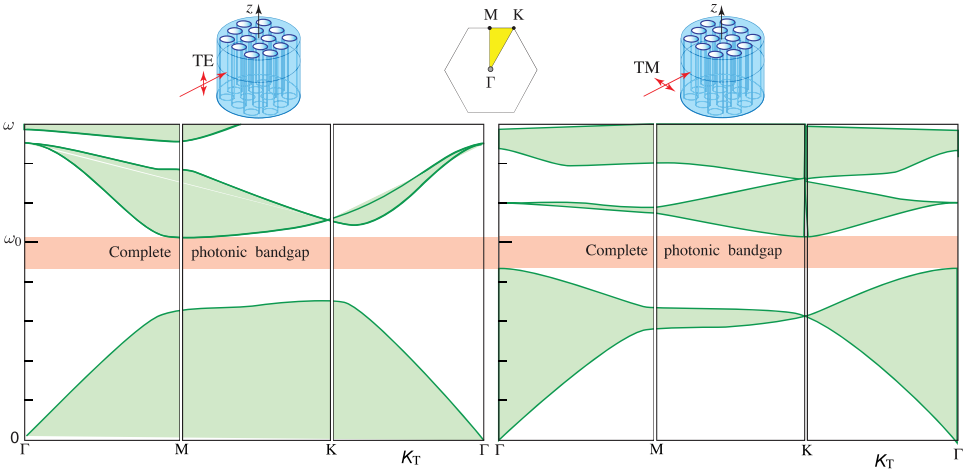


Figure 7.3-3 Calculated band structure of a 2D photonic crystal consisting of a homogeneous medium ($n = 3.6$) with air-filled cylindrical holes of radius $0.48\,a$ organized at the points of a triangular lattice with lattice constant a. The abscissa spans Bloch wavevectors defined by points on the periphery of the irreducible Brillouin zone, the ΓMK triangle. The ordinate is plotted in units of $\omega_0 = \pi c_o/a$. The wave travels in the plane of periodicity and has TE polarization (left) and TM polarization (right). The complete 2D photonic bandgap in the vicinity of ω_0 is highlighted.

For an oblique wave traveling at an angle with respect to the x–y plane, the Bloch wave in (7.3-4) becomes

$$U(\mathbf{r}_\mathrm{T}) = p_\mathbf{K}(\mathbf{r}_\mathrm{T}) \exp(-j\mathbf{K}_\mathrm{T} \cdot \mathbf{r}_\mathrm{T}) \exp(-jk_z z), \tag{7.3-5}$$

where k_z is a constant. The band structure then takes the form of a set of surfaces of $\omega = \omega(\mathbf{K}_\mathrm{T}, k_z)$.

A complete 3D photonic bandgap is a range of frequencies ω crossed by none of these surfaces, i.e., values of ω that are not obtained by any combination of real $\mathbf{K}_\mathrm{T}$ and k_z. While a complete 2D photonic bandgap exists for $k_z = 0$, as illustrated by the example in Fig. 7.3-3, a photonic bandgap for all off-axis waves is not attainable in 2D periodic structures.

*B. Three-Dimensional Photonic Crystals

Crystal Structure

A 3D photonic crystal is generated by the placement of copies of a basic dielectric structure, such as a sphere or a cube, at points of a 3D lattice generated by the lattice vectors $\mathbf{R} = m_1\mathbf{a}_1 + m_2\mathbf{a}_2 + m_3\mathbf{a}_3$, where m_1, m_2, and m_3 are integers, and $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are primitive vectors defining the lattice unit cell. The overall structure is periodic

[†] See S. G. Johnson and J. D. Joannopoulos, Block-Iterative Frequency-Domain Methods for Maxwell's Equations in a Planewave Basis, *Optics Express*, vol. 8, pp. 173–190, 2001.

and its physical properties, such as the permittivity $\epsilon(\mathbf{r})$ and the impermeability $\eta(\mathbf{r}) = \epsilon_o/\epsilon(\mathbf{r})$, are invariant to translation by $\mathbf{R}$, so that

$$\eta(\mathbf{r} + \mathbf{R}) = \eta(\mathbf{r}) \tag{7.3-6}$$

for all positions $\mathbf{r}$.

This periodic function may therefore be expanded in a 3D Fourier series,

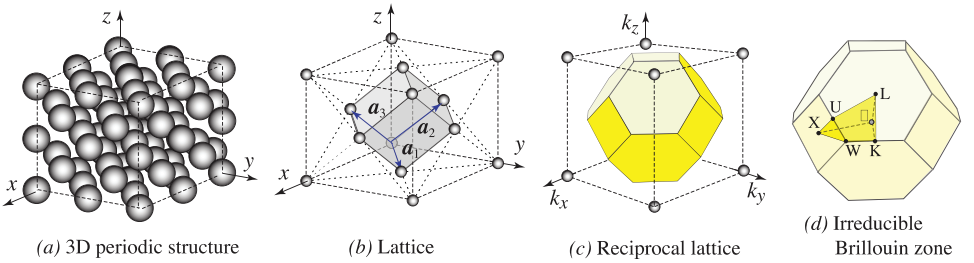
$$\eta(\mathbf{r}) = \sum_{\mathbf{G}} \eta_{\mathbf{G}} \exp(-j\,\mathbf{G}\cdot\mathbf{r}), \tag{7.3-7}$$

where $\mathbf{G} = \ell_1\mathbf{g}_1 + \ell_2\mathbf{g}_2 + \ell_3\mathbf{g}_3$ is a vector defined by the primitive vectors $\mathbf{g}_1$, $\mathbf{g}_2$, and $\mathbf{g}_3$, of another lattice, the **reciprocal lattice**, and ℓ_1, ℓ_2, and ℓ_3, are integers. The $\mathbf{g}$ vectors are related to the $\mathbf{a}$ vectors via

$$\mathbf{g}_1 = 2\pi\frac{\mathbf{a}_2\times\mathbf{a}_3}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3}, \quad \mathbf{g}_2 = 2\pi\frac{\mathbf{a}_3\times\mathbf{a}_1}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3}, \quad \mathbf{g}_3 = 2\pi\frac{\mathbf{a}_1\times\mathbf{a}_2}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3}, \tag{7.3-8}$$

so that $\mathbf{g}_1\cdot\mathbf{a}_1 = 2\pi$, $\mathbf{g}_1\cdot\mathbf{a}_2 = 0$, and $\mathbf{g}_1\cdot\mathbf{a}_3 = 0$, i.e., $\mathbf{g}_1$ is orthogonal to $\mathbf{a}_2$ and $\mathbf{a}_3$ and its length is inversely proportional to $\mathbf{a}_1$. Similar properties apply to $\mathbf{g}_2$ and $\mathbf{g}_3$. It can also be shown that $\mathbf{G}\cdot\mathbf{R} = 2\pi$.

If $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are mutually orthogonal, then $\mathbf{g}_1$, $\mathbf{g}_2$, and $\mathbf{g}_3$ are also mutually orthogonal and the magnitudes $g_1 = 2\pi/a_1$, $g_2 = 2\pi/a_2$, and $g_3 = 2\pi/a_3$ are the spatial frequencies associated with the periodicities in the three directions, respectively. An example of a 3D crystal lattice and its corresponding reciprocal lattice is shown in Fig. 7.3-4.



(a) 3D periodic structure *(b)* Lattice *(c)* Reciprocal lattice *(d)* Irreducible Brillouin zone

Figure 7.3-4 (a) A 3D periodic structure comprising dielectric spheres. (b) The spheres are placed at the points of a diamond (face-centered cubic) lattice for which $\mathbf{a}_1 = (a/\sqrt{2})(\hat{\mathbf{x}}+\hat{\mathbf{y}})$, $\mathbf{a}_2 = (a/\sqrt{2})(\hat{\mathbf{y}}+\hat{\mathbf{z}})$, and $\mathbf{a}_3 = (a/\sqrt{2})(\hat{\mathbf{x}}+\hat{\mathbf{z}})$, where a is the lattice constant. (c) The corresponding reciprocal lattice is a body-centered cubic lattice with a Brillouin zone indicated by the shaded volume, known as a Wigner–Seitz cell. (d) The irreducible Brillouin zone is the polyhedron whose corner points are marked by the crystallographic symbols ΓXULKW.

Bloch Modes

The modes of a 3D periodic medium are waves that maintain their shape upon translation by a lattice vector $\mathbf{R}$, changing only by a multiplicative constant of unity magnitude. These modes have the Bloch form $p_{\mathbf{K}}(\mathbf{r})\exp(-j\mathbf{K}\cdot\mathbf{r})$ where $p_{\mathbf{K}}(\mathbf{r})$ is a 3D periodic function, with the periodicity described by the same lattice vector $\mathbf{R}$; $\mathbf{K}$ is the **Bloch wavevector**; and $\hat{\mathbf{e}}$ is a unit vector in the direction of polarization. The Bloch mode is a traveling plane wave $\exp(-j\mathbf{K}\cdot\mathbf{r})$ modulated by a periodic function $p_{\mathbf{K}}(\mathbf{r})$.

Translation by $\mathbf{R}$ results in multiplication by a phase factor $\exp(-j\mathbf{K}\cdot\mathbf{R})$, which depends on $\mathbf{K}$.

Two modes with Bloch wavevectors $\mathbf{K}$ and $\mathbf{K}' = \mathbf{K} + \mathbf{G}$ are equivalent since $\exp(-j\mathbf{K}'\cdot\mathbf{R}) = \exp(-j\mathbf{K}\cdot\mathbf{R})$, i.e., translation by $\mathbf{R}$ is equivalent to multiplication by the same phase factor. This is because $\exp(-j\mathbf{G}\cdot\mathbf{R}) = \exp(-j2\pi) = 1$. Therefore, for the complete specification of all modes, we need only consider values of $\mathbf{K}$ within a finite volume of the reciprocal lattice, the Brillouin zone. The Brillouin zone is the volume of points that are closer to one specific reciprocal lattice point (the origin of the zone, denoted Γ) than to any other lattice point. Other symmetries of the lattice permit further reduction of that volume to the irreducible Brillouin zone, as illustrated by the example in Fig. 7.3-4.

Photonic Band Structure

To determine the ω-$\mathbf{K}$ dispersion relation for a 3D periodic medium, we begin with the eigenvalue problem described by the generalized Helmholtz equation (7.0-2). One approach for solving this problem is to generalize the Fourier method that was introduced in Sec. 7.2C for 1D periodic structures. By expanding the periodic functions $\eta(\mathbf{r})$ and $p_{\mathbf{K}}(\mathbf{r})$ in Fourier series, the differential equation (7.0-2) is converted into a set of algebraic equations leading to a matrix eigenvalue problem that can be solved numerically using matrix methods. As discussed at the end of Sec. 7.2C, we work with the magnetic field to ensure Hermiticity of the matrix representation.

Expanding the periodic function $p_{\mathbf{K}}(\mathbf{r})$ in the Bloch wave into a 3D Fourier series

$$p_{\mathbf{K}}(\mathbf{r}) = \sum_{\mathbf{G}} C_{\mathbf{G}}\, \exp(-j\,\mathbf{G}\cdot\mathbf{r}), \tag{7.3-9}$$

we write the magnetic-field vector in the Bloch form

$$\mathbf{H}(\mathbf{r}) = p_{\mathbf{K}}(\mathbf{r})\exp(-j\mathbf{K}\cdot\mathbf{r})\hat{\mathbf{e}} = \sum_{\mathbf{G}} C_{\mathbf{G}}\, \exp[-j(\mathbf{K}+\mathbf{G})\cdot\mathbf{r}]\,\hat{\mathbf{e}}. \tag{7.3-10}$$

For notational simplicity, the dependence of the Fourier coefficients $C_{\mathbf{G}}$ on the Bloch wavevector $\mathbf{K}$ is not explicitly indicated. Substituting (7.3-7) and (7.3-10) into (7.0-2), using the relation $\nabla\times\exp(-j\mathbf{K}\cdot\mathbf{r})\,\hat{\mathbf{e}} = -j(\mathbf{K}\times\hat{\mathbf{e}})\exp(-j\mathbf{K}\cdot\mathbf{r})$, and equating harmonic terms of the same spatial frequency yields

$$-\sum_{\mathbf{G}'}(\mathbf{K}+\mathbf{G})\times\left[(\mathbf{K}+\mathbf{G}')\times\hat{\mathbf{e}}\right]\eta_{\mathbf{G}-\mathbf{G}'}\,C_{\mathbf{G}'} = \frac{\omega^2}{c_o^2}\,C_{\mathbf{G}}\,\hat{\mathbf{e}}\,. \tag{7.3-11}$$

Forming a dot product with $\hat{\mathbf{e}}$ on both sides, and using the vector identity $\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C}) = -(\mathbf{B}\times\mathbf{A})\cdot\mathbf{C}$ leads to

$$\sum_{\mathbf{G}'} F_{\mathbf{G}\mathbf{G}'}\,C_{\mathbf{G}'} = \frac{\omega^2}{c_o^2}\,C_{\mathbf{G}}\,, \quad F_{\mathbf{G}\mathbf{G}'} = \left[(\mathbf{K}+\mathbf{G})\times\hat{\mathbf{e}}\right]\cdot\left[(\mathbf{K}+\mathbf{G}')\times\hat{\mathbf{e}}\right]\,\eta_{\mathbf{G}-\mathbf{G}'}\,. \tag{7.3-12}$$

The Helmholtz differential equation has now been converted into a set of linear equations for the Fourier coefficients $\{C_{\mathbf{G}}\}$. Since $\eta(z)$ is real, $\eta_{\mathbf{G}-\mathbf{G}'} = \eta^*_{\mathbf{G}'-\mathbf{G}}$, and the matrix $F_{\mathbf{G}\mathbf{G}'}$ is Hermitian. Hence, (7.3-12) represents an eigenvalue problem for a Hermitian matrix. For each Bloch wavevector $\mathbf{K}$, the eigenvalues ω^2/c_o^2 provide multiple values of ω, which are used to construct the ω-$\mathbf{K}$ diagram and the photonic band structure. The eigenvectors $\{C_{\mathbf{G}}\}$ determine the periodic function $p_{\mathbf{K}}(\mathbf{r})$ of the Bloch wave.

Examples of Structures

Spherical holes on a diamond lattice. An example of a 3D photonic crystal that has been shown to exhibit a complete 3D photonic bandgap comprises air spheres embedded in a high-index material at the points of a diamond (face-centered cubic) lattice (see Fig. 7.3-4). The radii of the air spheres are sufficiently large such that the spheres overlap, thereby creating intersecting veins. The calculated band structure, shown in Fig. 7.3-5, has a relatively wide complete 3D photonic bandgap between the two lowest bands. Silicon photonic crystals with an inverse-opal structure have been fabricated by inserting silicon into the voids of a self-assembled opal template comprising close-packed silica spheres; these are connected by small "necks" formed during sintering. As a final step the silica template is removed.[†]

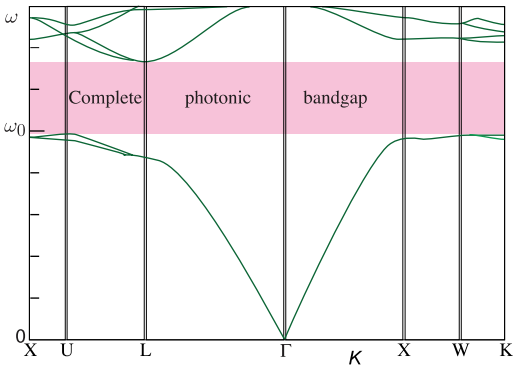


Figure 7.3-5 Calculated band structure of a 3D photonic crystal with a diamond (face-centered cubic) lattice of lattice constant a. The structure comprises air spheres of radius $0.325\,a$ embedded in a homogeneous material of refractive index $n = 3.6$. The complete photonic bandgap extends from approximately $\omega_0 = \pi c_o/a$ to $1.32\,\omega_0$ (see footnote on page 294).

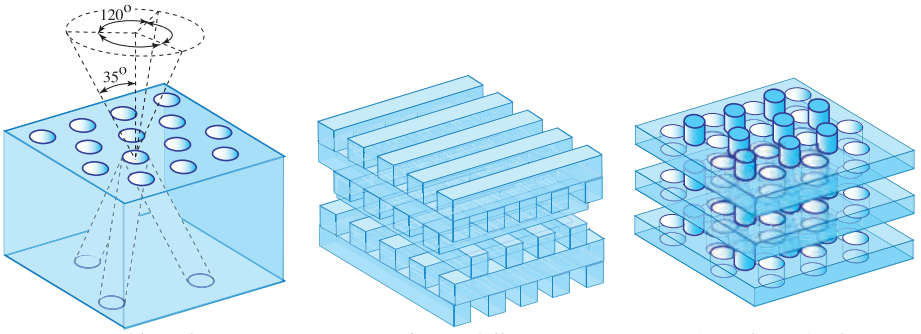
Yablonovite. The first experimental observation of a complete 3D photonic bandgap was made by Eli Yablonovitch and colleagues in 1991 using a variant of the diamond lattice structure now known as **Yablonovite**.[‡] This slanted-pore structure is fabricated by drilling a periodic array of cylindrical holes at specified angles in a dielectric slab. Three holes are drilled at each point of a 2D triangular lattice at the surface of the slab; the directions of the holes are parallel to three of the axes of the diamond lattice, as shown in Fig. 7.3-6(a). This structure exhibits a complete 3D photonic bandgap with a gap-midgap ratio of 0.19 when the refractive index of the material is $n = 3.6$.

Woodpile. Another 3D photonic-crystal structure, which is simpler to fabricate, is made of a 1D periodic stack of alternating layers, each of which is itself a 2D photonic crystal. For example, the woodpile structure illustrated in Fig. 7.3-6(b) uses layers of parallel logs with a stacking sequence that repeats itself every four layers. The orientation of the logs in adjacent layers is rotated 90°, and the logs are shifted by half the pitch every two layers. The resulting structure has a face-centered-tetragonal lattice symmetry. Fabricated using silicon micromachining, with a minimum feature size of 180 nm, this structure manifested a complete 3D photonic bandgap in the wavelength range $\lambda = 1.35$–$1.95\ \mu$m.[*]

[†] See A. Blanco, E. Chomski, S. Grabtchak, M. Ibisate, S. John, S. W. Leonard, C. Lopez, F. Meseguer, H. Miguez, J. P. Mondia, G. A. Ozin, O. Toader, and H. M. van Driel, Large-Scale Synthesis of a Silicon Photonic Crystal with a Complete Three-Dimensional Bandgap Near 1.5 Micrometres, *Nature*, vol. 405, pp. 437–440, 2000.

[‡] See E. Yablonovitch, T. J. Gmitter, and K. M. Leung, Photonic Band Structures: The Face-Centered Cubic Case Employing Non-Spherical Atoms, *Physical Review Letters*, vol. 67, pp. 2295–2298, 1991.

[*] See J. G. Fleming and S.-Y. Lin, Three-Dimensional Photonic Crystal with a Stop Band from 1.35 to 1.95 μm, *Optics Letters*, vol. 24, pp. 49–51, 1999.



(*a*) Yablonovite (*b*) Woodpile (*c*) Holes and poles

Figure 7.3-6 (*a*) The Yablonovite photonic crystal is fabricated by drilling cylindrical holes through a dielectric slab. At each point of a 2D triangular lattice at the surface, three holes are drilled along directions that make an angle of 35° with the normal and are separated azimuthally by 120°. (*b*) The woodpile photonic crystal comprises alternating layers of parallel rods, with adjacent layers oriented at 90°. (*c*) The holes-and-poles structure consists of alternating layers of 2D periodic structures: a layer of parallel cylindrical holes on a hexagonal lattice, followed by a layer of parallel rods lined up to fit between the holes.

Holes and poles. Yet another example is the holes-and-poles structure illustrated in Fig. 7.3-6(*c*). Here, two complementary types of 2D-periodic photonic-crystal slabs are used: dielectric rods in air and air holes in a dielectric. Fabricated in silicon, this structure exhibited a stop-band for all tilt angles in the $\lambda = 1.15$–1.6 μm telecommunications band.[†]

Both the holes-and-poles and the woodpile structures offer the opportunity to introduce arbitrary **point defects**, such as a missing hole or rod, thereby providing means for fabricating devices such as photonic-crystal waveguides (Sec. 9.5), photonic-crystal resonators (Sec. 11.4D), and specially controlled light emitters[‡] (Sec. 18.5C). Indeed, the ability to insert a defect at will may well be the most valuable feature of 2D and 3D photonic structures since 1D periodic media serve admirably as omnidirectional reflectors.

Fabrication Methods

In the early 1990s, photonic crystals were fabricated using adaptations of conventional semiconductor nanofabrication techniques. A decade later, by the early 2000s, a host of new ("bottom-up" and "top-down") techniques for the fabrication of 3D photonic crystals had been developed and implemented with varying degrees of success. The most prominent of these are colloidal self-assembly, holographic lithography, and direct writing via 3D multiphoton microlithography.

Submicrometer colloidal spheres of uniform size tend to spontaneously assemble on a face-centered cubic lattice. The resulting material is a synthetic opal, which serves as a template that can be impregnated with a semiconductor material such as silicon. Subsequent removal of the template yields a 3D photonic crystal, known as inverse opal, that consists of the semiconductor material containing periodic spheres of air. A challenge in using this method is growing the initial opal without forming polycrystalline segments, which suffer from deleterious lattice defects.

[†] See M. Qi, E. Lidorikis, P. T. Rakich, S. G. Johnson, J. D. Joannopoulos, E. P. Ippen, and H. I. Smith, A Three-Dimensional Optical Photonic Crystal with Designed Point Defects, *Nature*, vol. 429, pp. 538–542, 2004.

[‡] See S. P. Ogawa, M. Imada, S. Yoshimoto, M. Okano, and S. Noda, Control of Light Emission by 3D Photonic Crystals, *Science*, vol. 305, pp. 227–229, 2004.

In the 3D holographic-lithography approach, multiple laser beams generate interference patterns that illuminate a film of photoresist. Highly exposed regions of the photoresist become insoluble and the unexposed regions are washed away. The result is a periodic structure of cross-linked polymer containing air-filled voids. The use of liquid-crystal spatial light modulators to govern the phase properties of the various beams offers (dynamically tunable) interference patterns of arbitrary form. Four beams, generated by a single laser, are sufficient to create any desired photonic band structure.

In 3D multiphoton microlithography, femtosecond laser pulses are delivered via a lens to a particular location in a specially designed transparent polymeric material. The laser power is set at a level that effects multiphoton polymerization only in the vicinity of the focal region of the lens, where the optical intensity is sufficiently high (see Sec. 14.5B). The light is able to reach that region without polymerizing the intervening material because its intensity lies below the polymerization threshold outside the focal region. The photonic-crystal structure is fabricated by moving the focal point of the lens to all desired locations, thereby writing the three-dimensional microstructure. The strong thresholding behavior of the polymerization nonlinearity serves to enhance the resolution of the process beyond the diffraction limit.

Novel variations accommodate the fabrication of electrically responsive and graded-index (GRIN) photonic crystals.

READING LIST

Layered and Periodic Media

O. Stenzel, *The Physics of Thin Film Optical Spectra: An Introduction*, Springer-Verlag, 2nd ed. 2015.

A. B. Shvartsburg and A. A. Maradudin, *Waves in Gradient Metamaterials*, World Scientific/Imperial College Press, 2013.

R. Kashyap, *Fiber Bragg Gratings*, Academic Press, 2nd ed. 2010.

Š. Višňovský, *Optics in Magnetic Multilayers and Nanostructures*, CRC Press, 2006.

P. Yeh, *Optical Waves in Layered Media*, Wiley, 2005.

M. Nevière and E. Popov, *Light Propagation in Periodic Media: Differential Theory and Design*, CRC Press, 2003.

M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th expanded and corrected ed. 2002, Sec. 1.6.

W. C. Chew, *Waves and Fields in Inhomogeneous Media*, Van Nostrand Reinhold, 1990; IEEE Press, reprinted 1995.

A. Yariv and P. Yeh, *Optical Waves in Crystals: Propagation and Control of Laser Radiation*, Wiley, 1985, paperback reprint 2002.

L. Brillouin, *Wave Propagation in Periodic Structures: Electric Filters and Crystal Lattices*, Dover, 2nd ed. 1953, reprinted 2003.

Photonic Crystals

Q. Gong and X. Hu, *Photonic Crystals: Principles and Applications*, CRC Press/Taylor & Francis, 2013.

M. F. Limonov and R. M. De La Rue, *Optical Properties of Photonic Structures: Interplay of Order and Disorder*, CRC Press/Taylor & Francis, 2012.

D. W. Prather, S. Shi, A. Sharkawy, J. Murakowski, and G. J. Schneider, *Photonic Crystals: Theory, Applications, and Fabrication*, Wiley, 2009.

I. A. Sukhoivanov and I. V. Guryev, *Photonic Crystals: Physics and Practical Modeling*, Springer-Verlag, 2009.

M. Skorobogatiy and J. Yang, *Fundamentals of Photonic Crystal Guiding*, Cambridge University Press, 2009.

J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade, *Photonic Crystals: Molding the Flow of Light*, Princeton University Press, 1995, 2nd ed. 2008.

J.-M. Lourtioz, H. Benisty, V. Berger, J.-M. Gérard, D. Maystre, and A. Tchelnokov, *Photonic Crystals: Towards Nanoscale Photonic Devices*, Springer-Verlag, 2nd ed. 2008.

P. Markoš and C. M. Soukoulis, *Wave Propagation: From Electrons to Photonic Crystals and Left-Handed Materials*, Princeton University Press, 2008.

K. Sakoda, *Optical Properties of Photonic Crystals*, Springer-Verlag, 2nd ed. 2005.

K. Inoue and K. Ohtaka, eds., *Photonic Crystals: Physics, Fabrication and Applications*, Springer-Verlag, 2004.

S. Noda and T. Baba, eds., *Roadmap on Photonic Crystals*, Kluwer, 2003.

V. Kochergin, *Omnidirectional Optical Filters*, Kluwer, 2003.

Seminal and Review Articles

L. Nucara, F. Greco, and V. Mattoli, Electrically Responsive Photonic Crystals: A Review, *Journal of Materials Chemistry C*, vol. 3, pp. 8449–8467, 2015.

Q. Zhu, L. Jin, and Y. Fu, Graded Index Photonic Crystals: A Review, *Annalen der Physik*, vol. 527, pp. 205–218, 2015.

G. von Freymann, V. Kitaev, B. V. Lotsch, and G. A. Ozin, Bottom-Up Assembly Of Photonic Crystals, *Chemical Society Reviews*, vol. 42, pp. 2528–2554, 2013.

D. Xia, J. Zhang, X. He, and S. R. J. Brueck, Fabrication of Three-Dimensional Photonic Crystal Structures by Interferometric Lithography and Nanoparticle Self-Assembly, *Applied Physics Letters*, vol. 93, 071105, 2008.

T. Asano, B.-S. Song, Y. Akahane, and S. Noda, Ultrahigh-Q Nanocavities in Two-Dimensional Photonic Crystal Slabs, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 12, pp. 1123–1134, 2006.

Á. Blanco, P. D. García, D. Golmayo, B. H. Juárez, and C. López, Opals for Photonic Band-Gap Applications, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 12, pp. 1143–1150, 2006.

M. Qi, E. Lidorikis, P. T. Rakich, S. G. Johnson, J. D. Joannopoulos, E. P. Ippen, and H. I. Smith, A Three-Dimensional Optical Photonic Crystal with Designed Point Defects, *Nature*, vol. 429, pp. 538–542, 2004.

S. P. Ogawa, M. Imada, S. Yoshimoto, M. Okano, and S. Noda, Control of Light Emission by 3D Photonic Crystals, *Science*, vol. 305, pp. 227–229, 2004.

E. Yablonovitch, Photonic Crystals: Semiconductors of Light, *Scientific American*, vol. 285, no. 6, pp. 47–55, 2001.

Y. A. Vlasov, X. Z. Bo, J. C. Sturm, and D. J. Norris, On-Chip Natural Assembly of Silicon Photonic Bandgap Crystals, *Nature*, vol. 414, pp. 289–293, 2001.

S. G. Johnson and J. D. Joannopoulos, Block-Iterative Frequency-Domain Methods for Maxwell's Equations in a Planewave Basis, *Optics Express*, vol. 8, pp. 173–190, 2001.

A. Blanco, E. Chomski, S. Grabtchak, M. Ibisate, S. John, S. W. Leonard, C. Lopez, F. Meseguer, H. Miguez, J. P. Mondia, G. A. Ozin, O. Toader, and H. M. van Driel, Large-Scale Synthesis of a Silicon Photonic Crystal with a Complete Three-Dimensional Bandgap Near 1.5 Micrometres, *Nature*, vol. 405, pp. 437–440, 2000.

M. Campbell, D. N. Sharp, M. T. Harrison, R. G. Denning, and A. J. Turberfield, Fabrication of Photonic Crystals for the Visible Spectrum by Holographic Lithography, *Nature*, vol. 404, pp. 53–56, 2000.

H. B. Sun, S. Matsuo, and H. Misawa, Three-Dimensional Photonic Crystal Structures Achieved with Two-Photon-Absorption Photopolymerization of Resin, *Applied Physics Letters*, vol. 74, pp. 786–788, 1999.

J. G. Fleming and S.-Y. Lin, Three-Dimensional Photonic Crystal with a Stop Band from 1.35 to 1.95 μm, *Optics Letters*, vol. 24, pp. 49–51, 1999.

J. D. Joannopoulos, P. R. Villeneuve, and S. Fan, Photonic Crystals: Putting a New Twist on Light, *Nature*, vol. 386, pp. 143–149, 1997.

P. R. Villeneuve and M. Piché, Photonic Bandgaps in Periodic Dielectric Structures, *Progress in Quantum Electronics*, vol. 18, pp. 153–200, 1994.

S. John, Localization of Light, *Physics Today*, vol. 44, no. 5, pp. 32–40, 1991.

E. Yablonovitch, T. J. Gmitter, and K. M. Leung, Photonic Band Structures: The Face-Centered Cubic Case Employing Non-Spherical Atoms, *Physical Review Letters*, vol. 67, pp. 2295–2298, 1991.

S. John, Strong Localization of Photons in Certain Disordered Dielectric Superlattices, *Physical Review Letters*, vol. 58, pp. 2486–2489, 1987.

E. Yablonovitch, Inhibited Spontaneous Emission in Solid-State Physics and Electronics, *Physical Review Letters*, vol. 58, pp. 2059–2062, 1987.

PROBLEMS

7.1-2 **Beamsplitter Slab.** A lossless, dielectric slab of refractive index n and width d, oriented at 45° with respect to an incident beam, is used as a beamsplitter. Derive expressions for the transmittance and reflectance for TM polarization and sketch their spectral dependence. Compare the results with those expected for TE polarization, as provided in Example 7.1-6.

7.1-3 **Air Gap in Glass.** Determine the transmittance through a thin planar air gap of width $d = \lambda/2$ in glass of refractive index n. Assume (a) normal incidence, and (b) a TE wave incident at an angle greater than the critical angle. Can the wave penetrate (tunnel) through the gap?

7.1-4 **Multilayer Device in an Unmatched Medium.** The complex amplitude reflectance of a multilayer device is r_m when it is placed in a medium with refractive index n_1 matching its front layer. If the device is instead placed in a medium with refractive index n, show that the amplitude reflectance is $\mathrm{r} = (\mathrm{r}_b + \mathrm{r}_m)/(1 + \mathrm{r}_b\mathrm{r}_m)$, where $\mathrm{r}_b = (n - n_1)/(n + n_1)$ is the reflectance of the new boundary. Determine r in each of the following limiting cases: $\mathrm{r}_b = 0$, $\mathrm{r}_b = 1$, $\mathrm{r}_m = 0$, and $\mathrm{r}_m = 1$.

7.1-5 **Quarter-Wave Film: Angular Dependence of Reflectance.** Consider the quarter-wave antireflection coating described in Exercise 7.1-1. Derive an expression for the reflectance as a function of the angle of incidence.

7.1-6 **Transmittance of a Fabry–Perot Etalon.** The transmittance of a symmetric Fabry–Perot etalon was measured by using light from a tunable monochromatic light source. The transmittance versus frequency exhibits periodic peaks of period 150 MHz, each of width (FWHM) 5 MHz. Assuming that the medium within the resonator mirrors is a gas with $n = 1$, determine the length and finesse of the resonator. Assuming further that the only source of loss is associated with the mirrors, find their reflectances.

7.1-7 **Quarter-Wave and Half-Wave Stacks.** Derive expressions for the reflectance of a stack of N double layers of dielectric materials of equal optical thickness, $n_1 d_1 = n_2 d_2$, equal to $\lambda_o/4$ and $\lambda_o/2$.

7.1-8 **GaAs/AlAs Bragg Grating Reflector.** A Bragg grating reflector comprises N units of alternating layers of GaAs ($n_1 = 3.57$) and AlAs ($n_2 = 2.94$) of widths d_1 and d_2 equal to a quarter wavelength in each medium. The grating is placed in an extended GaAs medium. Calculate and plot the transmittance and reflectance of the grating as functions of N, for $N = 1, 2, \ldots, 10$, at a frequency equal to the Bragg frequency.

7.1-9 **Bragg Grating: Angular and Spectral Dependence of Reflectance.** Based on matrix algebra, determine the wave-transfer matrix and the reflectance of an N-layer alternating-layer dielectric Bragg grating. Use your program to verify the graphs presented in Fig. 7.1-12 and Fig. 7.1-13 for the spectral and angular dependence of the reflectance, respectively.

7.2-1 **Gap–Midgap Ratio.** Using a Fourier optics approach, determine the Bragg frequency and the gap–midgap ratio for the lowest bandgap of a 1D periodic structure comprising a stack of dielectric layers of equal optical thickness, with $n_1 = 1.5$ and $n_2 = 3.5$, and period $\Lambda = 2\ \mu$m. Assume that the wave travels along the axis of periodicity. Repeat the process for $n_1 = 3.4$ and $n_2 = 3.6$. Compare your results.

7.2-2 **Off-Axis Wave in 1D Periodic Medium.** Derive equations analogous to those provided in (7.2-24)–(7.2-28) for an off-axis wave traveling through a 1D periodic medium with a transverse wavevector k_x.

7.2-3 **Normal-to-Axis Wave in a 1D Periodic Medium.** Use the results of Prob. 7.2-2 to show that there are no bandgaps for a wave traveling along the lateral direction of a 1D periodic medium, i.e., for $K = 0$.

7.2-4 **Omnidirectional Reflector.** A periodic stack of double layers of dielectric materials with $n_1 d_1 = n_2 d_2$, $n_2 = 2n_1$ and $\Lambda = d_1 + d_2$ is to be used as an omnidirectional reflector in air. Plot the projected dispersion relation showing the light line for air (a diagram similar to Fig. 7.2-10). Determine the frequency range (in units of $\omega_{\mathcal{B}}$) for omnidirectional reflection.

CHAPTER

8

METAL AND METAMATERIAL OPTICS

8.1 SINGLE- AND DOUBLE-NEGATIVE MEDIA 306
- A. Wave Propagation in SNG and DNG Media
- B. Waves at Boundaries Between DPS, SNG, and DNG Media
- *C. Hyperbolic Media

8.2 METAL OPTICS: PLASMONICS 320
- A. Optical Properties of Metals
- B. Metal–Dielectric Boundary: Surface Plasmon Polaritons
- C. The Metallic Nanosphere: Localized Surface Plasmons
- D. Optical Antennas

8.3 METAMATERIAL OPTICS 334
- A. Metamaterials
- B. Metasurfaces

*8.4 TRANSFORMATION OPTICS 343
- A. Transformation Optics
- B. Invisibility Cloaks

Paul Karl Ludwig Drude (1863–1906), a German physicist, worked toward integrating optics with Maxwell's electromagnetics. He forged a theory, commonly known as the Drude model, for describing the behavior of electrons in metals.

Viktor Georgievich Veselago (born 1929), a Russian physicist, established theoretically in the 1960s that materials whose electric permittivity and magnetic permeability were both negative would exhibit unexpected and unusual properties.

Sir John Pendry (born 1943), a British theoretical physicist, showed in 2000 that a slab of negative-index material acts as a lens with theoretically perfect focus. In 2006, he proposed the use of transformation optics for creating invisibility cloaks.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Visible light cannot propagate through highly conductive media such as metals. When a light beam crosses the boundary into such a medium, its intensity rapidly diminishes within a short distance known as the penetration depth, which can be substantially smaller than a wavelength. A metallic surface acts rather like a mirror from which light is fully reflected back into the contiguous dielectric medium whence it came.

In the previous chapters of this book, metallic components have indeed played the role of simple mirrors. It turns out, however, that metals *can* support light waves, provided that they travel along the boundaries of the metal, in regions confined to subwavelength dimensions. Such light waves travel along a metal surface in the form of a guided surface wave. It may propagate *on*, but not *in*, metal wires, and it may be guided by subwavelength metallic structures configured as integrated-photonic circuits. Such structures may also serve as resonators within which light may be confined, or from which light can scatter strongly at specific resonance frequencies.

Advances in nanotechnology permit such subwavelength metallic structures to be embedded and distributed within dielectric materials to form synthetic photonic materials known as **metamaterials**. Unavailable in nature, these materials have the merit that they may be endowed with highly useful optical properties. Metamaterials offer numerous novel applications and have come to play an important role in photonics.

Metals and metamaterials possess a number of unique optical characteristics:

- A metal–dielectric boundary can serve as a waveguide that supports an optical wave traveling along the boundary. Such a surface wave, known as a **surface plasmon polariton** (**SPP**), is highly confined to the vicinity of the boundary and is accompanied by an electric-charge-density longitudinal wave (a plasma wave) of the same frequency that concomitantly travels along the metal surface. The tight confinement and short wavelength of SPPs can provide a significant increase in local field intensity. Biosensing applications are predicated on the sensitivity of the SPP to the properties of the surrounding dielectric media.
- Metallic structures of subwavelength dimensions (e.g., nanospheres) embedded in dielectric media support plasmonic oscillations at their boundaries. These oscillations, called **localized surface plasmons** (**LSPs**), exhibit resonance when the excitation frequency matches the resonance frequency of the structure, which typically falls in the visible or ultraviolet region of the spectrum. As a result of resonantly enhanced absorption and scattering, such nanoparticles can exhibit intense colors in the visible, both in transmission and reflection. This metal-optics technology, known as **plasmonics**, finds use in applications ranging from fabricating stained glass to probing the dielectric properties of a host medium.
- Metallic nanostructures may be fabricated in the form of circuit elements and **nanoantennas** that are analogous to their radiowave and microwave counterparts, but that operate in the infrared and visible. This plasmonics technology seeks to couple the domains of highly integrated compact electronics (dimensions < 100 nm) and optical-frequency ultrafast photonics (bandwidths > 100 THz), and is expected to find use in applications such as intrachip interconnects.
- An array of metallic structures printed at the boundary between two dielectric media creates a **metasurface**, which exhibits unique optical properties that are dictated by the shape of the elements and the geometrical configuration of the array. Behaving much like arrays of optical antennas, metasurfaces bend light in unusual ways that do not obey the ordinary laws of reflection and refraction prevailing at dielectric boundaries.
- Materials with embedded metallic and dielectric structures of subwavelength dimensions, distributed at wavelength or subwavelength distances throughout the volume, exhibit novel electrical and magnetic properties that result from electric

charges and currents induced in the constituent conductive metal components. Such metamaterials may be engineered to exhibit remarkable optical properties such as **negative refractive index**, whereby a negative angle of refraction is imparted at a boundary with a conventional dielectric medium.

- Metamaterials may be designed with spatially varying (graded) properties that transform optical waves in unusual ways so that they can, for example, wrap around objects and render them invisible. A particularly intriguing application is **optical cloaking**.

The optical properties of metals and metamaterials are described by the electromagnetic theory of light introduced in Chapter 5, just as for dielectric media. The principal distinction is that the electric permittivity ϵ and the magnetic permeability μ may take on negative values for metals and metamaterials. Materials in which one of these parameters is negative are referred to as **single-negative (SNG) media**, while those in which both are negative are known as **double-negative (DNG) media**. Conventional lossless dielectric media are **double-positive (DPS) media**.

This Chapter

The chapter opens with an introductory section that examines some of the optical properties of SNG and DNG media (Sec. 8.1). The existence of surface guided waves at a DPS-SNG boundary is demonstrated, negative refraction at a DPS-DNG boundary is highlighted, and some of the potential applications of negative-refractive-index materials are described.

Section 8.2 provides an introduction to metal optics and plasmonics. It begins by presenting a physical model for the optics of metals, the **Drude model**, and shows that a metal can behave as a SNG material under certain conditions. SPP waves at a metal–dielectric boundary are examined. The Rayleigh scattering of light from a metal nanosphere is considered and contrasted with Rayleigh scattering from a dielectric sphere (as discussed in Sec. 5.6B). Optical antennas are briefly described.

Section 8.3 considers metamaterial optics. It deals with the various shapes and organizational patterns of metal objects that can be embedded in a dielectric medium, or on a dielectric surface, to endow it with particular macroscopic properties: $\epsilon(\omega)$ and $\mu(\omega)$ with real and imaginary components of desired signs. Of special interest is the design of negative-refractive-index media.

Section 8.4 introduces transformation optics, a mathematical tool that facilitates the design of special graded optical materials that guide light along desired trajectories. The goal is often to fabricate metamaterials that effect a desired result, such as creating trajectories that bypass an object so that it is rendered invisible (cloaked).

Since metamaterial optics, one of the principal topics of this chapter, involves the interaction of light with components of subwavelength (nanometer) scale, it lies in the domain of **nanophotonics**, also called **nano-optics**. Previous chapters, in contrast, have focused principally on the propagation of light through dielectric materials and components with dimensions substantially greater than the wavelength of light, which is the domain of **bulk optics**.

Upcoming chapters on guided-wave optics (Chapter 9), fiber optics (Chapter 10), and resonator optics (Chapter 11), have traditionally dealt with light propagating through structures whose dimensions are of the order of micrometers, a domain known as **micro-optics**. In recent years, however, the reach of nano-optics, via plasmonics, has penetrated these domains, as illustrated by the following examples: The ability of metal–dielectric structures to support surface plasmon polaritons at subwavelength spatial scales has led to the use of plasmonic waveguides, as reported in Chapter 9. Similarly, the ability of metallic structures of subwavelength dimensions to support localized surface plasmons at their boundaries with dielectrics has led to the development of plasmonic resonators and plasmonic nanolasers, considered in Chapters 11 and 18, respectively.

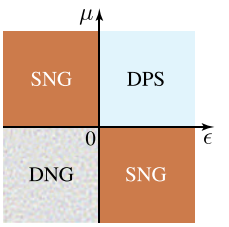
8.1 SINGLE- AND DOUBLE-NEGATIVE MEDIA

As described in Chapter 5, the propagation of an electromagnetic wave through a linear, isotropic medium is governed by the electric permittivity ϵ and magnetic permeability μ of the material. In general, these quantities are frequency-dependent and complex-valued. Wave properties, such as the propagation constant, velocity, attenuation coefficient, impedance, and dispersion relation, can be readily determined from ϵ and μ. The signs of the real and imaginary components of ϵ and μ at a given frequency govern the various regimes of wave propagation:

- For media in which μ is real and positive (indicating that there is neither magnetic absorption nor amplification), the wave-propagation characteristics depend on the signs of the real and imaginary parts of ϵ, as set forth in Chapter 5:
  - If ϵ is real, the medium exhibits neither dielectric absorption nor gain (it is lossless and passive), and waves propagate without attenuation, as described in Secs. 5.1–5.4.
  - In the presence of absorption (Sec. 5.5), ϵ is complex and $\text{Im}\{\epsilon\} \equiv \epsilon''$ is negative, but $\text{Re}\{\epsilon\} \equiv \epsilon'$ can be either positive or negative. For the resonant medium described in Sec. 5.5C, for example, the imaginary part of the electric susceptibility, $\text{Im}\{\chi\} \equiv \chi''$, is negative (see Fig. 5.5-6) and therefore so too is $\epsilon'' = \epsilon_o\chi''$. The behavior of the real part of the susceptibility χ', also displayed in Fig. 5.5-6, is such that $\epsilon' = \epsilon_o(1 + \chi')$ is positive for frequencies below the resonance frequency but may be negative above the resonance frequency.
  - For an active medium that exhibits gain, such as a laser medium, χ'' is positive (gain represents negative absorption; see Sec. 15.1A). In this case, ϵ'' is positive, while ϵ' may be either positive or negative.
- Similarly, for media with real and positive ϵ, the magnetic properties, described by μ, dictate the nature of wave propagation. Magnetic media, including media with metal components that carry induced electric currents and generate magnetic fields, generally have complex values of μ, with real and imaginary parts that may be either positive or negative.
- In the most general case, the manner in which the signs of the real and imaginary components of ϵ and μ dictate the characteristics of wave propagation is more subtle.

In the exposition that follows, we confine ourselves principally to lossless and passive media, in which there is neither absorption nor gain, indicating that we are, for example, away from dielectric and magnetic resonances. Under these circumstances, both ϵ and μ are real, and their signs may be positive or negative at a given frequency. Four regimes ensue:

- Double-positive (**DPS**) materials (both ϵ and μ are positive). These materials are transparent and have positive refractive index. Ordinary dielectric media fall into this category.
- Single-negative (**SNG**) materials (either ϵ or μ is negative). These materials are opaque but they support optical surface waves at boundaries with DPS materials. As will become apparent in Sec. 8.2, metals such as gold and silver exhibit negative ϵ while maintaining positive μ in the infrared and visible spectral regions. Ferrites have positive ϵ and negative μ at microwave frequencies.



DPS, SNG, and DNG media are defined by the signs of ϵ and μ.

- Double-negative (**DNG**) materials (both ϵ and μ are negative). These materials, also called **left-handed media** for reasons that will become apparent in the sequel,

are transparent and have negative refractive index, signifying that the application of Snell's law at a DPS-DNG boundary results in a negative angle of refraction. The ramifications of this property for optical components with multiple boundaries are most interesting. Metamaterials may be designed to exhibit such properties in specific frequency bands.

We begin by examining the optical properties of linear, lossless, passive media in the DPS, SNG, and DNG regimes. The ramifications of the presence of loss in a DNG medium are considered at the end of Sec. 8.1A. Initially we pay little heed to whether such media exist naturally or must be fabricated synthetically; this issue is addressed in Secs. 8.2 and 8.3.

A. Wave Propagation in SNG and DNG Media

The propagation of a monochromatic electromagnetic wave in a linear, homogeneous, and isotropic medium with electric permittivity ϵ and magnetic permeability μ is described in Secs. 5.3 and 5.4. Maxwell's equations (5.3-12)–(5.3-15), along with the Helmholtz equation (5.3-16), are applicable for arbitrary complex ϵ and μ, regardless of the signs of their real and imaginary parts.

For simplicity, we consider a monochromatic plane wave with electric and magnetic complex-amplitude vectors given by $\mathbf{E}(\mathbf{r}) = \mathbf{E}_0 \exp(-j\mathbf{k}\cdot\mathbf{r})$ and $\mathbf{H}(\mathbf{r}) = \mathbf{H}_0 \exp(-j\mathbf{k}\cdot\mathbf{r})$, respectively, and with wavevector $\mathbf{k}$. Maxwell's equations then require

$$\mathbf{k}\times\mathbf{H}_0 = -\omega\,\epsilon\,\mathbf{E}_0 \tag{8.1-1}$$

$$\mathbf{k}\times\mathbf{E}_0 = \omega\,\mu\,\mathbf{H}_0, \tag{8.1-2}$$

in accordance with (5.4-1)–(5.4-4). The associated wavenumber (magnitude of the vector $\mathbf{k}$) is

$$k = \omega\sqrt{\epsilon\,\mu}\,, \tag{8.1-3}$$

and the impedance (ratio of the magnitudes of E_0 and H_0) is given by

$$\eta = \frac{\omega\mu}{k} = \sqrt{\frac{\mu}{\epsilon}}\,, \tag{8.1-4}$$

as specified in (5.4-5). Equation (8.1-1) indicates that $\mathbf{E}_0$ is orthogonal to both $\mathbf{k}$ and $\mathbf{H}_0$, while (8.1-2) indicates that $\mathbf{H}_0$ is orthogonal to both $\mathbf{k}$ and $\mathbf{E}_0$, so that the three vectors form a mutually orthogonal set. For fixed orthogonal directions of the fields $\mathbf{E}_0$ and $\mathbf{H}_0$, the wavevector $\mathbf{k}$ is thus orthogonal to the plane defined by these field vectors, but its actual direction depends on the signs of ϵ and μ, as we shall see shortly.

Since k is in general complex, we write $k = \beta - j\gamma$, where β and γ are real, so that

$$\beta - j\gamma = \omega\sqrt{\epsilon\,\mu}\,. \tag{8.1-5}$$

The propagation constant $\beta = \omega/c$ determines both the wave velocity $c = c_o/n$ and the refractive index n, whereas γ represents the field attenuation coefficient ($\gamma = \frac{1}{2}\alpha$, where α is the intensity attenuation coefficient; see Sec. 5.5A).

We now consider the implications of these equations for media in which ϵ and μ are real, where either or both may be negative.

DPS Medium

The double-positive (DPS) medium provides a simple and familiar benchmark. Both ϵ and μ are positive, so that k and η are real, whereupon

$$\gamma = 0, \qquad \beta = nk_o\,, \qquad n = \sqrt{\frac{\epsilon}{\epsilon_o}\frac{\mu}{\mu_o}}\,, \qquad \eta = \sqrt{\frac{\mu}{\epsilon}}\,. \tag{8.1-6}$$

As discussed in Sec. 5.4A, such media support transverse electromagnetic (TEM) waves for which the vectors $\mathbf{E}_0$, $\mathbf{H}_0$, and $\mathbf{k}$ are mutually orthogonal and form a right-handed system, as illustrated in Fig. 8.1-1(a). The Poynting vector $\mathbf{S} = \frac{1}{2}\mathbf{E}_0 \times \mathbf{H}_0^*$ points along the same direction as the wave vector $\mathbf{k}$, and the intensity of the wave (power flow per unit area) is given by $I = \text{Re}\{S\} = |E_0|^2/2\eta$.

SNG Medium

In a single-negative (SNG) medium, either ϵ or μ is negative so that k and η are both imaginary, whereupon (8.1-5) provides

$$\gamma = \omega\sqrt{|\epsilon||\mu|}, \qquad \beta = 0, \qquad \eta = j\sqrt{\frac{|\mu|}{|\epsilon|}}\,. \tag{8.1-7}$$

These parameters correspond to an exponentially decaying field that behaves as $\exp(-\gamma z)$, where z is the propagation distance. Since $\beta = 0$, a SNG medium does not support propagating waves. The optical intensity is attenuated by the factor e^{-1} at a depth $d_p = 1/2\gamma = \lambda_o/4\pi\sqrt{|\epsilon/\epsilon_o|\,|\mu/\mu_o|}$. The quantity d_p is known as the **penetration depth** or **skin depth**.[†] The imaginary impedance η indicates that there is a $\pi/2$ phase shift between the electric and magnetic fields. Moreover, the Poynting vector $\mathbf{S} = \frac{1}{2}\mathbf{E}_0 \times \mathbf{H}_0^*$ is imaginary so that the intensity $I = \text{Re}\{S\} = 0$, indicating that no power is transported through such a medium.

DNG Medium

In a double-negative (DNG) medium, both ϵ and μ are negative so that (8.1-5) leads to $k = \omega\sqrt{|\epsilon||\mu|}$, which is real, whereupon

$$\gamma = 0, \qquad \beta = nk_o, \qquad n = -\sqrt{\frac{|\epsilon|}{\epsilon_o}\frac{|\mu|}{\mu_o}}, \qquad \eta = \sqrt{\frac{|\mu|}{|\epsilon|}}\,, \tag{8.1-8}$$

indicating that the refractive index is negative. Since $\gamma = 0$, the medium sustains wave propagation without attenuation. The choice of signs for the square roots in (8.1-8) is established by examining the directions of the vectors $\mathbf{E}_0$, $\mathbf{H}_0$, and $\mathbf{k}$, which may be determined directly from Maxwell's equations. For the DNG medium, (8.1-1) and (8.1-2) yield

$$\mathbf{k} \times \mathbf{H}_0 = \quad \omega\,|\epsilon|\,\mathbf{E}_0 \tag{8.1-9}$$

$$\mathbf{k} \times \mathbf{E}_0 = -\omega\,|\mu|\,\mathbf{H}_0. \tag{8.1-10}$$

As with the DPS medium, the vectors $\mathbf{E}_0$, $\mathbf{H}_0$, and $\mathbf{k}$ are mutually orthogonal. However, the reversal of signs in (8.1-9) and (8.1-10), relative to those in (8.1-1) and (8.1-2)

[†] The penetration depth is sometimes defined as the distance over which the field, rather than the intensity, is attenuated by a factor e^{-1}, in which case $d_p' = 1/\gamma$.

for the DPS medium, is tantamount to exchanging the roles of the electric and magnetic fields.

Figure 8.1-1 illustrates the directions of flow of the wavefronts and power in DPS and DNG media. Comparing Figs. 8.1-1(a) and (c), it is apparent that $\mathbf{E}_0$, $\mathbf{H}_0$, and $\mathbf{k}$ in a DPS medium form the usual right-handed set of vectors, whereas in a DNG medium they form a left-handed set (the medium is then said to be **left-handed**). This has profound implications since it signifies that in a DNG medium the complex Poynting vector $\mathbf{S} = \frac{1}{2}\mathbf{E}_0 \times \mathbf{H}_0^*$ introduced in (5.3-10) is antiparallel to the wavevector $\mathbf{k}$. Since the impedance η positive, the wavenumber k is taken to be negative, and therefore so too is the refractive index n, as stipulated in (8.1-8). The DNG material is therefore a **negative-index material** (**NIM**). The implications of negative refractive index will be considered in Sec. 8.1B.

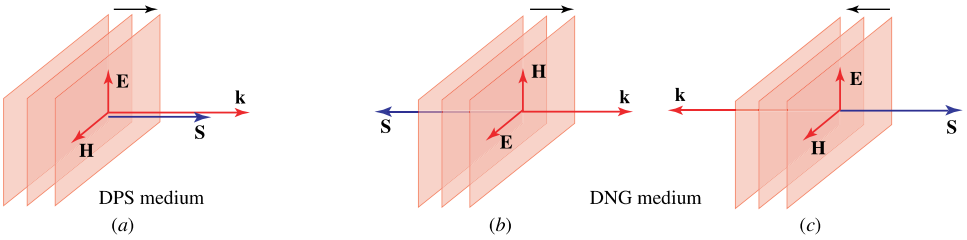


Figure 8.1-1 (a) Plane wave propagating in an ordinary double-positive (DPS) medium. The vectors **E**, **H**, and **k** form a right-handed set and the wavefronts travel in the same direction as the power flow. (b) Plane wave propagating in a double-negative (DNG) medium. The vectors **E**, **H**, and **k** form a left-handed set and the wavefronts travel in a direction opposite to that of the power flow. (c) Equivalent representation to that depicted in (b), obtained by effecting a 90° rotation about the horizontal (**S** and **k**) axis that renders **E** vertical, followed by a 180° rotation about the new vertical (**E**) axis.

Media with Complex ϵ and μ

The prospects of ϵ and μ both being *real* and *negative* at some frequency are remote. For example, if both parameters are described by a resonant-medium model, such as that considered in Sec. 5.5C, then throughout the frequency range over which the real part is negative, the imaginary part cannot be zero, by virtue of the Kramers–Kronig relations (see Sec. 5.5B and Sec. B.1 of Appendix B). It turns out, however, that left-handedness, and thus negative refractive index, can be exhibited in conjunction with absorption.

For absorptive media, $\epsilon = \epsilon' + j\epsilon''$ and $\mu = \mu' + j\mu''$ are complex with negative imaginary parts ϵ'' and μ'', respectively. We now demonstrate that if the real parts, ϵ' and μ', are both negative, the medium is indeed left-handed even for nonvanishing ϵ'' and μ''. As previously, consider a wave $\exp(-j\beta z)\exp(-\gamma z)$ with positive γ so that it decays along the $+z$ direction. The propagation constant β may be obtained by writing (8.1-5) in the form $(\beta - j\gamma)^2 = \omega^2(\epsilon' + j\epsilon'')(\mu' + j\mu'')$ and then matching the imaginary parts to obtain $2\gamma\beta = \omega^2(-\mu''\epsilon' - \epsilon''\mu')$. If both ϵ' and μ' are negative, then β is negative, and so too is the refractive index. The wavefront then travels in the $-z$ direction, opposite to the direction of decay. We now proceed to show that the power flow is always in the same direction as the power decay, so that the direction of the wavevector is opposite to that of the power flow and the medium is left-handed.

Power flow is determined by the Poynting vector $\text{Re}\{\frac{1}{2}\mathbf{E} \times \mathbf{H}^*\}$. Since $E_0 = \eta H_0$, where $\eta = \sqrt{\mu/\epsilon}$ is the characteristic impedance, and since $\text{Re}\{\eta\}$ is always positive for passive media, the power flow must be in the $+z$ direction for the field configuration

shown in Fig. 8.1-1(*c*). It can be explicitly shown that $\text{Re}\{\eta\} > 0$ by writing $\arg\{\eta\} = \frac{1}{2}(\arg\{\mu\} - \arg\{\epsilon\})$; since ϵ'' and μ'' are both negative, we have $\pi < \arg\{\epsilon\} < 2\pi$ and $\pi < \arg\{\mu\} < 2\pi$ whereupon $-\frac{1}{2}\pi < \arg\{\eta\} < \frac{1}{2}\pi$ so that $\text{Re}\{\eta\} > 0$.

Although the condition that the real part of both ϵ and μ be negative is *sufficient* for achieving left handedness, it is not *necessary*. It is possible for left-handedness to be exhibited for an absorptive medium with only one of the two parameters, ϵ' or μ', being negative, indicating that the class of left-handed media transcends that of double-negative media. The definitive *necessary and sufficient* condition for left-handedness turns out to be[†]

$$\epsilon'/|\epsilon| + \mu'/|\mu| < 0. \tag{8.1-11}$$

Materials for which both ϵ and μ are real, but only one is negative, do not satisfy (8.1-11) and therefore cannot be left-handed. Nor do media for which one of the material parameters, ϵ or μ, is real and positive, whatever the real and imaginary values of the other. It is clear, then, that nonmagnetic materials cannot be left-handed.

B. Waves at Boundaries Between DPS, SNG, and DNG Media

We now consider waves at boundaries between DPS and SNG media, and between DPS and DNG media.

Reflection at a DPS-SNG Boundary

A wave from an ordinary DPS medium that impinges on its boundary with a SNG medium is fully reflected since it cannot propagate through the latter. Since the impedance of the DPS medium is real, and that of the SNG medium is imaginary, the magnitude of the reflection coefficient must be unity (see Sec. 6.2). Such reflection is accompanied by the creation of an evanescent field in the SNG medium; this bears a resemblance to total internal reflection at the boundary between two dielectric (DPS) media. However, total internal reflection occurs only for angles of incidence greater than the critical angle whereas reflection at the DPS–SNG boundary is total regardless of the angle of incidence, as illustrated in Fig. 8.1-2. Reflection at the DPS–SNG boundary is therefore more akin to that at a perfect mirror.

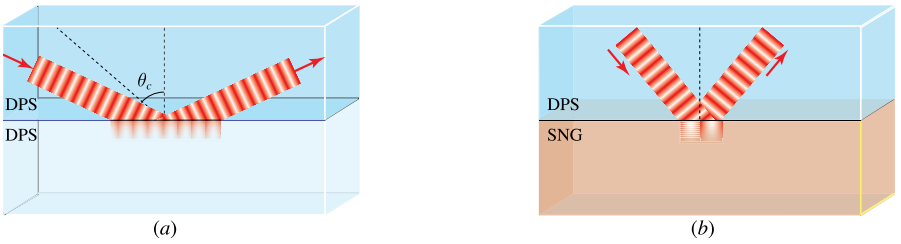


Figure 8.1-2 (*a*) Total internal reflection at a boundary between two DPS media takes place for angles of incidence greater than the critical angle θ_c. (*b*) Total reflection at a DPS–SNG boundary occurs for all angles of incidence. In both cases, an evanescent wave is created in the vicinity of the boundary.

[†] See R. A. Depine and A. Lakhtakia, A New Condition to Identify Isotropic Dielectric–Magnetic Materials Displaying Negative Phase Velocity, *Microwave and Optical Technology Letters*, vol. 41, pp. 315–316, 2004.

Surface Waves at a DPS-SNG Boundary

In the limit of grazing incidence (angle of incidence approaching 90°) at a DPS–SNG boundary, a **surface wave** may be created that propagates along the boundary but is evanescent on both sides of it. As a specific example, consider the boundary between a medium with positive electric permittivity ϵ_1 (medium 1) and another with negative permittivity ϵ_2 (medium 2). Both media are assumed to have the same (positive) magnetic permeability μ and both are lossless and passive so that ϵ_1, ϵ_2, and μ are real.

We proceed to demonstrate that such a DPS–SNG boundary can support a surface guided wave that travels along the boundary without changing its shape, as illustrated in Fig. 8.1-3(a). We assume that the wave is a TM wave, for which the magnetic field is parallel to the boundary and orthogonal to the direction of propagation. Each of the three field components H_x, E_y, and E_z varies as

$$\exp(-\gamma_1 y)\exp(-j\beta z), \quad y > 0 \text{ (medium 1)} \tag{8.1-12}$$

$$\exp(+\gamma_2 y)\exp(-j\beta z), \quad y < 0 \text{ (medium 2)}, \tag{8.1-13}$$

where β is a common propagation constant and γ_1 and γ_2 are positive field extinction coefficients. To satisfy the Helmholz equation (5.3-16) in each medium, we must have

$$-\gamma_1^2 + \beta^2 = \omega^2\mu\epsilon_1, \qquad -\gamma_2^2 + \beta^2 = \omega^2\mu\epsilon_2. \tag{8.1-14}$$

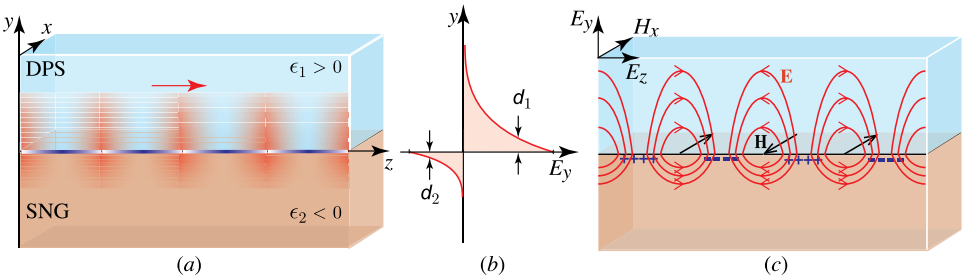


Figure 8.1-3 (a) Schematic representation of an optical surface wave traveling along the boundary between a medium of positive electric permittivity (DPS medium, such as a dielectric material) and another of negative permittivity (SNG medium, such as a metal below the plasma frequency). The associated longitudinal surface-charge wave is shown in dark blue; the width of each segment represents its wavelength, which is given by $2\pi/\beta$. (b) The electric field E_y as a function of the distance y from the boundary. (c) Electric and magnetic field lines and the associated charge. The plasmon wave penetrates to a depth d_1 and d_2 in the DPS and SNG media, respectively, and extends a distance d_b along the boundary.

The amplitudes of the three field components in each medium are related by Maxwell's equations, and the amplitudes in different media are related by the boundary conditions. Since the H_x component must be continuous, this component has a common amplitude, say H_0, in both media. Maxwell's equations (8.1-1) dictate that the amplitudes of the E_y components in the two media are $(-\beta/\omega\epsilon_1)H_0$ and $(-\beta/\omega\epsilon_2)H_0$. The condition that $D_y = \epsilon E_y$ is continuous at the boundary is therefore automatically satisfied. The amplitudes of the E_z components are $(-\gamma_1/j\omega\epsilon_1)H_0$ in medium 1 and $(\gamma_2/j\omega\epsilon_2)H_0$ in medium 2. Since the E_z components must be continuous at the boundary, it follows that

$$-\frac{\gamma_1}{\epsilon_1} = \frac{\gamma_2}{\epsilon_2}. \tag{8.1-15}$$

Because γ_1 and γ_2 are positive, this can only be attained if ϵ_1 and ϵ_2 have opposite signs. We therefore conclude that such a surface wave cannot exist at the boundary

between two media with positive electric permittivities, but may exist if the media have electric permittivities of opposite sign.

Since ϵE_y is continuous and ϵ changes sign at the boundary, E_y must also reverse sign at the boundary, as shown in Fig. 8.1-3(b). This in turn requires the existence of surface electric charge in the form of a charge-density longitudinal wave that oscillates at the optical frequency ω. The field lines and the charge distribution are illustrated in Fig. 8.1-3(c). The combination of the charge-density wave and the optical wave is known as a **surface plasmon polariton** (**SPP**).[†]

The properties of the SPP surface wave may be deduced from γ_1 and γ_2, in accordance with (8.1-14):

$$\beta = n_{\rm b} k_o\,, \qquad n_{\rm b} = \sqrt{\frac{\epsilon_{\rm b}}{\epsilon_o}}\,, \qquad \epsilon_{\rm b} = \frac{\epsilon_1 \epsilon_2}{\epsilon_1 + \epsilon_2}\,, \tag{8.1-16}$$

$$\gamma_1 = \sqrt{\frac{-\epsilon_1^2}{\epsilon_o(\epsilon_1 + \epsilon_2)}}\; k_o\,, \qquad \gamma_2 = \sqrt{\frac{-\epsilon_2^2}{\epsilon_o(\epsilon_1 + \epsilon_2)}}\; k_o\,, \tag{8.1-17}$$

SPP Wave
DPS-SNG Boundary

where $k_o = \omega\sqrt{\epsilon_o \mu}$ is the free-space wavenumber; and $n_{\rm b}$ and $\epsilon_{\rm b}$ are, respectively, the refractive index and electric permittivity associated with the SPP (the subscript "b" signifies "boundary"). For the surface wave to be a traveling wave, β must be real, which requires that $\epsilon_{\rm b}$ be positive. This is possible only if $|\epsilon_2| > \epsilon_1$, in which case the SPP surface wave exists. This same condition also renders γ_1 and γ_2 positive, as required. The properties of the surface wave, summarized below, are substantially influenced by the ratio $|\epsilon_2|/\epsilon_1$:

- The velocity is $c_o/n_{\rm b}$, and the propagation wavelength, called the plasmon wavelength, is $\lambda_o/n_{\rm b}$. If $|\epsilon_2| \approx \epsilon_1$, then $n_{\rm b}$ is large, the wave is slow, and the plasmon wavelength is much smaller than the free-space wavelength λ_o.
- The field extinction coefficient in the SNG medium, γ_2, is greater than that in the DNG medium, γ_1, by the factor $|\epsilon_2|/\epsilon_1$, which is greater than unity. The penetration depth in the SNG medium, $d_2 = 1/2\gamma_2$, is therefore always smaller than that in the DPS medium, $d_1 = 1/2\gamma_1$, by the same factor, as depicted in Fig. 8.1-3. If $|\epsilon_2| \approx \epsilon_1$, then $\epsilon_1 + \epsilon_2$ is small and both penetration depths are significantly smaller than the free-space wavelength λ_o so that the SPP is highly confined to the boundary. This is a remarkable property that can be exploited in many applications.
- The optical power flow in the two media may be determined from $\text{Re}\{\mathbf{S}\}$, where $\mathbf{S} = \frac{1}{2}\mathbf{E} \times \mathbf{H}^*$ is the complex Poynting vector [see (5.3-10)]. Since the components H_x and E_z are out of phase by $90°$, there is no power flow in the y direction. The power flows in the z and $-z$ directions in the DPS and SNG media, respectively, with intensities given by the magnitude of $\text{Re}\{\mathbf{S}\}$:

$$I_1(y) = \frac{\beta}{2\omega\epsilon_1}|H_0|^2 \exp(-2\gamma_1 y), \quad I_2(y) = \frac{\beta}{2\omega|\epsilon_2|}|H_0|^2 \exp(+2\gamma_2 y). \tag{8.1-18}$$

[†] A plasma is a collection of positive ions and free electrons whose overall net charge is approximately zero. A plasmon is a quasiparticle (a quantum) of the plasma oscillation, just as a photon is a quantum of the electromagnetic field. A polariton is a quasiparticle that results from the coupling of the electromagnetic field (photons) with the electric or magnetic excitation of a material. Surface plasmon polaritons result from the coupling of photons with surface plasmons.

The powers in the two media (areas under the $I_1(y)$ and $I_2(y)$ distributions) are:

$$P_1 = \frac{\beta}{4\omega\epsilon_1\gamma_1}|H_0|^2, \qquad P_2 = \frac{\beta}{4\omega|\epsilon_2|\gamma_2}|H_0|^2, \tag{8.1-19}$$

so that the net power flow $P_1 - P_2$ is proportional to $[(\epsilon_1\gamma_1)^{-1} - (|\epsilon_2|\gamma_2)^{-1}]$, which, in view of (8.1-15), is proportional to $(\epsilon_2^2 - \epsilon_1^2)$. Therefore, in the limit of $|\epsilon_2| \approx \epsilon_1$, the net power flow approaches zero.

EXAMPLE 8.1-1. ***SPP Wave.*** The boundary between two media with equal (positive) magnetic permeabilities $\mu_1 = \mu_2$, and with permittivities $\epsilon_1 = 1.41\,\epsilon_o$ and $\epsilon_2 = -47\,\epsilon_o$, supports a SPP wave at a free-space wavelength $\lambda_o = 1000$ nm with the following characteristics:

$$n_b = 1.206, \quad \lambda_o/n_b = 829.4 \text{ nm}, \quad d_1 = 381.1 \text{ nm}, \quad d_2 = 11.43 \text{ nm}.$$

Summary: SPP Waves at a DPS-SNG Boundary

If the condition $-\epsilon_2 > \epsilon_1$ is satisfied, the boundary between a DPS medium and a SNG medium with negative permittivity supports a TM optical surface wave along with an associated longitudinal surface-charge wave; the combination is a surface plasmon polariton (SPP). Because the penetration depths of the SPP into the two media are substantially smaller than an optical wavelength, the SPP is tightly confined to the boundary, resulting in a significant increase in local field intensity. The SPP has the distinct merit that it can be manipulated at the nanometer spatial scale while oscillating at an optical frequency. Using a similar analysis, it can be shown that the boundary between a DPS medium and a SNG medium with negative permeability supports a TE surface wave if the condition $-\mu_2 > \mu_1$ is satisfied. It can also be shown, as illustrated diagrammatically in Fig. 8.1-4, that the boundary between two SNG media may support a surface wave if ϵ_1 and ϵ_2 are of opposite signs, μ_1 and μ_2 are also of opposite signs, and a number of other conditions on the magnitudes of these parameters are satisfied. Surface waves are not supported at the boundary between two DPS media nor at the boundary between two DNG media.

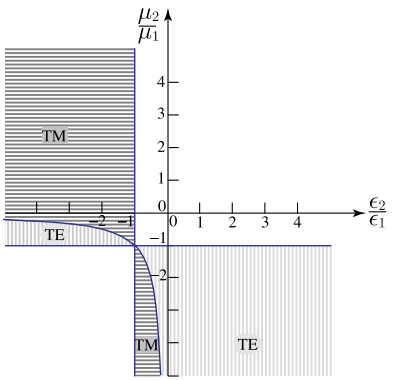


Figure 8.1-4 The boundary between two media with real permittivities and permeabilities can support TM or TE surface waves for values of the ratios ϵ_2/ϵ_1 and μ_2/μ_1 that fall in the shaded regions. Example 8.1-1 corresponds to $\mu_2/\mu_1 = 1$ and $\epsilon_2/\epsilon_1 < -1$.

SPP at boundary between DPS medium and lossy SNG medium. The result for a lossless SNG medium may be extended to a lossy medium by taking ϵ_2 to be complex ($\epsilon_2 = \epsilon_2' + j\epsilon_2''$), with negative real part ϵ_2', while maintaining $\mu_1 = \mu_2$ and ϵ_1 real and positive. Equations (8.1-16) and (8.1-17) for the parameters of the SPP wave in the lossless medium remain valid with slight modifications: the parameters β, ϵ_b, γ_1, and γ_2 become complex. Following the approach used in writing (5.5-5) and (8.1-5), we rewrite (8.1-16) as

$$n_\mathrm{b} - j\frac{\gamma_\mathrm{b}}{k_o} = \sqrt{\frac{\epsilon_\mathrm{b}}{\epsilon_o}}\,, \qquad \epsilon_\mathrm{b} = \frac{\epsilon_1\epsilon_2}{\epsilon_1+\epsilon_2}\,, \tag{8.1-20}$$

where γ_b is the amplitude attenuation coefficient of the traveling SPP wave. It is clear that the SPP-wave refractive index then becomes $n_\mathrm{b} = \mathrm{Re}\{\sqrt{\epsilon_\mathrm{b}/\epsilon_o}\}$. Equation (8.1-20) can be used to calculate the velocity of the SPP wave c_o/n_b, the plasmon wavelength λ_o/n_b, the intensity attenuation coefficient $\alpha_\mathrm{b} = 2\gamma_\mathrm{b}$, and the propagation length $d_\mathrm{b} = 1/\alpha_\mathrm{b} = 1/2\gamma_\mathrm{b}$. Equation (8.1-17) can be used to calculate the complex parameters γ_1 and γ_2, from which the penetration depths on both sides of the boundary, $d_1 = 1/2\,\mathrm{Re}\{\gamma_1\}$ and $d_2 = 1/2\,\mathrm{Re}\{\gamma_2\}$, may be determined. The dimensions of the plasmon wave are d_1 and d_2 in the transverse direction, and d_b along the boundary.

EXAMPLE 8.1-2. ***SPP at Gold-Si_3N_4 Interface.*** At a free-space wavelength $\lambda_o = 1000$ nm, the permittivities of Si_3N_4 and gold are $\epsilon_1 = 1.41\,\epsilon_o$ and $\epsilon_2 = (-47 + j3.4)\,\epsilon_o$, respectively. These values are identical to those set forth in Example 8.1-1 except that we now accommodate a small imaginary component of ϵ_2. An SPP wave propagating at the gold-Si_3N_4 interface thus has approximately the same values of n_b, d_1, and d_2 as those cited in Example 8.1-1. The propagation length, on the other hand, is determined by the imaginary component of ϵ_2 via

$$d_\mathrm{b} = 1/2\gamma_\mathrm{b} = \lambda_o/4\pi\,\mathrm{Im}\{\sqrt{\epsilon_\mathrm{b}/\epsilon_o}\}\,. \tag{8.1-21}$$

With the help of (8.1-20) we obtain

$$\epsilon_\mathrm{b}/\epsilon_o = 1.453 + j\,0.003234 \quad \text{so that} \quad \sqrt{\epsilon_\mathrm{b}/\epsilon_o} = 1.206 + j0.001341,$$

which leads to $d_\mathrm{b} = 59.3$ μm for $\lambda_o = 1$ μm.

Negative Refraction at a DPS-DNG Boundary

The refraction of light at the boundary between two ordinary dielectric (DPS) media obeys Snell's law, $n_1 \sin\theta_1 = n_2 \sin\theta_2$, which results from matching the components of the wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ along the direction of the boundary [Figs. 8.1-5(a) and 2.4-2]. If one of the media, say medium 2, is instead a DNG medium with negative refractive index n_2 then we have $n_1 \sin\theta_1 = -|n_2| \sin\theta_2$, which reveals that *the angle of refraction θ_2 must be negative and the refracted and incident rays both lie on same side of the normal to the boundary*. This outcome also can be understood as arising from matching the components of the wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ along the direction of the boundary [Fig. 8.1-5(b)].

It is thus clear that the optics of planar boundaries and lenses is altered significantly when a DPS medium is replaced by a DNG medium. Indeed, a convex lens of DNG material behaves like a concave lens of DPS material, and vice-versa. Also unexpected is the observation that a planar boundary between positive- and negative-index materials possesses focusing power, as illustrated in Fig. 8.1-6(a) for the special case $n_2 = -n_1$, which provides $\theta_2 = -\theta_1$. The planar boundary then acts on optical rays in the same way as does a convex spherical boundary between two DPS media, as depicted in Fig. 1.2-13: it forms an uninverted image with unity magnification. Moreover, for DPS and DNG media with permittivities and permeabilities that have the same magnitudes ($\epsilon_2 = -\epsilon_1$ and $\mu_2 = -\mu_1$), the impedances $\eta_1 = \sqrt{\mu_1/\epsilon_1}$ and

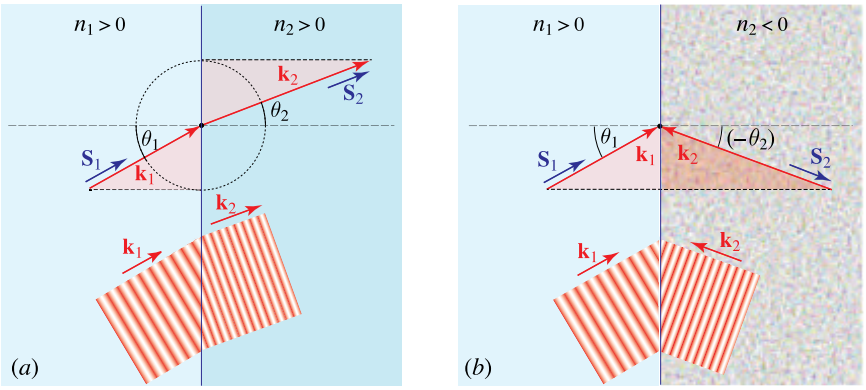


Figure 8.1-5 (a) Refraction at the boundary between two positive-index media. The directions of $\mathbf{S}_2$ and $\mathbf{k}_2$ are the same. (b) Refraction at the boundary between positive- and negative-index media. The directions of $\mathbf{S}_2$ and $\mathbf{k}_2$ are opposite. In both cases, the projections of the wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ along the boundary are equal in magnitude and parallel in direction.

$\eta_2 = \sqrt{\mu_2/\epsilon_2}$ are of equal magnitude and sign, so that no reflection occurs at the boundary, at any inclination, regardless of the polarization.

NIM Slab as a Near-Field Imaging System

A slab of DNG material (NIM) with parameters $\epsilon = -\epsilon_o$, $\mu = -\mu_o$, $n = -1$, and $\eta = \eta_o$, located in free space, acts as a lens. As illustrated in Fig. 8.1-6(b), each of the two DPS-DNG boundaries has focusing power so that one image is formed inside the slab and another beyond it. For a slab of width d_0, the imaging equation is simply $d_1 + d_2 = d_0$. Since the impedances match at the boundaries, no reflection occurs and the NIM slab has unity transmittance.

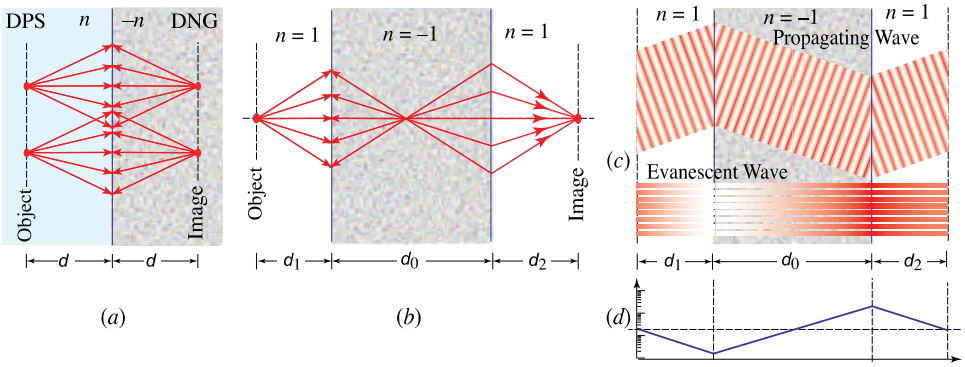


Figure 8.1-6 (a) Focusing of rays by a boundary between DPS and DNG media with refractive indices of equal magnitudes. The boundary forms an upright image inside the DNG medium. (b) Image formation using a slab with negative index $n = -1$ in free space. Each boundary has focusing power so that images are formed both inside and outside the slab. (c) Transmission of a propagating wave (spatial frequency smaller than the inverse wavelength), and an evanescent wave (spatial frequency greater than the inverse wavelength), through the slab. (d) The amplitude of an evanescent wave associated with a harmonic component of an object distribution with spatial frequency greater than the inverse wavelength is attenuated in free space but is amplified by the DNG medium so that its amplitude is restored in the image plane. A semilogarithmic sketch is provided.

As an imaging system, the NIM slab also has the remarkable property of forming an image of the object field distribution with subwavelength resolution, i.e., it offers imaging details finer than the wavelength (i.e., spatial frequencies greater than the inverse wavelength). As explained in Sec. 4.4D, imaging systems that make use of ordinary optics cannot transmit these high spatial frequencies since they correspond to attenuated evanescent waves. This remarkable property of the NIM slab means that, in principle, it is a "perfect lens" (also called a "superlens").[†]

The NIM slab offers resolution that exceeds the diffraction limit, as we now demonstrate with the help of the Fourier-optics approach described in Chapter 4. Referring to the illustration provided in Fig. 8.1-6(*c*), the transfer function for the propagation of a wave a distance d_1 through free space is

$$H_1(\nu_x, \nu_y) = \exp(-jk_z d_1), \quad k_z = \sqrt{k_o^2 - k_x^2 - k_y^2} = 2\pi\sqrt{\lambda^{-2} - \nu_x^2 - \nu_y^2}, \tag{8.1-22}$$

where $(\nu_x, \nu_y) = (k_x/2\pi, k_y/2\pi)$ are the spatial frequencies [see (4.1-9)]. Similarly, propagation a distance d_0 through a DNG medium is described by the transfer function $H_0(\nu_x, \nu_y) = \exp(-jk_z' d_0)$, where $k_z' = -k_z$. The third segment of the imaging system portrayed in Fig. 8.1-6(*c*) is propagation a distance d_2 through free space, which has the transfer function $H_2(\nu_x, \nu_y) = \exp(-jk_z d_2)$. The overall transfer function of this imaging system is thus the product $H = H_1 H_0 H_2$. When the imaging equation $d_1 + d_2 = d_0$ is satisfied, H becomes unity, which reveals that the system is an all-pass spatial filter and hence provides "perfect" imaging.

However, these three transfer functions have distinctly different behaviors for spatial frequencies that lie below and above λ^{-1} (i.e., for spatial frequencies smaller and larger than those for which $k_x^2 + k_y^2 = k_o^2$; see Fig. 4.1-11). Below λ^{-1}, the quantities under the square-root signs in (8.1-22) are positive so that all transfer functions are phase factors. In this domain, the phase of the function H_0 has a sign opposite to those of the free-space functions H_1 and H_2 so that the NIM phase factor compensates the phase shifts introduced by the two stretches of free space. The net result is a propagating wave [Fig. 8.1-6(*c*)].

For spatial frequencies larger than λ^{-1}, in contrast, the wavevector components[‡] k_z and $k_z' = -k_z$ become imaginary, $k_z = -j\sqrt{k_x^2 + k_y^2 - k_o^2} = -j2\pi\sqrt{\nu_x^2 + \nu_y^2 - \lambda^{-2}}$, whereupon

$$H_1 = \exp(-\gamma d_1), \; H_2 = \exp(-\gamma d_2), \; H_0 = \exp(+\gamma d_0), \; \gamma = 2\pi\sqrt{\nu_x^2 + \nu_y^2 - \lambda^{-2}}, \tag{8.1-23}$$

where γ is real. The factors H_1 and H_2 then represent attenuated evanescent waves, whereas H_0 represents an amplified evanescent wave. Consequently, the high spatial frequencies that are severely attenuated by propagation through free space, both before and after the slab, are amplified by an equal factor in the DNG medium, and are therefore fully restored [Fig. 8.1-6(*d*)]. Amplification of the evanescent wave in the NIM is not inconsistent with conservation of energy.

While the perfect restoration of an evanescent field that has diminished significantly as a result of exponential decay is possible in principle, slight energy dissipation in the NIM slab (evidenced by a small imaginary part of ϵ or μ) may thwart the restoration process. The distance from the object to the slab, the thickness of the slab, and the overall distance between the object and image planes, must be small in comparison with the wavelength, particularly if the spatial frequencies to be recovered are much

[†] See J. B. Pendry, Negative Refraction Makes a Perfect Lens, *Physical Review Letters*, vol. 85, pp. 3966–3969, 2000.

[‡] The choice of the minus sign for k_z, which is allowed by virtue of (4.1-3), ensures consistency with the development provided in Sec. 4.1B.

higher than the inverse wavelength. This requirement means that the NIM slab is a near-field imaging system.

Far-field imaging with subwavelength resolution. However, evanescent waves restored in the near field of the DNG slab may be converted into propagating waves, which can then be used to produce an image in the far field. This conversion may be implemented by making use of a periodic element of high spatial frequency, such as a nanoscale corrugation on the exit surface of the slab, as illustrated in Fig. 8.1-7. When an evanescent wave $\exp(-jk_x x)$ of high spatial frequency k_x is modulated by a harmonic function $\exp(jqx)$ of high spatial frequency q, the result is a propagating wave $\exp[-j(k_x - q)x]$ of lower spatial frequency $|k_x - q|$, provided that $|k_x - q| < k_o$. The far-field image formed by the downconverted spatial frequencies may, in principle, be processed to obtain a perfect replica of the original spatial distribution.

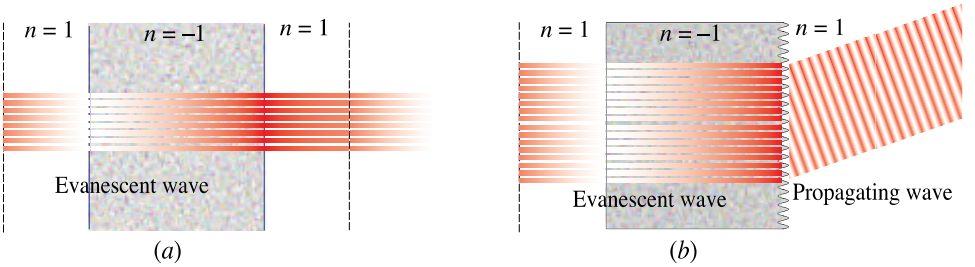


Figure 8.1-7 (*a*) An evanescent wave restored by a DNG slab attenuates in air beyond the near field. (*b*) An evanescent wave restored by a NIM slab can be converted into a propagating wave by making use of a periodic surface (grating) at the boundary.

*C. Hyperbolic Media

Anisotropic media may have permittivity and permeability tensors with positive or negative principal values. The designation of the medium as DPS, SNG, or DNG is then dependent on the direction of propagation of the wave as well as on its polarization. To demonstrate one facet of this rich behavior, we consider media endowed with isotropic magnetic properties and positive permeability μ, but with anisotropic dielectric properties.

Wave propagation in anisotropic media was described in Sec. 6.3 for media with positive principal values of the electric permittivity tensor $\boldsymbol{\epsilon}$, namely ϵ_1, ϵ_2 and ϵ_3. If these parameters take on mixed signs instead, wave propagation exhibits a number of unusual properties. For simplicity, we consider the special case of a uniaxial medium with $\epsilon_1 = \epsilon_2 > 0$, and compare wave propagation in media with positive and negative ϵ_3.

As shown in Sec. 6.3, the dispersion surfaces, also called the $\mathbf{k}$ surfaces or surfaces of constant $\omega(\mathbf{k})$, are, for the ordinary wave, a sphere of radius

$$k = n_o k_o \,, \tag{8.1-24}$$

where $n_o = \sqrt{\epsilon_1/\epsilon_o}$, and, for the extraordinary wave, a quadric surface

$$\frac{k_1^2 + k_2^2}{\epsilon_3} + \frac{k_3^3}{\epsilon_1} = \frac{k_o^2}{\epsilon_o}, \tag{8.1-25}$$

as provided in (6.3-21) and (6.3-22), respectively.

Considering the two cases at hand, we find:

- When ϵ_3 is positive, the extraordinary $\mathbf{k}$ surface is an ellipsoid of revolution, as displayed in Figs. 6.3-11(b) and 8.1-8(a). The ordinary and extraordinary waves have refractive indices n_o and $n(\theta)$, respectively, where $n(\theta)$, which is given by (6.3-15), varies between n_o and $n_e = \sqrt{\epsilon_3/\epsilon_o}$, as the angle θ varies between 0 and 90°.
- When ϵ_3 is negative, the extraordinary $\mathbf{k}$ surface is instead a hyperboloid of revolution (in two sheets), as portrayed in Fig. 8.1-8(b), and the material is known as a **hyperbolic medium**. The refractive index $n(\theta)$ for an extraordinary wave at an angle θ in the k_2–k_3 plane is then given by

$$\frac{1}{n^2(\theta)} = \frac{\cos^2\theta}{n_o^2} - \frac{\sin^2\theta}{n_e^2}, \tag{8.1-26}$$

where $n_e = \sqrt{|\epsilon_3|/\epsilon_o}$. As θ increases from 0 to $\theta_{\max} = \tan^{-1}(n_e/n_o)$, the refractive index $n(\theta)$ increases from n_o to ∞, as can be discerned from Fig. 8.1-8(b), signifying that the wavelength in the medium becomes progressively smaller and the wave slows to a halt at $\theta = \theta_{\max}$.

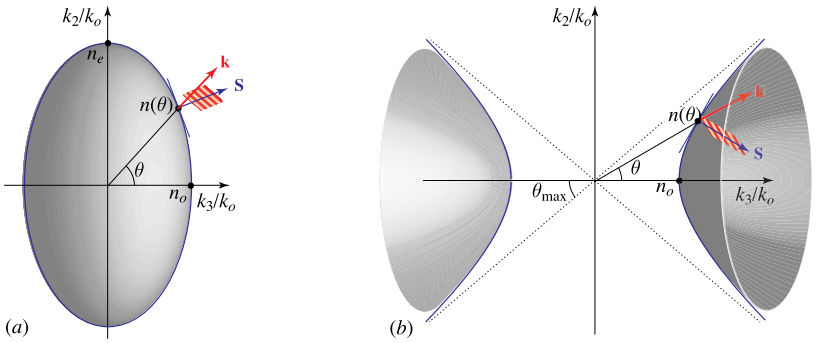


Figure 8.1-8 Contours of the $\mathbf{k}$ surfaces in the k_2–k_3 plane for a uniaxial anisotropic medium with principal values of the dielectric tensor $\epsilon_1 = \epsilon_2 > 0$ in two cases: (a) $\epsilon_3 > 0$, and (b) $\epsilon_3 < 0$. The contours are shown only for the extraordinary waves (for the ordinary waves, they are spheres in both cases). (a) The $\mathbf{k}$ contour is an ellipse and the medium is DPS for all directions of propagation. (b) The $\mathbf{k}$ contour is a hyperbola and the waves can propagate only in directions that lie within a cone of half-angle $\theta_{\max}$. Outside this cone, the medium acts like a SNG medium, which does not support propagating waves.

The Hyperbolic Slab as a Far-Field Imaging System with Subwavelength Resolution

An important property of a hyperbolic medium is that for a plane wave with wavevector components (k_1, k_2, k_3), no matter how large k_1 and k_2, there is a real value of k_3 that satisfies (8.1-25) when ϵ_3 is negative, indicating that the wave can propagate through the medium. This signifies that spatial frequencies greater than an inverse wavelength in any plane do not correspond to evanescent waves, as they do for ordinary media (see Sec. 4.1B), but rather can be transmitted over long distances. Although propagation is accompanied by Fresnel diffraction, the hyperbolic medium has a transfer function with no spatial frequency cutoff.

Moreover, Fresnel diffraction may be significantly reduced in a hyperbolic medium. If $n_0 \ll n_e$, then $\theta_{\max} = \tan^{-1}(n_e/n_o)$ approaches $\pi/2$, whereupon the hyperboloid of revolution in Fig. 8.1-8(b) flattens and becomes approximately planar, corresponding

to a constant $k_3 = n_o k_o$, for all k_1 and k_2. The transfer function $\exp(-jk_3 z)$ associated with propagation in the slab is then independent of the spatial frequencies (k_1, k_2) of the input-field distribution. A point in the input plane is thus imaged to a point in the output plane, and propagation may be described by ray optics. The slab then acts as perfect near-field imaging system (i.e., one with subwavelength resolution), as illustrated in Fig. 8.1-9(a). Furthermore, the slab may be curved in such a way that the image is geometrically magnified, as shown in Fig. 8.1-9(b). If the spatial-frequency components of the magnified image become smaller than the inverse wavelength, they generate propagating waves that may be captured with the help of a conventional lens, thereby forming a far-field image. A cylindrical slab such as this is known as a **hyperlens**.

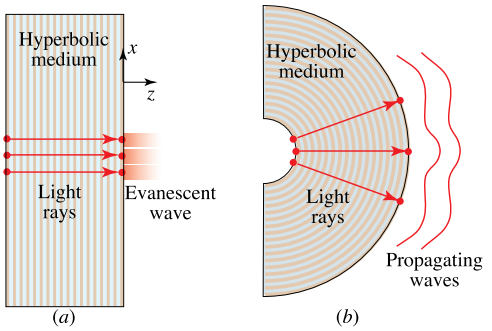


Figure 8.1-9 (a) A hyperbolic slab with $\epsilon_3 < 0$ and $0 < \epsilon_1 = \epsilon_2 \ll |\epsilon_3|$ has a planar dispersion relation ($k_3 =$ constant) so that propagation along the optic axis (z direction) is diffraction-free. (b) A hyperbolic slab curved to form an inhomogeneous cylindrical structure, with the local optic axis pointing in the radial direction, acts as a magnifier of subwavelength details. If the details of the magnified image are larger than the wavelength, it produces propagating waves in the outer medium.

Refraction at the Boundary of a Hyperbolic Medium

Refraction at a boundary between two media may be determined by drawing the $\mathbf{k}$ surfaces for the two media and matching the components of $\mathbf{k}$ along the boundary (see, e.g., the analysis of double refraction in Sec. 6.3E). The $\mathbf{k}$ surfaces for a boundary between an isotropic DPS medium and a hyperbolic medium are shown in Fig. 8.1-10(a), with only the extraordinary wave displayed for the hyperbolic medium.

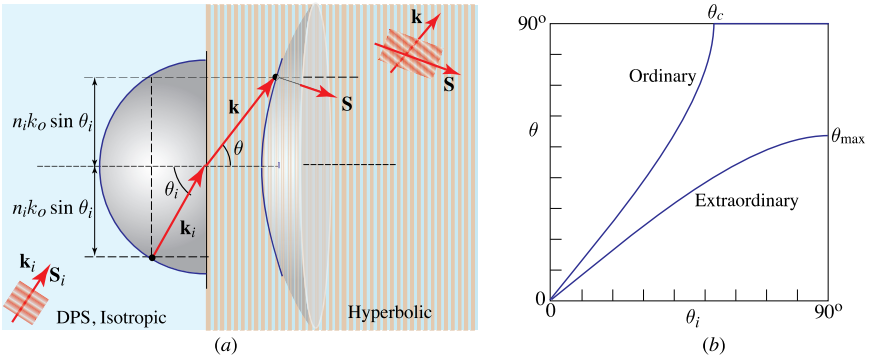


Figure 8.1-10 Refraction at a boundary between an isotropic DPS medium of refractive index n_i and a uniaxial hyperbolic medium with refractive indices n_o and n_e. (a) The $\mathbf{k}$ surface for the isotropic medium is a sphere, and that for the extraordinary wave of the hyperbolic medium is a hyperboloid of revolution. A form of negative refraction is evident since the Poynting vector $\mathbf{S}$ of the incident wave points upward whereas that of the refracted wave points downward. (b) Angle of refraction θ as a function of the angle of incidence θ_i for $n_o = 0.8\, n_i$ and $n_e = 1.25\, n_i$. For the ordinary wave, refraction occurs for angles of incidence θ_i smaller than the critical angle θ_c. For the extraordinary wave, refraction occurs for all angles θ_i, but the angle of refraction is limited to $\theta < \theta_{\max}$.

Because the $\mathbf{k}$ component along the boundary for the hyperbolic medium extends from 0 to ∞, it is always possible to find a match with that for any wave incident from the DPS medium, as may be confirmed by solving the matching equation, $n_i \sin\theta_i = n(\theta)\sin\theta$, together with (8.1-26). Total internal reflection therefore does not occur at this boundary. A plot of the angle of refraction θ as a function of the angle of incidence θ_i is displayed in Fig. 8.1-10(b). This type of **all-angle refraction** is consistent with the observation that the extraordinary wave in the hyperbolic medium cannot be evanescent.

8.2 METAL OPTICS: PLASMONICS

A. Optical Properties of Metals

Conductive Media

Conductive media such as metals, semiconductors, doped dielectrics, and ionized gases have free electric charges and an associated electric current density $\mathcal{J}$. In such materials, the first of the source-free Maxwell's equations, (5.1-7), must be modified by including the current density $\mathcal{J}$ along with the displacement current density $\partial\mathcal{D}/\partial t$, so that

$$\nabla \times \mathcal{H} = \frac{\partial \mathcal{D}}{\partial t} + \mathcal{J}. \tag{8.2-1}$$

The three other source-free Maxwell's equations, (5.1-8)–(5.1-10), remain unmodified, with $\mu = \mu_o$, since naturally occurring materials do not exhibit magnetic effects that vary at optical frequencies.

For a monochromatic wave of angular frequency ω, (8.2-1) takes the form

$$\nabla \times \mathbf{H} = j\omega\mathbf{D} + \mathbf{J}. \tag{8.2-2}$$

If the medium has linear dielectric properties, the electric flux density is given by $\mathbf{D} = \epsilon\mathbf{E} = \epsilon_o(1+\chi)\mathbf{E}$. Similarly, if the medium has linear conductive properties, and the **conductivity** is denoted σ, the electric current density is proportional to the electric field,

$$\mathbf{J} = \sigma\mathbf{E}, \tag{8.2-3}$$

which is a form of Ohm's law. Under these conditions, the right-hand side of (8.2-2) becomes $(j\omega\epsilon + \sigma)\mathbf{E} = j\omega(\epsilon + \sigma/j\omega)\mathbf{E}$, so that

$$\nabla \times \mathbf{H} = j\omega\epsilon_{\mathrm{c}}\mathbf{E}, \tag{8.2-4}$$

where the effective electric permittivity ϵ_{c} is given by

$$\boxed{\epsilon_{\mathrm{c}} = \epsilon\left(1 + \frac{\sigma}{j\omega\epsilon}\right).} \tag{8.2-5}$$

Effective Permittivity

The effective permittivity ϵ_{c} is a complex, frequency-dependent parameter that represents a combination of the dielectric and conductive properties of the medium. Since the second term in (8.2-5) varies inversely with frequency, the contribution of the conductive component diminishes with increasing frequency.

Since (8.2-4) takes the same form as the analogous equation for a dielectric medium, (5.3-12), the laws of wave propagation derived in Secs. 5.3–5.5 are applicable even in the presence of conductivity. Hence, the wavenumber in (5.5-2) and (5.5-3) takes the form $k = \beta - j\gamma = \omega\sqrt{\epsilon_c \mu_o}$, the impedance in (5.5-6) becomes $\eta = \sqrt{\mu_o/\epsilon_c}$, and the refractive index n and field attenuation coefficient $\gamma = \frac{1}{2}\alpha$ in (5.5-5) are determined from the complex equation

$$n - j\frac{\gamma}{k_o} = \sqrt{\frac{\epsilon_c}{\epsilon_o}} = \sqrt{\frac{\epsilon}{\epsilon_o}}\sqrt{1 + \frac{\sigma}{j\omega\epsilon}}, \tag{8.2-6}$$

where $k_o = \omega/c_o = \omega\sqrt{\epsilon_o \mu_o}$.

The ratio $\sigma/\omega\epsilon$ in (8.2-6) has two limiting regimes. When $\sigma/\omega\epsilon \ll 1$, dielectric effects dominate and conductive effects constitute only a minor correction to the wavenumber. When $\sigma/\omega\epsilon \gg 1$, in contrast, conductive effects dominate and $\epsilon_c \approx \sigma/j\omega$. We can use the Taylor-series expansion $\sqrt{1+x} \approx 1 + x/2$ for $x \ll 1$, along with the identity $\sqrt{j} = (1+j)/\sqrt{2}$, to obtain approximate expressions for the wave parameters when σ is real and frequency-independent. The results are provided in Table 8.2-1 for both limiting regimes:

Table 8.2-1 Refractive index n, attenuation coefficient α, and impedance η for a medium with real-valued, frequency-independent conductivity σ, in the limits of small and large values of $\sigma/\omega\epsilon$. These equations are not applicable when the conductivity σ is complex and/or frequency-dependent, as discussed in the next section in connection with the Drude model.

$\sigma/\omega\epsilon \ll 1$	$\sigma/\omega\epsilon \gg 1$	
$n \approx \sqrt{\epsilon/\epsilon_o}$	$n \approx \sqrt{\sigma/2\omega\epsilon_o}$	(8.2-7)
$\alpha \approx \sigma\sqrt{\mu_o/\epsilon}$	$\alpha \approx \sqrt{2\omega\mu_o\sigma}$	(8.2-8)
$\eta \approx \sqrt{\mu_o/\epsilon}$	$\eta \approx (1+j)\sqrt{\omega\mu_o/2\sigma}$	(8.2-9)

It is apparent from Table 8.2-1 that at a fixed frequency ω, the attenuation coefficient α is proportional to the conductivity σ for small values of σ, but becomes proportional to $\sqrt{\sigma}$ for large values of σ. The impedance η, initially real and independent of σ, eventually acquires a 45° phase shift and becomes inversely proportional to $\sqrt{\sigma}$ as σ becomes large. In this limit, η itself becomes small so that the material becomes highly reflective at the boundary with a non-conductive medium.

For a fixed value of σ at low frequencies, the refractive index n is proportional to $1/\sqrt{\omega}$, whereas the attenuation coefficient α and the impedance η are directly proportional to $\sqrt{\omega}$. As ω increases to a value such that the ratio $\sigma/\omega\epsilon$ becomes very small, all three of the wave propagation parameters become frequency independent and the material behaves as a nondispersive dielectric medium with loss.

For a perfect conductor, $\sigma \to \infty$ so that $\alpha \to \infty$ and the penetration depth $d_p = 1/\alpha \to 0$. Also, $\eta \to 0$ so that at the boundary with a dielectric medium the power reflectance $\mathcal{R} \to 1$; the material then behaves as a **perfect mirror**.

As discussed in the next section, the conductivities of real metals at optical frequencies are often complex-valued and frequency-dependent, in which case the results provided in Table 8.2-1 are not applicable. Under those conditions, the frequency dependencies of the wave parameters can differ considerably from those presented in the table.

Metals: The Drude Model

When the relation between $\mathcal{J}$ and $\mathcal{E}$ is dynamic rather than static, the conductivity σ must be frequency-dependent with a finite bandwidth and a finite response time. Treating the conduction electrons as an ideal gas of independent particles that move freely between scattering events, the **Drude model** (also called the **Drude–Lorentz** model) prescribes a frequency-dependent complex conductivity of the form

$$\sigma = \frac{\sigma_0}{1 + j\omega\tau}, \tag{8.2-10}$$

where σ_0 is the low-frequency conductivity and τ is a scattering time (or collision time). For sufficiently low frequencies such that $\omega \ll 1/\tau$, $\sigma \approx \sigma_0$ is real and frequency-independent, in which case the results provided in Table 8.2-1 apply.

If the medium has free-space-like dielectric properties with no other losses ($\epsilon = \epsilon_o$), inserting (8.2-10) into (8.2-5) leads to a relative effective permittivity given by

$$\frac{\epsilon_c}{\epsilon_o} = 1 + \frac{\omega_p^2}{-\omega^2 + j\omega/\tau} = 1 + \frac{\omega_p^2}{-\omega^2 + j\omega\zeta}, \tag{8.2-11}$$

where $\zeta = 1/\tau$ is the scattering rate (collision frequency). The **plasma frequency** ω_p (strictly speaking, the plasma angular frequency) is defined as

$$\omega_p = \sqrt{\frac{\sigma_0}{\epsilon_o \tau}} \tag{8.2-12}$$

and the free-space **plasma wavelength** λ_p is

$$\lambda_p = \frac{2\pi c_o}{\omega_p}. \tag{8.2-13}$$

Some lossless media exhibit dielectric properties that differ from those of free space, as manifested by a residual frequency-independent relative permittivity $1 + \chi_m$ that persists to frequencies $\omega \gg \omega_p$. In this case, the relative effective permittivity in (8.2-11) becomes

$$\frac{\epsilon_c}{\epsilon_o} = 1 + \chi_m + \frac{\omega_p^2}{-\omega^2 + j\omega\zeta}. \tag{8.2-14}$$

Observed values of ω_p, λ_p, τ, and ζ for a number of metals are provided in Table 8.2-2.

Table 8.2-2 Plasma frequency ω_p, free-space plasma wavelength λ_p, scattering time τ, and scattering rate $\zeta = 1/\tau$, for Al, Ag, Au, and Cu.[a]

	ω_p (rad/s)	λ_p (nm)	τ (fs)	ζ (s^{-1})
Al	1.83×10^{16}	103	5.10	1.96×10^{14}
Ag	1.37×10^{16}	138	31.3	0.32×10^{14}
Au	1.35×10^{16}	139	9.25	1.08×10^{14}
Cu	1.33×10^{16}	142	6.90	1.45×10^{14}

[a] See, e.g., E. J. Zeman and G. C. Schatz, An Accurate Electromagnetic Theory Study of Surface Enhancement Factors for Ag, Au, Cu, Li, Na, Al, Ga, In, Zn, and Cd, *Journal of Physical Chemistry*, vol. 91, pp. 634–643, 1987.

These results are related to those for the resonant medium described by the Lorentz oscillator model, which treats the motion of an electron bound to a nucleus as a harmonic oscillator (see Sec. 5.5C). However, here the electrons are free and there is no restoring force so that the elastic constant $\kappa = 0$ and the resonance frequency $\omega_0 = \sqrt{\kappa/m} = 0$. The Lorentz-model equation of motion (5.5-16) then becomes $d^2x/dt^2 + \zeta dx/dt = -e\mathcal{E}/m$, and the corresponding polarization density $\mathcal{P} = -Nex$ obeys the equation $d^2\mathcal{P}/dt^2 + \zeta d\mathcal{P}/dt = (Ne^2/m)\mathcal{E}$, where N is the electron density of the medium. For a field oscillating at an angular frequency ω, this gives rise to $-\omega^2 P + j\omega\zeta P = (Ne^2/m)E$, which leads to a susceptibility

$$\chi(\omega) = \frac{P}{\epsilon_o E} = \frac{\omega_p^2}{-\omega^2 + j\omega\zeta}, \tag{8.2-15}$$

where

$$\boxed{\omega_p = \sqrt{\frac{Ne^2}{\epsilon_o m}}.} \tag{8.2-16}$$

Plasma Frequency
Drude Model

Equation (8.2-15), together with the relation $\epsilon_c = \epsilon_o(1 + \chi)$ for a medium with free-space dielectric properties, leads to (8.2-11) and (8.2-12), provided that

$$\sigma_0 = \frac{Ne^2\tau}{m}. \tag{8.2-17}$$

Equation (8.2-15) is a special case of (5.5-19) for the Lorentz resonant medium, with $\omega_0 = 0$ and $\zeta = \Delta\omega = 2\pi\Delta\nu$, where $\Delta\nu$ is the spectral width. A comparison of the frequency dependence of the real and imaginary parts of the relative permittivity $\epsilon_r = \epsilon_c/\epsilon_o$ for a Lorentz dielectric resonant medium, based on (5.5-19), or (5.5-20) and (5.5-21), and for a Drude metal based on (8.2-11), is provided in Fig. 8.2-1.

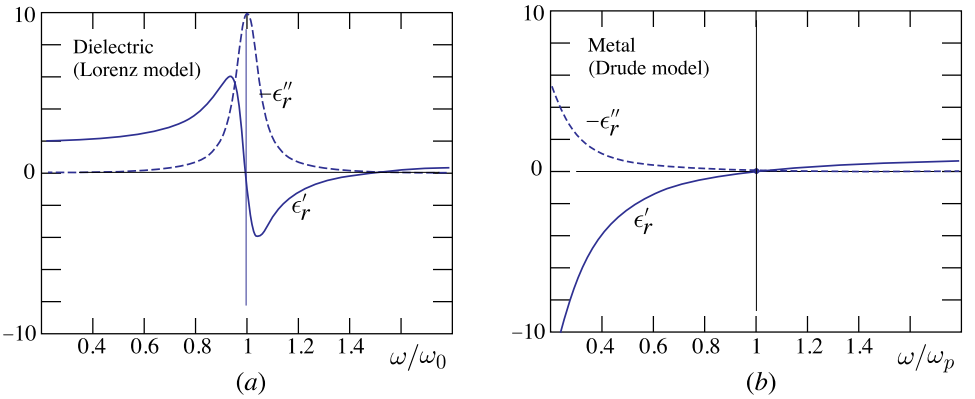


Figure 8.2-1 Frequency dependence of the real part (solid curve) and imaginary part (dashed curve) of the relative effective permittivity $\epsilon_r = \epsilon_c/\epsilon_o = \epsilon_r' + j\epsilon_r''$. (*a*) Dielectric resonant medium described by the Lorenz model, (5.5-20) and (5.5-21), with resonance frequency ω_0, $Q = \omega_0/\Delta\omega = 10$, and $\chi_o = 1$. (*b*) Metal described by the Drude model (8.2-11) with plasma frequency ω_p, $\omega_p\tau = 10$, and $\epsilon = \epsilon_o$. The real part of the permittivity is negative for $\omega < \omega_p\sqrt{1 - (\omega_p\tau)^{-2}} \approx \omega_p$.

EXAMPLE 8.2-1. ***Permittivity and Reflectance of Silver.*** As an example of the range of validity of the Drude model, Fig. 8.2-2(a) displays the measured relative effective permittivity $\epsilon_r = \epsilon_c/\epsilon_o$ for silver (dashed curves) along with the best-fitting functions (solid curves) consistent with the Drude model [see, e.g., Fig. 8.2-1(b)]. The experimental permittivities and reflectances for various metals, as functions of wavelength, are available from a number of sources.† It is apparent from Fig. 8.2-2(a) that the Drude-model relative permittivity provided in (8.2-14) fits the experimental data quite well over the 450–600-nm wavelength range, but is clearly inadequate at shorter wavelengths. The best-fitting model parameters over the wavelength range displayed (200–600 nm) are $\chi_m = 4.45$, $\omega_p = 1.47 \times 10^{16}$ rad/s ($\lambda_p = 128$ nm), and $\tau = 12$ fs ($\zeta = 0.84 \times 10^{14}$ s^{-1}). These values differ from those presented in Table 8.2-2 because of the constraints inherent in this analysis.

In Fig. 8.2-2(b), the measured power reflectance for light normally incident on the Ag-air boundary (dashed curve) is compared with the reflectance calculated on the basis of the Drude model (solid curve) using the relation $\mathcal{R} = |(\eta - \eta_o)/(\eta + \eta_o)|^2$, where $\eta = \sqrt{\epsilon_c/\mu_o}$ and $\eta_o = \sqrt{\epsilon_o/\mu_o}$ are the impedances of Ag and air, respectively [see (6.2-8)]. The fit is quite good over the 400–600-nm wavelength range, where Ag exhibits near-perfect reflectance (as it does at longer wavelengths). Other metals exhibit similar behavior. Gold and copper also have free-space plasma wavelengths that lie well into the ultraviolet (see Table 8.2-2) and those metals do not become good reflectors until the wavelength exceeds ~ 550 nm (thereby explaining their reddish color). The origin of this behavior is interband absorption, i.e., absorption by bound electrons, which is not accommodated in the Drude free-electron model (see Prob. 8.2-3). Aluminum, in contrast, hews quite closely to the predictions of the Drude model, exhibiting near-unity reflectance over a wavelength range that stretches from 200 nm to beyond 10 μm.

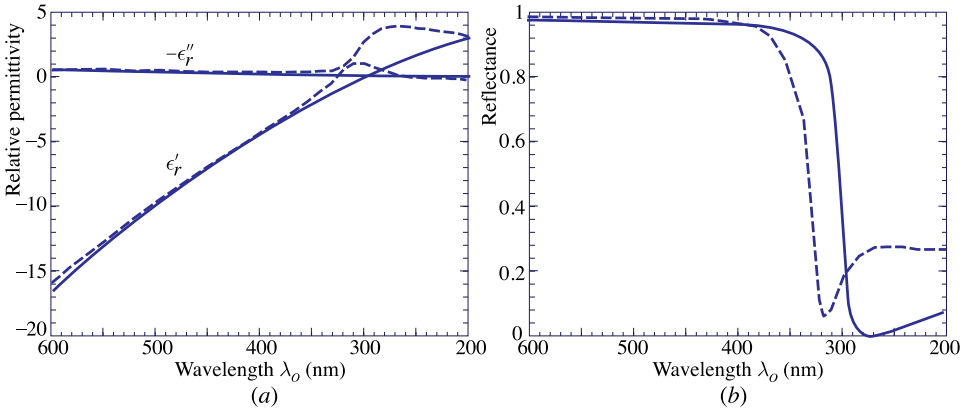


Figure 8.2-2 (a) Real and imaginary parts of the relative effective permittivity $\epsilon_r = \epsilon_r' + j\epsilon_r''$ as a function of wavelength for Ag. The dashed curves represent experimental values whereas the solid curves are fits to the Drude model. (b) Power reflectance $\mathcal{R}$ vs. wavelength at the Ag-air boundary. The dashed curve is the experimental reflectance while the solid curve is the calculated reflectance based on the Drude model. Note that the wavelength decreases from left to right (corresponding to increasing frequency); the free-space plasma wavelength λ_p lies beyond the right edges of the graphs.

Simplified Drude model. We now consider the behavior of the Drude model for angular frequencies sufficiently high such that $\omega \gg 1/\tau$ ($\omega \gg \zeta$), which is equivalent to neglecting damping ($\zeta \to 0$) in (5.5-16). Under these conditions, (8.2-10) becomes $\sigma \approx \sigma_0/j\omega\tau$, which is imaginary, and (8.2-14), with $\chi_m = 0$, reduces to the effective permittivity for the **simplified Drude model**, which is real:

† The experimental data presented in Fig. 8.2-2 were drawn from P. B. Johnson and R. W. Christy, Optical Constants of the Noble Metals, *Physical Review B*, vol. 6, pp. 4370–4379, 1972.

$$\epsilon_c \approx \epsilon_o \left(1 - \frac{\omega_p^2}{\omega^2}\right). \tag{8.2-18}$$

Effective Permittivity
Simplified Drude Model

It is clear from (8.2-18) that the presence of conductivity in the medium serves to suppress the permittivity to a value below ϵ_o and to impart to it a functional form that is inversely proportional to the square of the frequency.

The simplified Drude model is useful for describing the optical behavior of metals in the near-infrared and visible regions of the spectrum [see Fig. 2.0-1]. For wavelengths shorter than $\sim 1\mu$m, which corresponds to an angular frequency $\omega = 2\pi\nu = 1.9\times10^{15}$ rad/s, $\omega \gg \zeta$ for the metals listed in Table 8.2-2.

In summary, for frequencies that are sufficiently high such that $\omega \gg 1/\tau$, the simplified Drude model for a metal is a special case of the Lorentz model for a dielectric medium in which there is neither a restoring force ($\kappa = 0$) nor damping ($\zeta = 0$). The permittivity (8.2-18) is real and negative for frequencies below the plasma frequency ω_p, where the metal behaves as a SNG medium, and real and positive for frequencies above ω_p, where the metal behaves as a DPS medium.

Wave propagation in a medium described by the simplified Drude model is characterized by a propagation constant $\beta = nk_o = \sqrt{\epsilon_c/\epsilon_o}\,(\omega/c_o)$ and a dispersion relation

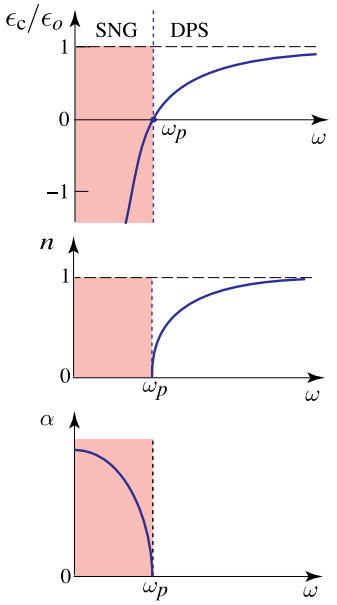
$$\beta = \frac{\omega}{c_o}\sqrt{1 - \frac{\omega_p^2}{\omega^2}}\,, \tag{8.2-19}$$

Dispersion Relation
Simplified Drude Model

as illustrated in Fig. 8.2-3. The corresponding relative permittivity ϵ_c/ϵ_o, refractive index n, and attenuation coefficient α are also displayed in the figure.

It is evident from Fig. 8.2-3 that wave propagation in a metal described by the simplified Drude model exhibits distinctly different behavior below, at, and above the plasma frequency ω_p:

- At frequencies below the plasma frequency ($\omega < \omega_p$), the effective permittivity ϵ_c is negative. The metal then behaves as a SNG medium with imaginary wavenumber, $k = j\omega\sqrt{|\epsilon_c|\mu_o}$, which corresponds to attenuation without propagation. This spectral region may therefore be regarded as a **forbidden band**. The attenuation coefficient $\alpha = 2k_o\sqrt{\omega_p^2/\omega^2 - 1}$ decreases monotonically with increasing frequency and vanishes at the plasma frequency. The free electrons then undergo longitudinal collective oscillations that take the form of **plasmons**, the quanta of the plasma wave (much as photons are the quanta of the electromagnetic wave). The negative permittivity corresponds to an imaginary impedance, which indicates that total reflection occurs at the boundary with a DPS medium (see Sec. 6.2), provided that $\lambda_o > \lambda_p$. Doped semiconductors, in contrast, are not reflective in the visible region. This is because the plasma frequency of such materials lies in the infrared since the free-electron density N is far smaller than that in metals [see (8.2-16)]. Although the Drude model provides a good starting point for determining the optical properties of metals and doped semiconductors, it is by no means the final word, as was illustrated in Example 8.2-1.



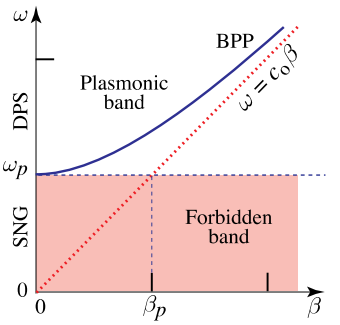


Figure 8.2-3 Left: Frequency dependence of the relative effective permittivity ϵ_c/ϵ_o, the refractive index n, and the attenuation coefficient α for a medium described by the simplified Drude model. Right: The dispersion relation $\omega = \sqrt{\omega_p^2 + c_o^2\beta^2}$. Metals support propagating optical waves, known as bulk plasmon polaritons (BPPs), but only when the frequency of the wave falls in the plasmonic band (which comprises frequencies greater than the plasma frequency ω_p).

- At frequencies above the plasma frequency ($\omega > \omega_p$), the effective permittivity is positive and real so that the medium behaves like a lossless dielectric material, albeit with unique dispersion characteristics. The propagation constant becomes $\beta = \sqrt{\omega^2 - \omega_p^2}\,/c_o$, while the refractive index $n = \sqrt{1 - \omega_p^2/\omega^2}$ lies below unity and is very small near the plasma frequency. This spectral band is referred to as the **plasmonic band** and waves that travel through the metal in this band are called **bulk plasmon polaritons** (**BPPs**).

B. Metal–Dielectric Boundary: Surface Plasmon Polaritons

Since a metal behaves as a SNG medium at optical frequencies below the plasma frequency ($\omega < \omega_p$), it can support a surface plasmon polariton (SPP) wave at the boundary with an ordinary dielectric (DPS) medium. As indicated in Sec. 8.1B, a SPP is a guided optical surface wave accompanied by a longitudinal electron-density wave that oscillates at the optical frequency. Because the SPP hugs the metal–dielectric boundary over distances far smaller than an optical wavelength, it can be controlled and manipulated at nanometer spatial scales without sacrificing its optical temporal frequency. The tight binding of the wave to the boundary carries a concomitant enhancement of the local field intensity.

In this section we apply the analysis of SPP waves at a DPS-SNG boundary, developed in Sec. 8.1B, to a metal–dielectric boundary. Specifically, we combine the effective permittivity of a metal described by the simplified Drude model, set forth in (8.2-18), with the general wave-parameter results summarized in (8.1-16) and (8.1-17). Based on the analysis provided in Sec. 8.1B, a SPP wave can exist at the boundary provided that the magnitude of the permittivity of the SNG medium $|\epsilon_2|$ is greater than that of the DPS medium ϵ_1. If the metal is described by the simplified Drude model, (8.2-18) reveals that its effective permittivity is $\epsilon_2 = \epsilon_o(1 - \omega_p^2/\omega^2)$, from which it is clear that the material behaves as a SNG medium if $\omega < \omega_p$. As illustrated in Fig. 8.2-

4(*a*), the condition $|\epsilon_2| > \epsilon_1$ is equivalent to the condition $\omega < \omega_s$, where

$$\omega_s = \frac{\omega_p}{\sqrt{1+\epsilon_{r1}}}\,, \tag{8.2-20}$$

with $\epsilon_{r1} = \epsilon_1/\epsilon_o$ representing the relative permittivity of the dielectric medium. The frequency band over which a SPP wave can be supported, $\omega < \omega_s$, is smaller than that over which the simple Drude metal behaves as a SNG medium, $0 < \omega < \omega_p$.

After working through the algebra involved in combining (8.1-16) and (8.2-18), it turns out that β, n_b, and ϵ_b obey (8.2-21). Equation (8.2-22) is carried over intact from (8.1-17). Again, $k_o = \omega\sqrt{\epsilon_o\,\mu} = 2\pi/\lambda_o$ represents the free-space wavenumber, and the penetration depth is given by $d_p = 1/2\gamma$.

$$\beta = n_\mathrm{b} k_o\,, \qquad n_\mathrm{b} = \sqrt{\frac{\epsilon_\mathrm{b}}{\epsilon_o}}\,, \qquad \epsilon_\mathrm{b} = \epsilon_1\,\frac{1-\omega^2/\omega_p^2}{1-\omega^2/\omega_s^2}\,, \tag{8.2-21}$$

$$\gamma_1 = \sqrt{\frac{-\epsilon_1^2}{\epsilon_o(\epsilon_1+\epsilon_2)}}\;k_o\,, \qquad \gamma_2 = \sqrt{\frac{-\epsilon_2^2}{\epsilon_o(\epsilon_1+\epsilon_2)}}\;k_o\,. \tag{8.2-22}$$

SPP Wave
Metal–Dielectric

As illustrated in Fig. 8.2-4, the behavior of the parameters provided in (8.2-21) differs for the three salient frequency bands:

- For $\omega < \omega_s$, the simple Drude metal behaves as a SNG medium, and SPP waves may be guided along its boundary with a dielectric medium. The frequency band $\omega < \omega_s$ lies within the forbidden band of the *bulk* metal, $0 < \omega < \omega_p$, in which bulk propagating waves are not permitted. The properties of the SPP wave are dependent on the ratio $|\epsilon_2|/\epsilon_1$, which is a monotonically decreasing function of the ratio ω/ω_s, and approaches the critical value of unity when $\omega = \omega_s$. Forging a comparison with waves in the bulk dielectric medium, the SPP velocity c_o/n_b is smaller and the plasmon wavelength λ_o/n_b is shorter. This is also evident from the dispersion curve displayed in Fig. 8.2-4(*c*) since $\beta > \omega/c_1$ for the SPP wave. In accordance with (8.1-17), the penetration depth in the metal, $d_2 = 1/2\gamma_2$, is smaller than that in the dielectric, $d_1 = 1/2\gamma_1$, and both are smaller than the wavelength in the bulk dielectric medium. As ω/ω_s increases, the SPP slows and the associated plasmon wavelength decreases. As explained in Sec. 8.1B, the SPP also becomes more localized, exhibiting smaller penetration depths in both the metal and the dielectric. At $\omega/\omega_s = 1$, the permittivities of the metal and dielectric become equal in magnitude and opposite in sign ($\epsilon_1 + \epsilon_2 = 0$), whereupon the velocity of the SPP becomes zero.
- For $\omega_s < \omega < \omega_p$, the simple Drude metal is a SNG medium, but SPP waves are not permitted since $|\epsilon_2| < \epsilon_1$. Hence, bulk waves propagating in the dielectric medium are totally reflected from the DPS-SNG boundary at all angles of incidence.
- For $\omega > \omega_p$, the simple Drude metal acts as a DPS medium. DPS-DPS boundaries cannot support SPP waves. However, much like a dielectric medium, the bulk metal supports propagating waves, namely bulk plasmon polaritons (BPPs). At these high frequencies, the reflection and refraction of waves at the metal–dielectric boundary comply with the conventional laws obeyed by two dielectric media. Total internal reflection occurs, but only for angles of incidence greater than the critical angle.

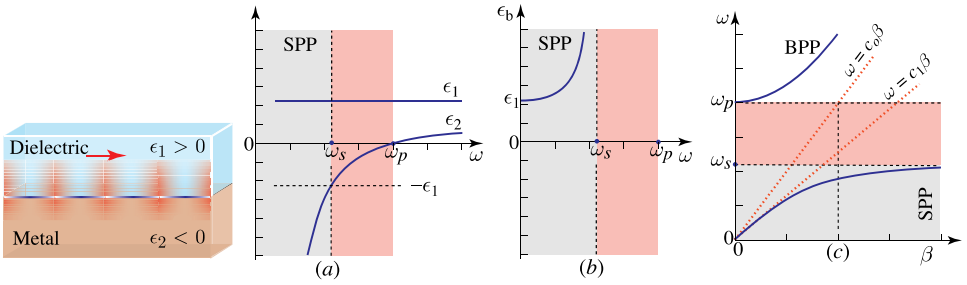


Figure 8.2-4 Surface plasmon polariton (SPP) wave at a metal–dielectric boundary, as depicted in Fig. 8.1-3. (*a*) Frequency dependence of the permittivities of the dielectric and metallic media, ϵ_1 and ϵ_2, respectively. The condition $|\epsilon_2| > \epsilon_1$, required for the existence of the SPP wave, is satisfied for $\omega < \omega_s$. (*b*) Frequency dependence of the effective permittivity ϵ_b of the SPP wave. The wave velocity is c_o/n_b, where $n_b = \sqrt{\epsilon_b/\epsilon_o}$. (*c*) Dispersion relations for the BPP (bulk plasmon polariton) and SPP waves. These plots were computed from (8.2-19) and (8.2-21), respectively, using $\epsilon_{r1} = 2.25$ (the relative permittivity of glass), so that $\omega_s = \omega_p/\sqrt{1+\epsilon_{r1}} = 0.55\,\omega_p$. Light lines in free space and in the bulk dielectric medium are shown as dotted red.

EXAMPLE 8.2-2. ***SPP at Ag-Air and Ag-SiO$_2$ Boundaries.*** SPP waves are supported at the boundaries of dielectrics with metals such as silver (Ag) and gold (Au). For an optical wave of free-space wavelength $\lambda_o = 800$ nm, the permittivity of Ag is $\epsilon_2 \approx -32.6\,\epsilon_o$.[†] The plasma frequency of Ag, $\omega_p = 1.37 \times 10^{16}$ s^{-1}, corresponds to a free-space plasma wavelength $\lambda_p = 2\pi c_o/\omega_p = 138$ nm (see Table 8.2-2). At a Ag-air boundary ($\epsilon_1 = \epsilon_o$), (8.2-20) provides that the free-space wavelength corresponding to ω_s is $\lambda_s = \sqrt{2}\,\lambda_p = 195$ nm. SPP waves must therefore have a free-space wavelength longer than 195 nm to be viable. In accordance with (8.2-21) and (8.2-22), a SPP wave of frequency corresponding to a free-space wavelength $\lambda_o = 800$ nm at this boundary has the following properties:

$$n_b = 1.016, \quad \lambda_o/n_b = 788 \text{ nm}, \quad d_1 = 358 \text{ nm}, \quad d_2 = 11 \text{ nm}.$$

At a Ag-SiO$_2$ boundary ($\epsilon_1 = 3.9\,\epsilon_o$ at $\lambda_o = 800$ nm), the free-space plasma wavelengths corresponding to ω_p and ω_s are $\lambda_p = 138$ nm and $\lambda_s = 305$ nm, respectively. At this boundary, a SPP wave of frequency corresponding to a free-space wavelength $\lambda_o = 800$ nm has the following properties:

$$n_b = 2.104, \quad \lambda_o/n_b = 380 \text{ nm}, \quad d_1 = 87 \text{ nm}, \quad d_2 = 10 \text{ nm}.$$

Comparing the results for both dielectric media, it is apparent that the higher-index material, SiO$_2$, results in a SPP wave with lower velocity, shorter wavelength, and shallower penetration into both the Ag and the dielectric medium. Moreover, $d_2 < d_1$ for both air and SiO$_2$, confirming the deeper penetration into the dielectric side of the boundary, as depicted in Figs. 8.1-3 and 8.2-4.

Generation and Detection of Surface Plasmon Polaritons

Since the propagation constant of a SPP wave traveling along a metal–dielectric boundary is greater than that of an ordinary optical wave of the same frequency propagating in the dielectric medium [see Fig. 8.2-4(*c*)], it is difficult to couple the two waves. One way in which this can be achieved, however, is to couple an evanescent wave, generated by total reflection at a metal–dielectric boundary, to the SPP wave at the

[†] The experimental value for the permittivity of Ag at a free-space wavelength of $\lambda_o = 800$ nm is $\epsilon_2 = (-32.6 + j\,0.5)\,\epsilon_o$. The calculated value of the effective permittivity at this wavelength, based on (8.2-18) for the simplified Drude model (no damping), provides $\epsilon_2' = -32.8\,\epsilon_o$, in close agreement with the experimental value.

opposite metal–dielectric boundary, as portrayed in the prism-coupler configuration shown in Fig. 8.2-5(a). A similar approach is used to couple light into a waveguide via a prism, as will become clear in Fig. 9.4-4.

The coupling takes place only when the two waves are phase matched, i.e., when their propagation constants are precisely equal. From the interior of a prism of refractive index $n_{\rm p}$, this condition is satisfied at an angle of incidence $\theta_{\rm p} = \theta_r$, where $n_{\rm p} k_o \sin\theta_r = \beta$. The parameter β is the propagation constant of the SPP wave, given in (8.2-21) and (8.2-20), which is dependent on the refractive index $n_1 = \sqrt{\epsilon_1/\epsilon_o}$ of the dielectric medium adjacent to the remote side of the metallic film [which is represented in Fig. 8.2-5(a) as air]. It is straightforward to verify that this condition is met at an angle of incidence θ_r given by

$$\sin\theta_r = \frac{n_1}{n_{\rm p}} \sqrt{\frac{1-\omega^2/\omega_p^2}{1-(1+n_1^2)\,\omega^2/\omega_p^2}}\,. \tag{8.2-23}$$

When the phase-matching condition is satisfied, optical power is transferred to launch the SPP wave via a form of **frustrated total internal reflection** (FTIR), so that the power of the reflected optical wave at the boundary of the prism diminishes significantly. The change in reflectance is manifested as a sharp, resonance-like function of the incidence angle, as portrayed in Fig. 8.2-5(b). Since the angle θ_r depends

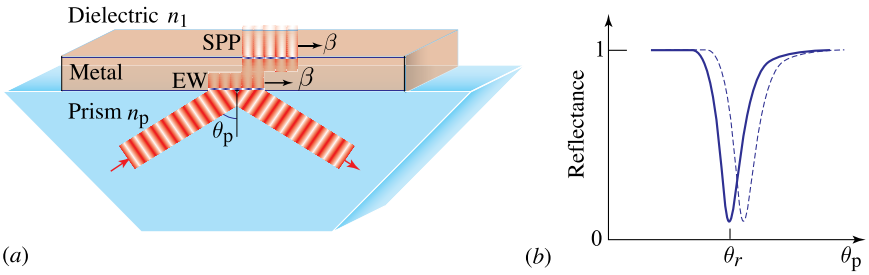


Figure 8.2-5 (a) Generation of a SPP wave by use of a prism coupler. An evanescent wave (EW) generated by total internal reflection of an optical wave at the prism–metal boundary excites a SPP wave at the opposite metal–dielectric boundary. (b) Dependence of the reflectance of the optical wave on the angle of incidence $\theta_{\rm p}$. When the phase-matching condition is satisfied, at $\theta_{\rm p} = \theta_r$, a SPP wave is launched, which results in a decrease of the reflected-wave intensity. A slight change in the refractive index n_1 alters θ_r in accordance with (8.2-23), which gives rise to the dashed curve. The system thus serves as a precise sensor.

on n_1, it is sensitive to changes in the physical environment surrounding the metal film, which determine n_1. This measurement technique, known as **surface plasmon resonance** (**SPR**) **spectroscopy**, has found extensive use for chemical and biological sensing applications; examples are gas detection and the measurement of molecular adsorption. Another method of exciting an SPP wave is to scatter light from a periodic subwavelength structure (grating) deposited on the metallic surface, which contributes a spatial-frequency component that compensates for the propagation-constant mismatch.

The detection of an SPP wave may be achieved by converting it into a proportional optical wave, using a prism coupler or grating operated in a reverse conversion process.

C. The Metallic Nanosphere: Localized Surface Plasmons

A metallic structure of subwavelength dimensions supports plasmonic oscillations at its (external or internal) boundary with a dielectric medium. Examples of such structures are metallic nanospheres, nanodisks, and other nanoparticles. These oscillations are known as **localized surface plasmon polaritons**, or simply **localized surface plasmons** (**LSPs**). When the excitation frequency matches the resonance frequency of the structure, the result is a **surface plasmon resonance** (**SPR**). A LSP is to be distinguished from a long-range SPP, which is a SPP wave that propagates along an extended metal–dielectric boundary, as described in Sec. 8.2B. Similarly, the surface plasmon resonance frequency is to be distinguished from the plasma frequency of the metal, although they are related. Gold and silver nanoparticles have plasmon resonance frequencies that lie in the visible region of the spectrum, whereas the associated plasma frequencies of these metals lie well into the ultraviolet. By virtue of the curved surfaces of nanoparticles, SPRs may be excited by direct-light illumination. The resultant intense colors exhibited by such particles, both in transmission and in reflection, are attributable to resonantly enhanced scattering and absorption.

The Metallic Nanosphere

A metallic nanosphere embedded in a surrounding dielectric medium supports LSP oscillations. The distribution of the optical field is obtained by solving Maxwell's equations in the metal and dielectric, which have negative and positive permittivities, respectively, and accommodating surface charges and the appropriate boundary conditions. The field distribution of the lowest-order plasmonic mode is depicted in Fig. 8.2-6.

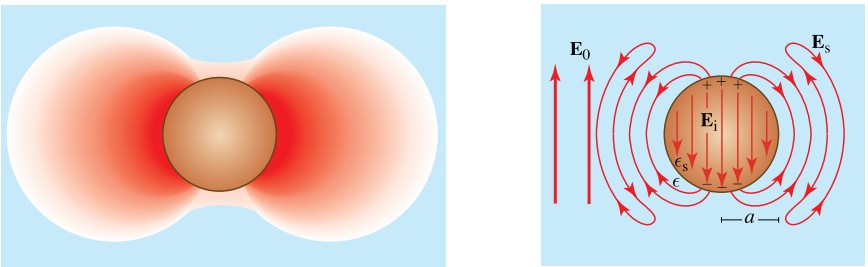


Figure 8.2-6 (*a*) Magnitude of the optical field outside a metallic nanosphere supporting the lowest-order plasmonic mode of this resonator. The internal field is not shown. (*b*) Field lines of a plasmonic mode excited by an incident plane wave.

The metallic nanosphere is a resonant scatterer. It is clear from the theory of Rayleigh scattering provided in Sec. 5.6B that a plane wave with electric field E_0 incident on a small sphere creates a parallel internal field E_i, which in turn generates an oscillating electric dipole that radiates a scattered dipole wave E_s (see Fig. 5.6-3). In accordance with (5.6-12), the total scattered optical power is $P_\mathrm{s} = \sigma_\mathrm{s} I_0$, where σ_s is the scattering cross section and I_0 is the intensity of the incident wave. Equations (5.6-16) and (5.6-17), which are applicable when the radius-to-wavelength ratio $a/\lambda \ll 1$, reveal that σ_s depends on the permittivities ϵ_s and ϵ of the nanosphere and the surrounding medium, respectively, as well as on a/λ:

$$\sigma_\mathrm{s} = \pi a^2 Q_\mathrm{s}, \qquad Q_\mathrm{s} = \frac{8}{3}\left|\frac{\epsilon_\mathrm{s} - \epsilon}{\epsilon_\mathrm{s} + 2\epsilon}\right|^2 \left(2\pi\frac{a}{\lambda}\right)^4, \tag{8.2-24}$$

$$E_{\rm i} = \frac{3\epsilon}{\epsilon_{\rm s} + 2\epsilon} E_0 \,. \tag{8.2-25}$$

For a metal described by the simplified Drude model, the effective permittivity of the metallic medium is given by $\epsilon_{\rm s} = \epsilon_o(1 - \omega_p^2/\omega^2)$, as provided in (8.2-18); $\epsilon_{\rm s}$ is thus negative or positive, depending on whether ω lies below or above the plasma frequency ω_p, respectively. It follows that the denominator $(\epsilon_{\rm s} + 2\epsilon)$ of (8.2-24) and (8.2-25) can be either negative or positive, but vanishes when $\epsilon_{\rm s} = -2\epsilon$, where $\sigma_{\rm s}$ and $E_{\rm i}$ increase without limit. The LSP resonance frequency at which this occurs is established by setting $\epsilon_o(1 - \omega_p^2/\omega^2) = -2\epsilon$, which gives rise to

$$\boxed{\omega_0 = \frac{\omega_p}{\sqrt{1 + 2\epsilon_r}} \,,} \tag{8.2-26}$$

LSP Resonance Frequency

with $\epsilon_r = \epsilon/\epsilon_o$. The surface plasmon resonance frequency ω_0 is to be distinguished from both the plasma frequency of the metal ω_p [see (8.2-16)], and the maximum frequency at which a SPP can exist ω_s [see (8.2-20)], although all three are closely related.

At frequencies near ω_0, the scattering cross section $\sigma_{\rm s}$ and the internal field $E_{\rm i}$ are substantially enhanced, as depicted in Fig. 8.2-7, which displays $\sigma_{\rm s}$ and $E_{\rm i}$ as functions of ω. Below resonance ($\omega < \omega_0$), the dipole created by the incident field points in the same direction as the incident field, while the internal field is in the opposite direction, i.e., out of phase since $E_{\rm i}/E_0$ is negative. The opposite situation prevails above resonance ($\omega > \omega_0$), as it does for the dielectric nanosphere (Fig. 5.6-3). In the

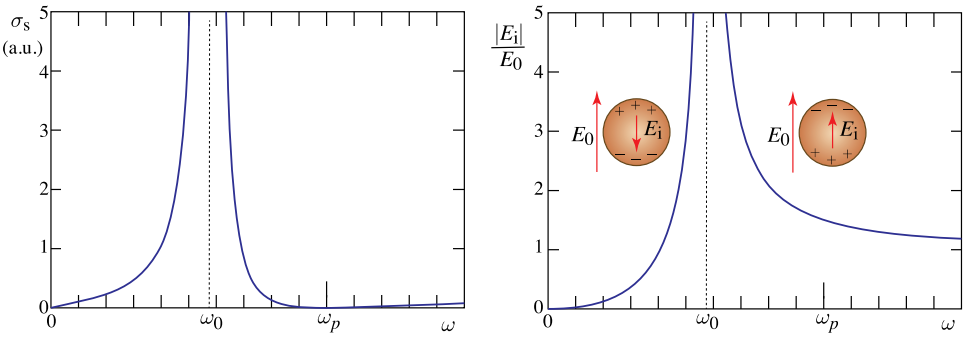


Figure 8.2-7 Resonance characteristics of the scattering cross section $\sigma_{\rm s}$ and the internal field $E_{\rm i}$ for a metallic nanosphere described by the simplified Drude model in air ($\epsilon_r = 1$). In accordance with (8.2-26), the resonance frequency is $\omega_0 = \omega_p/\sqrt{3}$, where ω_p is the plasma frequency of the metal. Below resonance, the internal field is opposite in direction to the incident external field (out of phase), whereas above resonance it is in the same direction (in phase).

vicinity of resonance, both the internal field $E_{\rm i}$ and the scattered field $E_{\rm s}$ in the near-field zone can be significantly enhanced with respect to the incident field E_0. This field enhancement is accompanied by spatial localization of electromechanical energy at a scale that corresponds to the size of the nanosphere.

The simplified Drude model used to generate the graphs presented in Fig. 8.2-7 assumes that the metal has no absorption; this idealization is the origin of the infinite cross section and infinite internal field at resonance. Of course, real metals have finite

resisitivity and finite absorption, which is mathematically accommodated by introducing an imaginary component into the permittivity. Denoting the complex permittivity as $\epsilon_s = \epsilon_s' + j\epsilon_s''$, resonance occurs when the real parts of the denominators in (8.2-24) and (8.2-25) vanish, namely when $\epsilon_s' = -2\epsilon$. Use of the complex permittivity thus results in residual denominators of $j\epsilon_s''$, rather than zero, which gives rise to a finite value of σ_s at resonance:

$$\sigma_s = \pi a^2 Q_s, \qquad Q_s = \frac{8}{3}\left(2\pi\frac{a}{\lambda}\right)^4 \frac{\epsilon_s''^2 + 9\epsilon^2}{\epsilon_s''^2}. \tag{8.2-27}$$

Aside from being a resonant scatterer, the metallic nanosphere is also a resonant absorber. As discussed in Sec. 5.6D, the Rayleigh scattering of an incident wave by a nanosphere of complex permittivity ϵ_s in a surrounding medium of real permittivity ϵ is accompanied by absorption. Both absorption and scattering contribute to the attenuation (extinction) of the incident wave. In accordance with (5.6-22), the absorption cross section of a small, spherical scatterer of radius a, $\sigma_a = \pi a^2 Q_a$, exhibits the same resonance condition as the scattering cross section σ_s set forth in (8.2-24), namely $\epsilon_s' = -2\epsilon$. Substituting $\epsilon_s = \epsilon_s' + j\epsilon_s''$ into (5.6-22) leads to an absorption efficiency Q_a given by

$$Q_a \approx -4\left(2\pi\frac{a}{\lambda}\right) \text{Im}\left\{\frac{\epsilon_s - \epsilon}{\epsilon_s + 2\epsilon}\right\} = \left(2\pi\frac{a}{\lambda}\right)\frac{-12\,\epsilon\epsilon_s''}{(\epsilon_s' + 2\epsilon)^2 + \epsilon_s''^2}. \tag{8.2-28}$$

At resonance, the denominator on the right becomes simply $\epsilon_s''^2$, which leads to a peak value of the absorption coefficient given by $Q_a = -(2\pi a/\lambda)(12\,\epsilon/\epsilon_s'')$. The larger the resistivity of the metal, which is represented by ϵ_s'', the broader the resonance profile and the smaller the peak values of σ_a and σ_s.

Metallic nanospheres whose localized surface plasmon (LSP) resonance frequencies lie in the visible and ultraviolet bands are used in applications that exploit their wavelength-selective absorption and scattering resonances, along with the attendant field enhancement and localization. Nanospheres embedded in stained glass, for example, produce brilliant colors as a result of the extinction of specific wavelengths; an example is provided by the North Rose Window at Notre Dame Cathedral in Paris, pictured at right. As another example, the dependence of the LSP resonance frequency on the relative permittivity ϵ_r of the host medium gives rise to a sensor responsive to the dielectric properties of the surrounding medium; a host medium with increased permittivity results in a decreased resonance frequency and an increased resonance wavelength, as is understood from (8.2-26).

The North Rose Window at Notre Dame Cathedral in Paris dates from 1260. Metallic nanostructures embedded in the stained glass exhibit plasmonic resonances with dramatic optical characteristics. (Adapted from a photograph by Krzysztof Mizera, August 30, 2008, Wikimedia Commons.)

D. Optical Antennas

An antenna is an electrically conductive structure that converts an oscillating electric current into an electromagnetic field, and vice versa. It is a key component in transmitters and receivers of electromagnetic radiation. At radiowave and microwave frequencies, antennas take the form of metallic wires, poles, loops, and microstrips whose dimensions are of the order of the wavelength [Fig. 8.2-8(a)]. Antennas such as these are resonant structures. A monopole antenna comprising a metal pole of length L mounted on a conducting plate, for example, has a resonance frequency $c/4L$,

corresponding to a wavelength $\lambda = 4L$. Equivalently, a dipole antenna comprising two poles separated by a small gap, each pole of length L, exhibits resonance when $L = \lambda/4$.

An antenna may also take the form of an electrically conductive structure that intercepts an electromagnetic wave and alters its angular distribution [Fig. 8.2-8(b)]. At microwave frequencies, these include the horn antenna (a metallic horn connected to the end of a waveguide) and the dish antenna (a paraboloidal metal surface with the end of a waveguide situated at its focus). These antennas are not necessarily resonant and their dimensions may be substantially greater than the wavelength of the radiation.

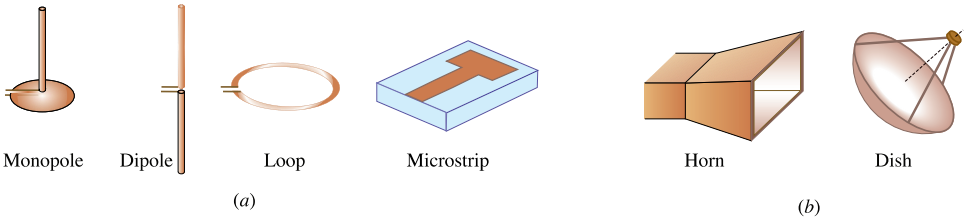


Figure 8.2-8 (a) Radiowave antennas. (b) Microwave antennas.

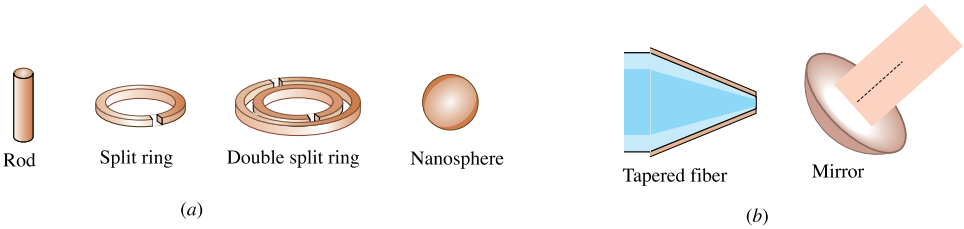


Figure 8.2-9 (a) Optical antennas made of metallic structures that exhibit resonance at optical frequencies. (b) Non-resonant optical antennas.

Resonant optical antennas may be constructed by fabricating metallic structures similar to those used for radiowave antennas, but with scaled-down dimensions [Fig. 8.2-9(a)]. However, since the length of an optical quarter-wave dipole antenna lies in the nanometer region, fabrication can be challenging. At optical frequencies, the optical field interacts with metallic antennas such as these via SPP waves, which have propagation wavelengths that are even smaller than the free-space optical wavelength. These **plasmonic antennas** operate as scatterers that convert the incoming light into localized SPP waves that in turn radiate light with a modified spatial distribution.

Optical antennas in the non-resonant category include the metal-coated tapered optical-fiber tip used in near-field microscopy and the paraboloidal mirror used in telescopes, as illustrated in Fig. 8.2-9(b). The dimensions of these optical antennas are typically far greater than the optical wavelength.

An example of an optical antenna that exhibits resonance is provided by an incoming planar optical wave illuminating a metallic nanosphere and exciting a localized SPP wave, which in turn radiates an optical dipole wave as discussed in Sec. 8.2C. At the resonance frequency, the field in the vicinity of the nanosphere is enhanced and localized, and the scattering cross-section increases sharply so that more of the incoming light is captured and scattered. The nanosphere thus functions as a resonant optical antenna. Other metallic structures with nanoscale dimensions, such as the split ring and the double split ring shown in Fig. 8.2-9(a), also exhibit resonances at optical frequencies; their resonance properties are shape- and material-dependent.

Resonant optical antennas may be used to localize and couple light into small absorbers, such as single molecules. In the near-field microscopy arrangement depicted in Fig. 8.2-10(a), for example, a metal rod on a conducting pedestal may be placed at the end of a tapered optical fiber to create a monopole antenna. A nanosphere at the end of a pointed glass tip, as illustrated in Fig. 8.2-10(b), carries out a similar function. In general, a resonant optical antenna placed between an emitter and an absorber can serve to enhance their interaction by facilitating the processes of radiation and detection.

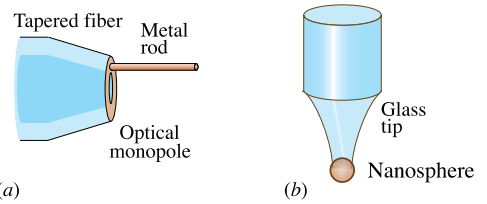


Figure 8.2-10 Optical antennas used to localize light in near-field microscopy. (a) Monopole antenna at the end of a tapered fiber. (b) Nanosphere antenna at the end of a glass tip.

The greatest challenge in modeling the interaction of an optical wave with a metallic structure arises for structures whose dimensions are comparable with the optical wavelength ($\sim\mu$m). This typically requires a complete analysis incorporating the electromagnetic fields and electric-charge distributions. Conventional bulk optical structures, which have far larger dimensions, are readily analyzed using the usual techniques of optics. At the opposite extreme, metallic nanostructures are handled quite well with effective-circuit models, as will become evident in Sec. 8.3A.

8.3 METAMATERIAL OPTICS

Optical metamaterials are synthetic composite materials constructed with carefully designed spatial patterns and with dimensions that are smaller than the optical wavelength. They owe their special optical properties to the atomic and molecular structures of the constituent materials as well as to the geometry and dimensions of the spatial patterns in relation to the wavelength. They can be engineered to have distinct and unusual optical properties that are not available in natural materials. Metamaterials form the basis for various exotic optical devices.

Photonic crystals, considered in Chapter 7, are a special class of optical metamaterials. These periodic *dielectric* structures exhibit photonic bandgaps similar to the electronic bandgaps observed in semiconductor materials. Another special class of metamaterials, described in this section, makes use of *metallic* elements of subwavelength dimensions, such as rods and rings, that are embedded in dielectric media and organized in periodic or random patterns at a subwavelength spatial scale. The shapes of these building blocks, and the patterns in which they are arranged, are designed so that the effective electric and magnetic material parameters, ϵ and μ respectively, are rendered either positive or negative. This in turn allows the synthesis of SNG and DNG materials, with their attendant special optical properties, as detailed in Sec. 8.1.

When the metallic elements exhibit resonance, the effective behavior of ϵ and μ can display the familiar frequency dependence of the susceptibility of a resonant medium, portrayed in Fig. 5.5-6. At frequencies above the resonance frequency, as shown in Fig. 8.3-1, the real part can be negative so that the medium can be SNG or DNG. Although the imaginary part, corresponding to attenuation, is significant near resonance, a narrow band with negative real part can persist at sufficiently high frequencies where the attenuation is minimal.

Since $\mu = \mu_o$ at optical frequencies for naturally occurring nonmagnetic materials, DNG (negative-index) media cannot be readily created without the use of metamaterials. The fabrication of optical DNG metamaterials is undeniably challenging. First, the metallic elements must be of subwavelength dimensions so that resonance frequencies lie in the optical band. This requires nanoscale fabrication technology. Second, the electric and magnetic resonance frequencies must be sufficiently close to each other so that the bands of negative ϵ and μ align, as depicted in Fig. 8.3-1. A great deal of effort has been directed toward addressing these challenges since the field of metamaterials has come to the fore. The principles underlying metamaterials also offer promise in the domains of acoustics, mechanics, and thermodynamics.

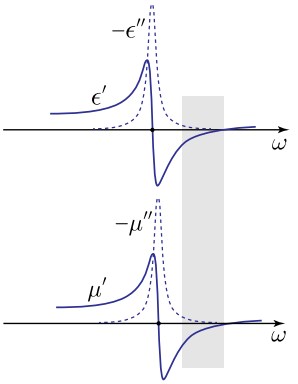


Figure 8.3-1 A medium with resonant permittivity and permeability can behave as a DNG medium in a frequency band above both resonance frequencies (shaded area).

We consider three-dimensional optical metamaterials, and metamaterials of reduced dimensionality, known as metasurfaces, in turn.

A. Metamaterials

The effective electromagnetic parameters ϵ and μ for metamaterials comprising three-dimensional distributions of metallic elements embedded in a dielectric host material may be determined by making use of approximate models, complex analytical techniques, or numerical methods. Approximate models include the effective-medium approach and the effective-circuit approach. The effective-medium approach based on the use of the Maxwell-Garnett formula described in Sec. 5.6D is, strictly speaking, applicable only for nanospheres. However, it is widely used in practice for metallic particles of other shapes as well.

The effective-circuit approach, on the other hand, is applicable for metallic elements of various shapes, provided they are sufficiently small. The dielectric/magnetic properties of a natural material are usually determined by summing the electric/magnetic dipoles of the constituent atoms induced by the applied electric/magnetic fields. This enables the polarization and magnetization densities to be determined, and thence the electric permittivity ϵ and magnetic permeability μ. A similar approach may be adopted for metamaterials, in which each constituent element, regarded as a Rayleigh scatterer, is modeled by electric and/or magnetic dipoles. To determine the electric and magnetic dipole moments, metallic structures of subwavelength dimensions are considered as electric circuit elements, an approach known as the **point-dipole approximation**.

Larger elements may be modeled using Mie scattering theory (see Sec. 5.6C), in which the dipole contributions become the leading terms of multipole series expansions. More complex circuit models can also be used. However, when metallic nanoparticles are juxtaposed in sufficiently close proximity, such that their localized plasmonic fields overlap, the foregoing approximations fail to account for element-to-element interactions and inter-element resonances. Such effects may be accommodated via the

circuit approach by considering mutual couplings between neighboring elements, such as mutual inductances, and by treating the composite circuits as transmission lines or circuit networks. When approximations such as these fail, numerical methods offer a fallback position.

We proceed to examine several approximate models based on both the effective-medium and effective-circuit approaches. The models we consider describe metamaterials exhibiting negative permittivity, negative permeability, negative index, as well as hyperbolic properties.

Negative-Permittivity Metamaterial: Metallic Nanospheres in a Dielectric Medium

As described in Sec. 5.6D and illustrated in Fig. 8.3-2, a dielectric medium of electric permittivity ϵ that is uniformly filled with small nanospheres of complex permittivity $\epsilon_{\rm s}$ yields an isotropic composite medium of effective permittivity $\epsilon_{\rm e}$ that obeys the Maxwell-Garnett mixing rule (5.6-20),

$$\epsilon_{\rm e} \approx \epsilon \, \frac{2(1-{\rm f})\epsilon + (1+2{\rm f})\epsilon_{\rm s}}{(2+{\rm f})\epsilon + (1-{\rm f})\epsilon_{\rm s}} \,, \tag{8.3-1}$$

where f is the volume fraction of the inclusions (filling ratio).

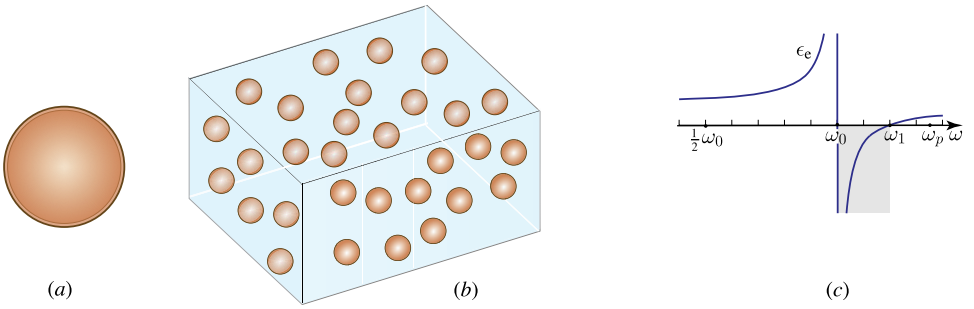


Figure 8.3-2 Negative-permittivity metamaterial. (a) Metallic nanosphere. (b) Metamaterial comprising uniformly distributed metallic nanospheres embedded in a dielectric medium. (c) The effective permittivity $\epsilon_{\rm e}$ has a pole at the resonance frequency ω_0 and a zero at ω_1. Since $\epsilon_{\rm e}$ is negative in the band lying between $\omega_0 < \omega < \omega_1$, this metamaterial is single-negative (SNG) if μ is positive.

For metallic elements described by the simplified Drude model, in accordance with (8.2-18) we have $\epsilon_{\rm s} = \epsilon_o(1-\omega_p^2/\omega^2)$, where ω_p is the plasma frequency, so that (8.3-1) yields

$$\epsilon_{\rm e} = \epsilon_{\rm L} \, \frac{1-\omega^2/\omega_1^2}{1-\omega^2/\omega_0^2} \,, \tag{8.3-2}$$

where

$$\omega_0 = \frac{\omega_p}{\sqrt{1+\epsilon_{r0}}} \,, \qquad \omega_1 = \frac{\omega_p}{\sqrt{1+\epsilon_{r1}}} \,, \tag{8.3-3}$$

$$\epsilon_{\rm L} = \frac{1+2{\rm f}}{1-{\rm f}} \, \epsilon \,, \qquad \epsilon_{r0} = \frac{2+{\rm f}}{1-{\rm f}} \, \epsilon_r \,, \qquad \epsilon_{r1} = \frac{2(1-{\rm f})}{1+2{\rm f}} \, \epsilon_r \,, \tag{8.3-4}$$

and where $\epsilon_r = \epsilon/\epsilon_o$ is the relative permittivity of the host medium, which is assumed to be frequency-independent.

As shown in Fig. 8.3-2(c), the effective permittivity $\epsilon_{\rm e}$ has a pole at ω_0 and a zero at ω_1. Since $\epsilon_{r0} > \epsilon_r$, the resonance frequency ω_0 falls below that of the isolated

nanosphere, which is given by (8.2-26). Also, since $\epsilon_{r1} < \epsilon_{r0}$, we see that $\omega_1 > \omega_0$, so that ϵ_e is negative within the spectral band between ω_0 and ω_1, which lies below the plasma frequency ω_p of the metal. With μ positive, this metamaterial is therefore single-negative (SNG), much like a homogeneous metal below its plasma frequency [see Fig. 8.2-1(b)].

Negative-Permittivity Metamaterial: Thin Metallic Rods Isotropically Distributed in a Dielectric Medium

The inductance L of a cylindrical metallic rod of length a and radius w ($a \gg w$) [Fig. 8.3-3(a)] is given by $L \approx (\mu_o a/2\pi)\,[\ln(2a/w) - 3/4]$. The effective electric permittivity of a medium comprising parallel rods separated by a distance a, as depicted in Fig. 8.3-3(b), is determined by observing that an electric field E along a rod develops a voltage $V = aE$ between its two ends. This in turn generates an electric current $i = V/j\omega L$ in the inductor, which corresponds to a charge $q = i/j\omega$ and an electric dipole moment $\mathrm{p} = qa$. Since the number of rods per unit volume is $N = 1/a^3$, the polarization density is given by $P = N\mathrm{p} = \mathrm{p}/a^3$. The effective susceptibility of the medium is thus $\chi_e = P/\epsilon_o E$ and the effective permittivity is $\epsilon_e = \epsilon_o(1 + \chi_e)$.

Combining these equations leads to an expression for the effective permittivity that is identical in form to that of a simple Drude metal,

$$\epsilon_e = \epsilon_o\left(1 - \frac{\omega_p^2}{\omega^2}\right), \qquad \omega_p = \frac{1}{\sqrt{\epsilon_o a L}} = 2\pi\frac{c_o}{a}\frac{1}{\sqrt{2\pi\ln(2a/w) - \frac{3}{2}\pi}}, \tag{8.3-5}$$

with a plasma frequency ω_p determined by the dimensions of the rod, a and w, via its inductance L, where we have assumed that the dielectric medium has the permittivity of free space. With μ positive, this metamaterial is thus single-negative (SNG). Though the rod is assumed to be a perfect conductor in the calculations set forth above, loss is readily accommodated by adding a resistance R to the impedance $j\omega L$ of the rod.

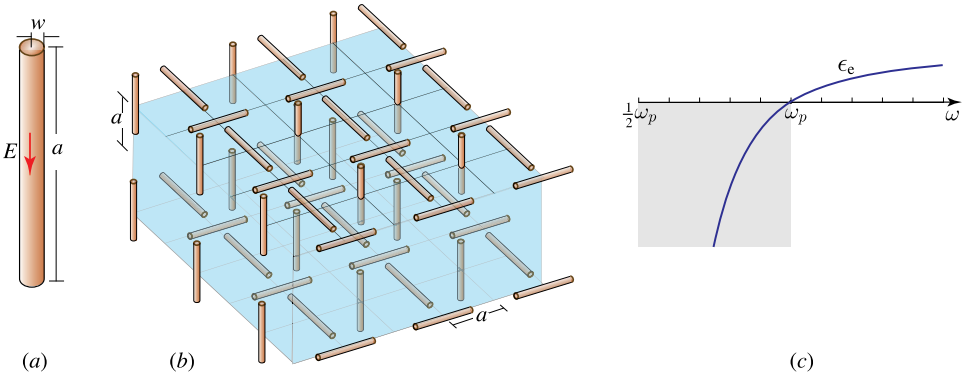


Figure 8.3-3 Negative-permittivity metamaterial. (a) Thin metallic rods of length a and radius w, oriented in (b) three orthogonal directions at every point of a cubic lattice of dimension a, create an isotropic metamaterial. (c) The effective electric permittivity ϵ_e has a frequency dependence identical to that of the simplified Drude model. This metamaterial is thus single-negative (SNG), provided that μ is positive.

Negative-Permeability Metamaterial: Split-Ring Metallic Elements in a Dielectric Medium

A metallic split ring [Fig. 8.3-4(a)], modeled as an inductor L in series with a capacitor C (the open section of the ring), forms a resonant circuit of resonance frequency $\omega_0 =$

$1/\sqrt{LC}$. When the size of the split ring is $\lesssim 100$ nm and the gap is $\lesssim 10$ nm, ω_0 lies in the optical region of the spectrum.

The effective magnetic permeability μ_e of a metamaterial consisting of a collection of such split rings, organized uniformly and in three directions at the vertices of a periodic lattice, as shown in Fig. 8.3-4(b), may be established by calculating the magnetic dipole moment $\mathfrak{m}$ induced by a magnetic field H applied along an axis normal to the plane of the ring [Fig. 8.3-4(a)]. The voltage V induced in the loop is equal to the rate of change of the magnetic flux so that $V = -j\omega\mu_o AH$, where A is the area of the ring. This voltage generates an electric current $i = V/Z$, where the circuit impedance is $Z = j\omega L + 1/j\omega C$. This electric current in turn results in a magnetic dipole moment $\mathfrak{m} = Ai$.

A density of N split rings per unit volume gives rise to a magnetization density $M = N\mathfrak{m}$ so that the effective magnetic permeability $\mu_e = \mu_o(H + M)/H$ is given by

$$\mu_e = \mu_o \frac{1 - \omega^2/\omega_1^2}{1 - \omega^2/\omega_0^2}, \qquad \omega_0 = \frac{1}{\sqrt{LC}}, \qquad \omega_1 = \frac{\omega_0}{\sqrt{1 - \mu_o NA^2/L}}. \tag{8.3-6}$$

The inductance of the ring is $L \approx \mu_o b\,[\ln(8b/a) - 7/4]$, where b and a are the ring and wire radii, respectively ($b \gg a$). As displayed in Fig. 8.3-4(c), μ_e exhibits a resonance at ω_0 and a zero at ω_1, and is negative in the intervening region. For a positive electric permittivity ϵ, this structure behaves as a single-negative (SNG) metamaterial. The frequency dependence of μ_e revealed in (8.3-6) is the same as that of the effective permittivity ϵ_e for metallic nanospheres set forth in (8.3-2).

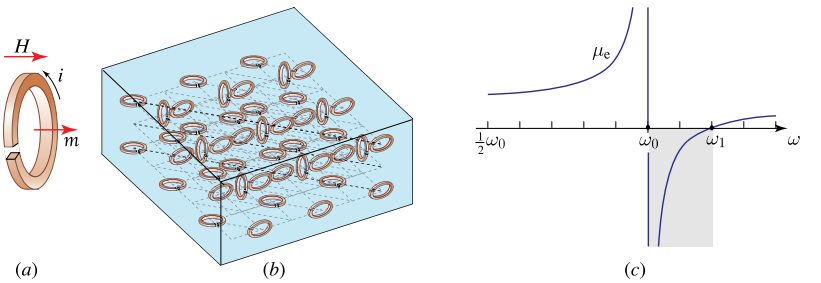


Figure 8.3-4 Negative-permeability metamaterial. (a) A metallic split ring excited by a magnetic field exhibits a magnetic dipole moment $\mathfrak{m}$. (b) An isotropic metamaterial fabricated by configuring such split rings in three directions at the vertices of a cubic lattice. (c) The frequency dependence of the effective magnetic permeability μ_e exhibits a pole at ω_0, a zero at ω_1, and is negative in the intervening range. When ϵ is positive, this structure serves as a single-negative (SNG) metamaterial.

Negative-Index Metamaterials

The negative electric permittivity of the metallic-rod metamaterial [Fig. 8.3-3(c)] may be combined with the negative magnetic permeability of the metallic split-ring metamaterial [Fig. 8.3-4(c)] to create a double-negative (DNG) metamaterial that serves as a negative-index material (NIM). Implementation is achieved by repeating the combined rod and double split-ring element, displayed in Fig. 8.3-5(a), in two directions, as illustrated in Fig. 8.3-5(b). This approach requires oblique incidence of the wave, as indicated, so that surface plasmon resonances (SPR) can be excited with out-of-plane magnetic fields. This design was first experimentally demonstrated in the microwave

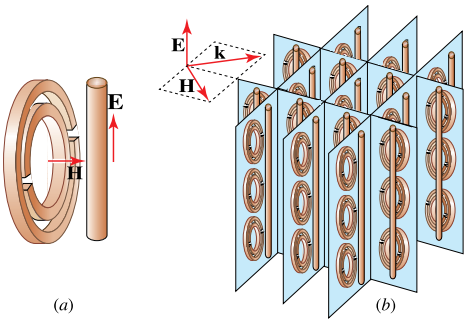


Figure 8.3-5 Negative-index metamaterial. (*a*) Combined rod and double split-ring element. (*b*) DNG metamaterial comprising an array of the elements shown in (*a*), oriented along two orthogonal directions. For waves traveling along directions in the horizontal plane, the medium is double-negative, and the refractive index is negative, at frequencies above the permittivity and permeability resonances, as schematized in Fig. 8.3-1.

region, and its dimensions were subsequently scaled down for operation at optical frequencies.

Alternative designs for optical NIMs that are easier to fabricate have subsequently been developed. One such design is a "fishnet" metal–dielectric multilayer structure, a simplified version of which is illustrated in Fig. 8.3-6. In this configuration, the optical wave is normally incident on the fishnet surface and the electric and magnetic fields are aligned with the metal strips, as shown. The strips aligned with the electric field are responsible for the negative permittivity. The strips aligned with the magnetic field support anti-symmetric resonant modes between pairs of coupled strips, which results in negative permeability above the resonance frequency. Fishnet nanostructures serve as NIMs that operate in the visible region.

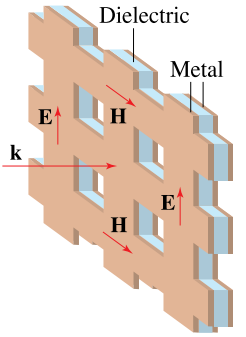


Figure 8.3-6 Simplified version of a "fishnet" metal–dielectric nanostructured composite metamaterial consisting of a stacked network of intersecting subwavelength plasmonic waveguides. The structure serves as a negative-index material (NIM) in the visible region of the spectrum.

*Hyperbolic Metamaterial: Parallel Metallic Rods in a Dielectric Medium

As described in Sec. 8.1C, an anisotropic medium is said to be hyperbolic if the effective principal values of its permittivity (or permeability) tensor have mixed signs. We demonstrate that an array of parallel metallic rods embedded in a dielectric medium of permittivity ϵ, such as those illustrated in Fig. 8.3-7, may exhibit hyperbolic behavior when the rods have extreme anisotropy by virtue of a large aspect ratio.

For this purpose, we use the effective-medium approach to write expressions for the principal values of the permittivity tensor along the x, y, z directions (denoted 1, 2, 3, respectively). An analysis similar to that carried out for the dielectric sphere, the result of which is provided in (5.6-17), reveals that the ratio of internal and external fields for a dielectric cylinder takes the form $E_\mathrm{i}/E_0 = 2\epsilon/(\epsilon_s + \epsilon)$. Using the Maxwell–Garnett mixing rule then yields effective permittivities for fields in the plane of the cross section, and along the axial direction (where $E_\mathrm{i}/E_0 = 1$), that are given by:

$$\epsilon_{\mathrm{e}1} = \epsilon_{\mathrm{e}2} = \epsilon\,\frac{2\mathrm{f}\epsilon_\mathrm{s} + (1-\mathrm{f})(\epsilon_\mathrm{s}+\epsilon)}{\epsilon_\mathrm{s} + \epsilon + \mathrm{f}(\epsilon - \epsilon_\mathrm{s})}, \qquad \epsilon_{\mathrm{e}3} = (1-\mathrm{f})\epsilon + \mathrm{f}\epsilon_\mathrm{s}\,. \tag{8.3-7}$$

Making use of the simplified Drude model (8.2-18) for $\epsilon_{\rm s}$, combining the results for metallic nanospheres (8.3-2)–(8.3-4) and metallic rods (8.3-5), and using (8.2-20), we obtain the following expressions for the permittivity principal values:

$$\epsilon_{\rm e1} = \epsilon_{\rm e2} = \epsilon_{\rm L}\,\frac{1-\omega^2/\omega_1^2}{1-\omega^2/\omega_0^2}\,, \qquad \epsilon_{\rm e3} = \epsilon_{\rm H}\left(1-\frac{\omega_3^2}{\omega^2}\right), \tag{8.3-8}$$

where

$$\omega_0 = \frac{\omega_p}{\sqrt{1+\epsilon_{r0}}}, \qquad \omega_1 = \frac{\omega_p}{\sqrt{1+\epsilon_{r1}}}, \qquad \omega_3 = \frac{\omega_p}{\sqrt{1+\epsilon_{r3}}}, \tag{8.3-9}$$

$$\epsilon_{r0} = \frac{1+{\rm f}}{1-{\rm f}}\,\epsilon_r, \qquad \epsilon_{r1} = \frac{1-{\rm f}}{1+{\rm f}}\,\epsilon_r, \qquad \epsilon_{r3} = \frac{1-{\rm f}}{\rm f}\,\epsilon_r, \tag{8.3-10}$$

$$\epsilon_{\rm L} = \frac{1+{\rm f}}{1-{\rm f}}\,\epsilon, \qquad \epsilon_{\rm H} = \epsilon_o\left[{\rm f}+(1-{\rm f})\epsilon_r\right], \tag{8.3-11}$$

with $\epsilon_r = \epsilon/\epsilon_o$.

The effective ordinary permittivity $\epsilon_{\rm e1} = \epsilon_{\rm e2}$ exhibits a pole at the resonance frequency ω_0 and a zero at ω_1, while the effective extraordinary permittivity $\epsilon_{\rm e3}$ varies from negative values at low frequencies to positive values at high frequencies, passing through zero at ω_3, much like a pure metal. As illustrated in Fig. 8.3-7, an extensive frequency band exists for which $\epsilon_{\rm e1} = \epsilon_{\rm e2}$ and $\epsilon_{\rm e3}$ have opposite signs, so that the anisotropic medium is hyperbolic. Within this band, the material behaves like a dielectric in one direction and a metal in the orthogonal direction.

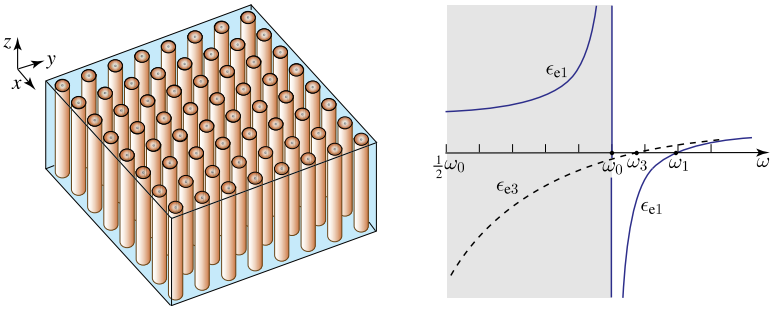


Figure 8.3-7 Hyperbolic metamaterial. Frequency dependence of the principal components of the effective electric permittivity tensor for a metamaterial comprising parallel metallic rods described by the simplified Drude model embedded in a host medium of permittivity ϵ. The ordinary effective permittivities $\epsilon_{\rm e1} = \epsilon_{\rm e2}$ have resonance frequencies ω_0 and exhibit positive values below resonance. The extraordinary effective permittivity $\epsilon_{\rm e3}$ is negative in the frequency range $\omega < \omega_3$. The spectral band over which the effective medium is hyperbolic is delineated by the shaded region, where $\epsilon_{\rm e1} = \epsilon_{\rm e2}$ and $\epsilon_{\rm e3}$ have opposite signs.

B. Metasurfaces

Metasurfaces are metamaterials whose dimensionality is reduced from three to two. Examples are ultrathin arrays of subwavelength-scale metallic elements, deposited in periodic, aperiodic, or random patterns, on the surface of a dielectric substrate. **Complementary metasurfaces** comprise arrays of subwavelength-scale dielectric elements, such as holes and their separations, arrayed on an ultrathin metallic surface. The shapes of the individual elements, and the geometry of their layout on the surface, endow metasurfaces with distinctive optical properties that are a consequence of the coupling of light and SPP waves generated at the metal–dielectric boundary.

Metasurface as a Phase Modulator

As explained in Sec. 2.4, a wave traveling in the z direction, on transmission through a dielectric plate of fixed thickness d and graded refractive index $n(x, y)$ in the x–y plane, undergoes a spatially varying phase shift $\varphi(x, y) = n(x, y)k_o d$, which modifies its wavefront [see (2.4-14)]. Achieving a phase shift of 2π requires a local thickness equal to the wavelength of light in the medium. A planar metasurface has the merit that it can introduce a phase shift of similar magnitude with far less thickness. The metallic elements of the metasurface function much like optical antennas that modify the optical wavefront. A resonant antenna acts as a scatterer that introduces a frequency-dependent phase shift that ranges from $-\pi/2$ to $\pi/2$ for frequencies below to above resonance, respectively.

A spatially varying phase shift $\varphi(x, y)$ may be implemented by making use of a metasurface comprising elements of spatially graded sizes and geometries that correspond to spatially varying resonance frequencies. An incoming wave of fixed frequency is then subjected to a spatially varying phase shift so that the metasurface acts as a phase modulator. An example is provided in Fig. 8.3-8(a). Since the metasurface is ultra thin, it may be modeled mathematically as an optical component that introduces a spatially varying phase discontinuity (i.e., a phase shift that takes place over a distance $d \to 0$).

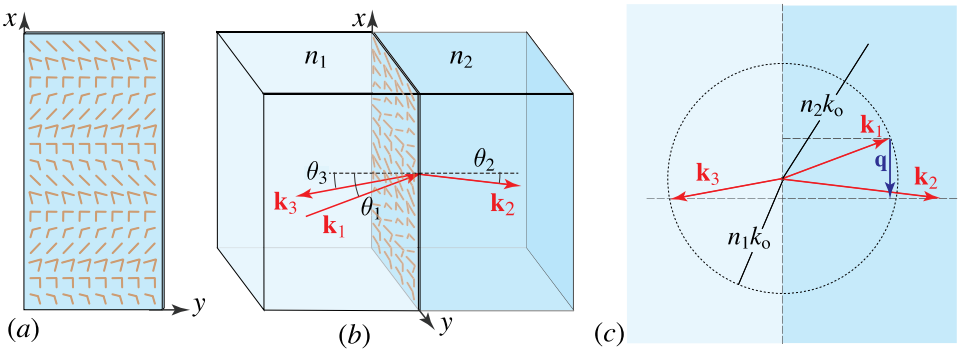


Figure 8.3-8 (a) A metasurface using an array of metallic elements whose shapes and resonance frequencies vary in the x direction. The shapes of the elements are engineered such that the phase shift they introduce is a linear function $\varphi = qx$ for one of the polarization components. (b) Negative reflection and negative refraction at a boundary between two media of refractive indices n_1 and n_2 by virtue of the presence of the metasurface portrayed in (a) between the two media. (c) Phase-matching condition for the incident and refracted waves, and for the incident and reflected waves, at the metasurface boundary.

The phase modulation introduced by such a metasurface may modify an incoming optical wave in any of the many ways described in Sec. 2.4. A salutary feature of this approach is that the wave undergoes minimal spatial spread (diffraction) as it crosses the infinitesimally thin metasurface. Consider, for example, a phase $\varphi(x, y)$ that varies linearly along the metasurface at a rate q, so that $\varphi = qx$. The complex amplitude of the incoming wave is then modulated by the factor $\exp(-jqx)$, which is a periodic function of the spatial frequency $\nu_x = q/2\pi$, as explained in Sec. 4.1A. An incoming plane wave of wavevector $\mathbf{k}_1$ will then generate refracted and reflected plane waves with wavevectors $\mathbf{k}_2$ and $\mathbf{k}_3$, respectively.

To ensure phase matching at both sides of the surface, as depicted in Fig. 8.3-8(c), the component of the vector $\mathbf{k}_2$ parallel to the surface must match that of $\mathbf{k}_1 + \mathbf{q}$, where $\mathbf{q}$ is a vector of magnitude q pointing in the x direction. Likewise, for the reflected

wave, the component of the vector $\mathbf{k}_3$ along the surface must match that of $\mathbf{k}_1 + \mathbf{q}$. Hence, if the metasurface lies at the boundary between two ordinary media of refractive indices n_1 and n_2, its presence causes the conventional Snell's law of refraction and reflection to assume the following modified form:

$$n_2 k_o \sin\theta_2 = n_1 k_o \sin\theta_1 + q, \tag{8.3-12}$$
Metasurface Refraction

$$n_1 k_o \sin\theta_3 = n_1 k_o \sin\theta_1 + q, \tag{8.3-13}$$
Metasurface Reflection

where θ_1, θ_2, and θ_3 are the angles of incidence, refraction, and reflection, respectively.

With appropriate choice of the magnitude and sign of q, the presence of the metasurface can result in *negative reflection* and *negative refraction* at the boundary, as illustrated in Fig. 8.3-8(b). Equations (8.3-12) and (8.3-13) properly reduce to Snell's law when $q = 0$.

For a phase discontinuity $\varphi(x)$ that varies slowly with the position x, the derivative $q = d\varphi/dx$ may be regarded as the local spatial frequency at x. This quantity determines the local tilt imparted to an incoming wavefront, and thus the angles of reflection and refraction as a function of x. This approach can clearly be generalized to metasurfaces that introduce a two-dimensional phase discontinuity $\varphi(x, y)$. In that case, the vector $\mathbf{q} = \nabla\varphi$ represents the magnitude and direction of the local spatial frequency of the phase modulation. The metasurface can therefore be designed to introduced desired local tilts of the wavefront in both the x–z and y–z planes, much like an antenna array or an optical phase plate. The metasurface can also be engineered to introduce position-dependent amplitude modulation, imparted by the shape of the local elements. The combination of phase and amplitude modulation can serve as a hologram with complex transmittance that is designed to simulate the wavefront of light generated by an object.

Extraordinary Optical Transmission Through Subwavelength Holes in a Metallic Film

A metasurface consisting of a periodic array of subwavelength holes and separations perforated in a planar metallic film may exhibit extraordinarily high optical-power transmittance at certain wavelengths and angles of incidence. The power transmittance $\mathcal{T}(\lambda, \theta)$, as a function of the wavelength λ and the angle of incidence θ, is found to exhibit sharp peaks with values that significantly exceed those predicated on the basis of conventional diffraction theory. Indeed, if $\mathcal{T}_h$ is the total hole area per unit area of the film, the peak values of the transmittance $\mathcal{T}(\lambda, 0)$ for a plane wave traveling in a direction orthogonal to the film may exceed $\mathcal{T}_h$ by orders of magnitude.

This phenomenon is attributable to the generation of SPP waves through the holes and the attendant radiation of light by the oscillating charges created in the metallic film. The array of subwavelength holes in the metal film should thus be considered as an active radiating element, rather than as a passive geometrical aperture through which the light is transmitted.

Maximum transmission occurs at frequencies for which the incident optical wave and the excited SPP wave are phase matched. For a periodic array of holes in the form of a square lattice with period a_0, the phase-matching condition is

$$\beta = k_\perp \pm m_x \boldsymbol{g}_x \pm m_y \boldsymbol{g}_y, \tag{8.3-14}$$

where $\beta = (\omega/c_o)\sqrt{\epsilon_\mathrm{b}/\epsilon_o}$ is the propagation constant of the SPP wave [see (8.2-21)]; $k_\perp = (2\pi/\lambda)\sin\theta$ is the component of the wavevector of the incident light in the plane

of the array; $g_x = g_y = 2\pi/a_0$ are the fundamental spatial frequencies of the periodic array; and m_x and m_y are integers representing associated spatial harmonics (the scattering order). The transmittance $\mathcal{T}(\lambda, \theta)$ as a function of the angle of incidence θ also exhibits photonic band gaps,[†] much like those observed in photonic crystals, which are three-dimensional periodic dielectric structures of subwavelength dimensions (see Sec. 7.3).

Similar extraordinary optical transmission is expected for holes in a perfect conductor, rather than a real metal. The medium is then characterized by an effective dielectric function that has a plasmon form, with a plasma frequency dictated by the geometry of the holes.[‡]

*8.4 TRANSFORMATION OPTICS

In other chapters of this book, graded-index optics is considered from an *analysis* perspective, with a mandate to *determine how light propagates in a medium endowed with particular dielectric and magnetic properties*. Examples are provided in the context of ray optics, wave optics, and electromagnetic optics:

- Graded-index (GRIN) materials allow optical rays to follow curved trajectories governed by the profile of the refractive index $n(\mathbf{r})$ (Sec. 1.3).
- The trajectories of scalar waves in GRIN materials can be described in terms of the Eikonal equation (Secs. 2.3 and 10.2C).
- For isotropic materials with graded electric permittivity $\epsilon(\mathbf{r})$ and magnetic permeability $\mu(\mathbf{r})$, Maxwell's equations give rise to the generalized Helmholtz equations (5.2-16) and (5.2-17); these equations were solved in Chapter 7 for layered and periodic structures such as photonic crystals.
- The optics of anisotropic graded media are described by Maxwell's equations with position-dependent tensors $\boldsymbol{\epsilon}(\mathbf{r})$ and $\boldsymbol{\mu}(\mathbf{r})$; the solution requires full vector analysis.
- GRIN optics is useful for fabricating a variety of optical components, including GRIN lenses (Exercise 1.3-1) and GRIN optical fibers (Sec. 10.1B).

In this section, in contrast, we consider graded-index optics from a *synthesis* or *design* perspective, where the goal is to *determine the dielectric and magnetic properties of a medium that realize a desired pattern of light propagation.* The synthesis problem is more challenging than the analysis problem in two respects: (1) the required mathematical tools are more advanced; and (2) physical implementation of the graded medium often requires the use of metamaterials constructed from components that are available, configured in a particular spatial arrangement, and amenable to being fabricated with current technology.

A. Transformation Optics

Transformation optics is a mathematical tool that facilitates the design of optical materials that guide light along desired trajectories. The underlying concept relies on a geometrical transformation that converts simple trajectories into desired ones. In order that Maxwell's equations remain valid, the optical parameters associated with

[†] T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio, and P. A. Wolff, Extraordinary Optical Transmission Through Sub-Wavelength Hole Arrays, *Nature*, vol. 391, pp. 667–669, 1998.

[‡] F. J. Garcia-Vidal, L. Martín-Moreno, and J. B. Pendry, Surfaces with Holes in Them: New Plasmonic Metamaterials, *Journal of Optics A: Pure and Applied Optics*, vol. 7, pp. S97–S101, 2005.

the transformed equivalent system must also be modified, and this establishes the character of the required optical material. As a simple example of such equivalence, a local compression of the coordinate system by a scaling factor is equivalent to a local increase of the refractive index by the same factor, so that the optical pathlength (product of length and refractive index) remains unchanged.

A three-step design procedure provides a guide:

- Begin with a *pilot* physical system for which the optical trajectories are known, such as a homogeneous and isotropic material.
- Find a coordinate transformation that converts these trajectories to the desired ones.
- Determine the transformed physical parameters of the equivalent material. The new material will implement the desired optical trajectories in the original coordinate system.

Since geometrical transformations generally involve changes of directions and introduce direction-dependent scaling, the transformed parameters are generally both anisotropic and spatially varying.

Transformation Principle

Let $\{\epsilon_{ij}\}$ and $\{\mu_{ij}\}$ be the elements of the permittivity and permeability tensors of the original material in the original coordinate system (x_1, x_2, x_3). The elements of the permittivity and permeability tensors of the equivalent material (denoted by the superscript "e") in the transformed coordinate system (u_1, u_2, u_3) are then related to the original elements by the matrix equations[†]

$$\boldsymbol{\epsilon}^{\mathrm{e}} = |\det \mathbf{A}|^{-1}\, \mathbf{A}^{\mathsf{T}} \boldsymbol{\epsilon}\, \mathbf{A}\,, \quad \boldsymbol{\mu}^{\mathrm{e}} = |\det \mathbf{A}|^{-1}\, \mathbf{A}^{\mathsf{T}} \boldsymbol{\mu}\, \mathbf{A}\,. \tag{8.4-1}$$

Here $\mathbf{A}$ is the 3×3 Jacobian transformation matrix, whose elements are the partial derivatives

$$A_{ij} = \frac{\partial u_i}{\partial x_j}, \quad i, j = 1, 2, 3\,. \tag{8.4-2}$$

The quantity $\mathbf{A}^{\mathsf{T}}$ is the transpose of $\mathbf{A}$, and $\boldsymbol{\epsilon}$ and $\boldsymbol{\mu}$ are 3×3 matrices whose elements are $\{\epsilon_{ij}\}$ and $\{\mu_{ij}\}$, respectively. Since $\mathbf{A}$ is generally dependent on (x_1, x_2, x_3), the equivalent material is generally inhomogeneous, even if the original material is homogeneous.

In the special case for which the original material is both homogeneous and isotropic, say free space, then $\boldsymbol{\epsilon}$ and $\boldsymbol{\mu}$ are diagonal with equal diagonal elements ϵ_o and μ_o, respectively, whereupon

$$\epsilon_o^{-1} \boldsymbol{\epsilon}^{\mathrm{e}} = \mu_o^{-1} \boldsymbol{\mu}^{\mathrm{e}} = |\det \mathbf{A}|^{-1}\, \mathbf{A}^{\mathsf{T}} \mathbf{A}. \tag{8.4-3}$$

The tensors $\boldsymbol{\epsilon}^{\mathrm{e}}$ and $\boldsymbol{\mu}^{\mathrm{e}}$ are then identical except for a scaling factor. Under these conditions the impedance, which depends on their ratio, remains unchanged for all polarizations, which in turn implies that the equivalent medium introduces no reflection at any boundary with free space.

We provide a number of examples to illustrate the transformation principle:

[†] See, e.g., J. B. Pendry, Y. Luo, and R. Zhao, Transforming the Optical Landscape, *Science*, vol. 348, pp. 521–524, 2015.

EXAMPLE 8.4-1. ***Refraction Without Reflection.*** In this example, we design an optical material implementing ray trajectories that refract without reflection at a planar surface, as shown in Fig. 8.4-1(a).

We begin with an initial homogeneous medium, say free space, with rays that follow parallel straight trajectories at an angle θ_1, as shown in Fig. 8.4-1(b). We now apply a geometrical transformation that stretches the coordinate system by a scale factor s along the x_3 direction in the region $x_3 > 0$. The desired refraction is achieved by choosing s as the ratio of the initial and desired slopes, $s = \tan\theta_1/\tan\theta_2$ [Fig. 8.4-1(c)]. This transformation is implemented by the relations

$$u_1 = x_1, \quad u_2 = x_2, \quad u_3 = s^{-1}x_3. \tag{8.4-4}$$

This type of scaling of the Cartesian coordinate system, in which the directions of the axes do not change, converts a cube into a cuboid. Based on (8.4-2), the Jacobian matrix $\mathbf{A}$ is diagonal with diagonal elements $(1, 1, s^{-1})$ and determinant $\det \mathbf{A} = s^{-1}$, so that (8.4-3) provides

$$\epsilon_o^{-1}\boldsymbol{\epsilon}^{\mathrm{e}} = \mu_o^{-1}\boldsymbol{\mu}^{\mathrm{e}} = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & s^{-1} \end{bmatrix}. \tag{8.4-5}$$

Since the matrices $\boldsymbol{\epsilon}^{\mathrm{e}}$ and $\boldsymbol{\mu}^{\mathrm{e}}$ are diagonal, the anisotropic material has principal axes pointing along the axes of the coordinate system. The principal values are: $\epsilon_1 = s\epsilon_o$, $\epsilon_2 = s\epsilon_o$, and $\epsilon_3 = s^{-1}\epsilon_o$ together with $\mu_1 = s\mu_o$, $\mu_2 = s\mu_o$, and $\mu_3 = s^{-1}\mu_o$.

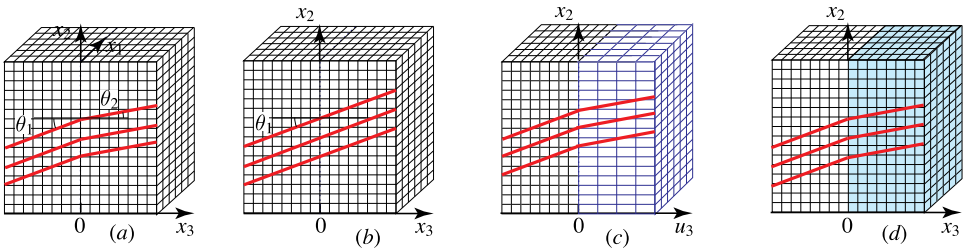


Figure 8.4-1 Geometrical transformation implementing refraction without reflection. (a) Desired optical trajectories. (b) Free space with straight-line optical trajectories. (c) Stretching the coordinate system by the factor s for $x_3 > 0$ causes the rays to change slope and follow the desired trajectories. (d) Equivalent anisotropic, homogeneous material that causes the rays to change slope in an identical way.

The parameters of the equivalent material may also be obtained by matching the phase shift encountered when a plane wave crosses the stretched free-space segment with that encountered when the wave is transmitted through an unstretched segment filled with the new material. To determine the parameters, we consider three waves in turn, each with the electric field along one of the coordinates:
– Wave 1 is a plane wave traveling along the x_3 direction with electric and magnetic fields in the x_1 and x_2 directions, respectively. The appropriate permittivity and permeability are thus ϵ_1 and μ_2 so that $k = \omega\sqrt{\epsilon_1\mu_2} = \omega\sqrt{s\epsilon_o s\mu_o} = sk_o$, corresponding to a refractive index $n_1 = s$. The phase shift accumulated over the distance d is therefore $sk_o d$, as expected, and the impedance is $\eta_1 = \sqrt{\mu_2/\epsilon_1} = \eta_o$.
– Wave 2 is also taken to travel along the x_3 direction but the electric and magnetic fields are now in the x_2 and $-x_1$ directions, respectively. This wave also travels with a refractive index $n_2 = s$ and has an impedance $\eta_2 = \eta_o$.
– Wave 3 travels along the x_2 direction with electric and magnetic fields in the x_3 and x_1 directions, respectively. The appropriate permittivity and permeability are ϵ_3 and μ_1 so that $k = \omega\sqrt{\epsilon_3\mu_1} = \omega\sqrt{s^{-1}\epsilon_o s\mu_o} = k_o$ corresponding to a refractive index $n_3 = 1$. The phase shift is $k_o d$, as expected, since there is no stretching in the x_2 direction. The impedance is $\eta_3 = \sqrt{\mu_1/\epsilon_3} = s\eta_o$.

Using these results, we conclude that the final design is a piecewise homogeneous medium with free space in the left half plane and an anisotropic uniaxial material in the right half plane [Fig. 8.4-1(d)]. The anisotropic material is birefringent with $n_1 = s$, $n_2 = s$, and $n_3 = 1$, but it introduces no reflection at the boundary with free space since the impedances are the same as that of free space: $\eta_1 = \eta_2 = \eta_o$.

Two factors distinguish refraction at the boundary of the synthesized anisotropic medium from conventional refraction at the boundary of a homogeneous and isotropic medium: (1) the refraction is not accompanied by reflection; and (2) the relationship between the angle of refraction and the angle of incidence, $s \tan\theta_2 = \tan\theta_1$, differs from Snell's law.

EXAMPLE 8.4-2. ***Refraction at Normal Incidence.*** We next consider the design of an optical material that implements refraction by an angle θ at a normal planar surface, as depicted in Fig. 8.4-2(a). This type of refraction cannot occur at the boundary between two isotropic dielectric materials, but can occur at the boundary between an isotropic and an anisotropic material, as described in Sec. 6.3E.

We begin with a pilot system of free space with ray trajectories along horizontal parallel straight lines [Fig. 8.4-2(b)], and implement the coordinate transformation

$$u_1 = x_1, \quad u_2 = x_2 + sx_3, \quad u_3 = x_3 \tag{8.4-6}$$

for $x_3 > 0$, with $s = \tan\theta$. This deflects the trajectories, as desired, by shearing along the x_2 direction [Fig. 8.4-1(c)]. The permittivity and permeability tensors of the equivalent anisotropic material corresponding to this coordinate transformation, as determined by use of (8.4-2) and (8.4-3), are:

$$\epsilon_o^{-1}\boldsymbol{\epsilon}^{\mathrm{e}} = \mu_o^{-1}\boldsymbol{\mu}^{\mathrm{e}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & s \\ 0 & s & 1+s^2 \end{bmatrix}. \tag{8.4-7}$$

This represents a homogeneous, but anisotropic, medium. When placed in the $x_3 > 0$ region, it introduces the desired refraction at normal incidence [Fig. 8.4-2(d)].

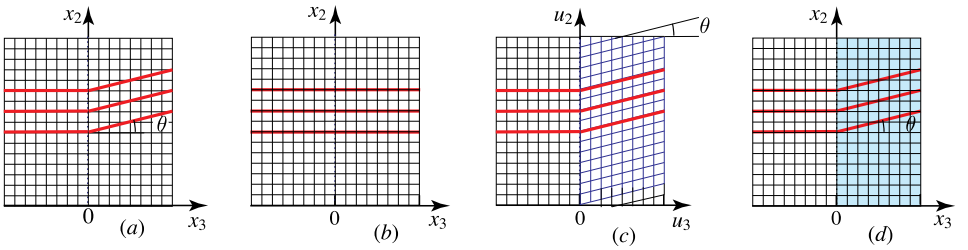


Figure 8.4-2 Geometrical transformation implementing refraction at normal incidence. (a) Desired optical trajectories. (b) Free space with horizontal straight-line optical trajectories. (c) Shearing the coordinate system along the x_2 direction for $x_3 > 0$ refracts the trajectories as desired. (d) Equivalent anisotropic material that exhibits identical refraction.

EXAMPLE 8.4-3. ***Cylindrical Focusing.*** In this case, parallel straight-line trajectories are to be refracted at a planar boundary such that they all meet at a common focal point at a distance f from the boundary, as shown in Fig. 8.4-3(a).

We begin with the straight trajectories shown in Fig. 8.4-3(b) in a Cartesian coordinate system and apply the coordinate transformation

$$u_1 = x_1, \quad u_2 = (f - x_3)\sin(x_2/f), \quad u_3 = f - (f - x_3)\cos(x_2/f), \tag{8.4-8}$$

for $x_3 > 0$. The result is a cylindrical coordinate system centered at $(u_2 = 0, u_3 = f)$, as shown in Fig. 8.4-3(c). This transformation converts a line $x_2 = a$ in the plane $x_1 = 0$ in the original coordinate system into a line $u_2 = (f - u_3)\tan(a/f)$ in the new coordinate system. It also converts a line $x_3 = b$ in the plane $x_1 = 0$ into a circle $u_2^2 + (f - u_3)^2 = (f - b)^2$ of radius $(f - b)$ centered

at the point $(u_2, u_3) = (0, f)$. Based on (8.4-1) and (8.4-2), the transformation yields the diagonal matrix

$$\epsilon_o^{-1}\boldsymbol{\epsilon}^{\mathrm{e}} = \mu_o^{-1}\boldsymbol{\mu}^{\mathrm{e}} = \begin{bmatrix} s & 0 & 0 \\ 0 & s^{-1} & 0 \\ 0 & 0 & s \end{bmatrix}, \qquad s = \frac{f}{|x_3 - f|}. \tag{8.4-9}$$

The permittivity and permeability tensors of the equivalent medium therefore have principal axes along the (x_1, x_2, x_3) axes, with principal values that are dependent on the position x_3, i.e., the equivalent material is graded along the x_3 direction with larger anisotropy near the focal line [Fig. 8.4-3(*d*)].

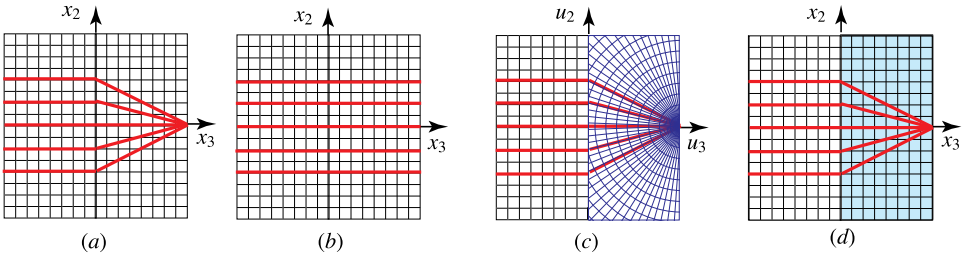


Figure 8.4-3 Geometrical transformation implementing cylindrical focusing. (*a*) Desired optical trajectories. (*b*) Free space with Cartesian coordinate system and parallel straight-line optical trajectories. (*c*) Conversion to a cylindrical coordinate system centered at $x_3 = f$ for $x_3 > 0$ produces the desired trajectories. (*d*) Equivalent anisotropic material with identical trajectories.

B. Invisibility Cloaks

An invisibility cloak is a device that guides light around an object such that the object appears transparent, and therefore invisible. For example, the trajectories shown in Fig. 8.4-4(*a*) avoid a sphere of radius a, emerging as if they had followed straight lines and passed right through it.

Following the prescribed design steps for transformation optics, we begin with the straight-line trajectories shown in Fig. 8.4-4(*b*) in a Cartesian-coordinate system (x_1, x_2, x_3). We next convert to a coordinate system (u_1, u_2, u_3) such that points of a sphere with radius $r = \sqrt{x_1^2 + x_2^2 + x_3^2}$ are mapped to points of a sphere of radius $u = \sqrt{u_1^2 + u_2^2 + u_3^2} > a$, thereby avoiding the sphere of radius a, as desired. This is accomplished for all points $r < b$, where $b > a$, via the linear relation

$$u = a + \frac{b-a}{b}\, r. \tag{8.4-10}$$

As r varies from 0 to b, u varies from a to b so that points of the sphere $0 < r < b$ in the original coordinate system are mapped into points in a spherical shell $a < u < b$ in the new coordinate system. This mapping may also be written as $u = s(r)r$, where

$$s(r) = \frac{a}{r} + \frac{b-a}{b} \tag{8.4-11}$$

is a position-dependent scaling factor.

When applied isotropically, this scaling produces the coordinate transformation

$$u_1 = s(r)\, x_1, \quad u_2 = s(r)\, x_2, \quad u_3 = s(r)\, x_3. \tag{8.4-12}$$

As can be shown by simple substitution, points on the straight line $x_2 = f$ in the plane $x_1 = 0$ are mapped into the curved trajectory

$$u_2^2 + u_3^2 = a^2 \left[\frac{u_2}{u_2 - (b-a)f/b} \right]^2 \tag{8.4-13}$$

in the u_2–u_3 plane. The red curved trajectories shown in Fig. 8.4-4 are computed from (8.4-13) for four values of f. The grid shown in Fig. 8.4-4(c) within the shell $a < u < b$ is computed by use of (8.4-13) and a similar equation determined by mapping the straight lines $x_3 = f$ in the plane $x_1 = 0$ into the u_2–u_3 plane. The parameters of the

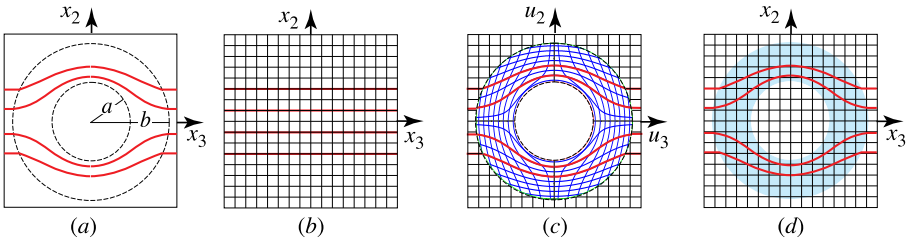


Figure 8.4-4 Geometrical transformation implementing cloaking of a sphere. (a) Desired optical trajectories. (b) Free space with Cartesian coordinate system and parallel straight-line optical trajectories. (c) Coordinate system transformation mapping points inside the sphere to points within a spherical shell outside the sphere, thereby producing the desired trajectories. (d) Equivalent anisotropic material with identical trajectories.

equivalent material to be placed in the $a < r < b$ spherical shell shown in Fig. 8.4-4(d) may be determined by use of (8.4-1) and (8.4-2) together with (8.4-12). The result is

$$\epsilon_o^{-1}\boldsymbol{\epsilon}^{\mathrm{e}} = \mu_o^{-1}\boldsymbol{\mu}^{\mathrm{e}} = \frac{b}{b-a}\frac{1}{v^2}\begin{bmatrix} v^2-u_1^2 & -u_1u_2 & -u_1u_3 \\ -u_2u_1 & v^2-u_2^2 & -u_2u_3 \\ -u_3u_1 & -u_3u_2 & v^2-u_3^2 \end{bmatrix}, \quad v^2 = \frac{u^4}{2au-a^2}. \tag{8.4-14}$$

Clearly, the dielectric and magnetic properties of the material in the spherical shell are both inhomogeneous and anisotropic. For example, at points on the x_1 axis $(u, 0, 0)$, we have

$$\epsilon_o^{-1}\boldsymbol{\epsilon}^{\mathrm{e}} = \mu_o^{-1}\boldsymbol{\mu}^{\mathrm{e}} = \frac{b}{b-a}\begin{bmatrix} 1-u^2/v^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{8.4-15}$$

At these points, the principal axes are aligned with the Cartesian coordinates (x_1, x_2, x_3). The principal value ϵ_1 varies from 0 to $\epsilon_o(b-a)/b$ as u varies from a to b, while the principal values ϵ_2 and ϵ_3 are fixed at the value $\epsilon_o b/(b-a)$. Similar results apply for μ.

At optical wavelengths, the implementation of invisibility cloaks via metamaterials requires the use of advanced nanofabrication technologies such as electron-beam or focused ion-beam lithography. The constituent dielectric and magnetic elements, which have various shapes and dimensions, must be intricately designed and precisely laid out. Since such elements are highly resonant, the electromagnetic properties of the metamaterial depend strongly on wavelength so that such devices typically operate only over narrow spectral bandwidths.

READING LIST

Single- and Double-Negative Media

S. A. Ramakrishna and T. M. Grzegorczyk, *Physics and Applications of Negative Refractive Index Materials*, CRC Press/Taylor & Francis, 2008.

J. B. Pendry and D. R. Smith, Reversing Light with Negative Refraction, *Physics Today*, vol. 57, no. 6, pp. 37–43, 2004.

R. A. Depine and A. Lakhtakia, A New Condition to Identify Isotropic Dielectric–Magnetic Materials Displaying Negative Phase Velocity, *Microwave and Optical Technology Letters*, vol. 41, pp. 315–316, 2004.

S. A. Darmanyan, M. Nevière, and A. A. Zakhidov, Surface Modes at the Interface of Conventional and Left-Handed Media, *Optics Communications*, vol. 225, pp. 233–240, 2003.

W. S. Weiglhofer and A. Lakhtakia, eds., *Introduction to Complex Mediums for Electromagnetics and Optics*, SPIE Optical Engineering Press, 2003.

M. W. McCall, A. Lakhtakia, and W. S. Weiglhofer, The Negative Index of Refraction Demystified, *European Journal of Physics*, vol. 23, pp. 353–359, 2002.

R. A. Shelby, D. R. Smith, and S. Schultz, Experimental Verification of a Negative Index of Refraction, *Science*, vol. 292, pp. 77–99, 2001.

J. B. Pendry, Negative Refraction Makes a Perfect Lens, *Physical Review Letters*, vol. 85, pp. 3966–3969, 2000.

V. G. Veselago, The Electrodynamics of Substances with Simultaneously Negative Values of ϵ and μ, *Soviet Physics–Uspekhi*, vol. 10, pp. 509–514, 1968 [*Uspekhi Fizicheskikh Nauk*, vol. 92, pp. 517–526, 1964].

Plasmonics

M. I. Stockman *et al.*, Roadmap on Plasmonics, *Journal of Optics*, vol. 20, 043001, 2018.

A. Trügler, *Optical Properties of Metallic Nanoparticles: Basic Principles and Simulation*, Springer-Verlag, 2016.

I. D. Mayergoyz, *Plasmon Resonances in Nanoparticles*, World Scientific, 2013.

M. Pelton and G. Bryant, *Introduction to Metal–Nanoparticle Plasmonics*, Wiley, 2013.

A. V. Zayats and S. A. Maier, eds., *Active Plasmonics and Tuneable Plasmonic Metamaterials*, Wiley–Science Wise, 2013.

Z. Han and S. I. Bozhevolnyi, Radiation Guiding with Surface Plasmon Polaritons, *Reports on Progress in Physics*, vol. 76, 016402, 2013.

S. Enoch and N. Bonod, eds., *Plasmonics: From Basics to Advanced Topics*, Springer-Verlag, 2012.

S. Adachi, *The Handbook on Optical Constants of Metals: In Tables and Figures*, World Scientific, 2012.

S. Roh, T. Chung, and B. Lee, Overview of the Characteristics of Micro- and Nano-Structured Surface Plasmon Resonance Sensors, *Sensors*, vol. 11, pp. 1565–1588, 2011.

D. Sarid and W. Challener, *Modern Introduction to Surface Plasmons*, Cambridge, 2010.

P. Berini, Long-Range Surface Plasmon Polaritons, *Advances in Optics and Photonics*, vol. 1, pp. 484–588, 2009.

S. A. Maier, *Plasmonics: Fundamentals and Applications*, Springer-Verlag, 2007.

M. L. Brongersma and P. G. Kik, eds., *Surface Plasmon Nanophotonics*, Springer-Verlag, 2007.

H. A. Atwater, The Promise of Plasmonics, *Scientific American*, vol. 296, no. 4, pp. 56–63, 2007.

W. L. Barnes, A. Dereux, and T. W. Ebbesen, Surface Plasmon Subwavelength Optics, *Nature*, vol. 424, pp. 824–830, 2003.

E. J. Zeman and G. C. Schatz, An Accurate Electromagnetic Theory Study of Surface Enhancement Factors for Ag, Au, Cu, Li, Na, Al, Ga, In, Zn, and Cd, *Journal of Physical Chemistry*, vol. 91, pp. 634–643, 1987.

H. J. Hagemann, W. Gudat, and C. Kunz, Optical Constants from the Far Infrared to the X-Ray Region: Mg, Al, Cu, Ag, Au, Bi, C, and Al_2O_3, *Deutsches Elektronen-Synchrotron Report DESY-SR74-7*, May 1974.

P. B. Johnson and R. W. Christy, Optical Constants of the Noble Metals, *Physical Review B*, vol. 6, pp. 4370–4379, 1972.

Near-Field Optics and Nanophotonics

M. Bertolotti, C. Sibilia, and A. M. Guzmán, *Evanescent Waves in Optics: An Introduction to Plasmonics*, Springer-Verlag, 2017.

N. Rotenberg and L. Kuipers, Mapping Nanoscale Light Fields, *Nature Photonics*, vol. 8, pp. 919–926, 2014.

L. Novotny and B. Hecht, *Principles of Nano-Optics*, Cambridge University Press, 2nd ed. 2012.

Z. Zalevsky and I. Abdulhalim, *Integrated Nanophotonic Devices*, Elsevier, 2nd ed. 2014.

L. Novotny, From Near-Field Optics to Optical Antennas, *Physics Today*, vol. 64, no. 7, pp. 47–52, 2011.

S. V. Gaponenko, *Introduction to Nanophotonics*, Cambridge University Press, 2010.

M. Ohtsu, K. Kobayashi, T. Kawazoe, T. Yatsui, and M. Naruse, *Principles of Nanophotonics*, CRC Press/Taylor & Francis, 2008.

D. Courjon, *Near-Field Microscopy and Near-Field Optics*, World Scientific, 2003.

S. Jutamulia, ed., *Selected Papers on Near-Field Optics*, SPIE Optical Engineering Press (Milestone Series Volume 172), 2002.

S. Kawata, M. Ohtsu, and M. Irie, eds., *Near-Field Optics and Surface Plasmon Polaritons*, Springer-Verlag, 2001.

Metamaterials

I. Liberal, A. M. Mahmoud, Y. Li, B. Edwards, and N. Engheta, Photonic Doping of Epsilon-Near-Zero Media, *Science*, vol. 355, pp. 1058–1062, 2017.

I. Liberal and N. Engheta, Zero-Index Platforms: Where Light Defies Geometry, *Optics & Photonics News*, vol. 27, no. 7/8, pp. 26–33, 2016.

N. Engheta, 150 Years of Maxwell's Equations, *Science*, vol. 349, pp. 136–137, 2015.

N. M. Litchinitser and V. M. Shalaev, Metamaterials: State-of-the Art and Future Directions, in D. L. Andrews, ed., *Photonics: Scientific Foundations, Technology and Applications*, Volume II, *Nanophotonic Structures and Materials*, Wiley–Science Wise, 2015.

A. Poddubny, I. Iorsh, P. Belov, and Y. Kivshar, Hyperbolic Metamaterials, *Nature Photonics*, vol. 7, pp. 958–967, 2013.

R. Marqués, F. Martín, and M. Sorolla, *Metamaterials with Negative Parameters: Theory, Design and Microwave Applications*, Wiley, 2013.

O. Hess, J. B. Pendry, S. A. Maier, R. F. Oulton, J. M. Hamm, and K. L. Tsakmakidis, Active Nanoplasmonic Metamaterials, *Nature Materials*, vol. 11, pp. 573–584, 2012.

N. I. Zheludev and Y. S. Kivshar, From Metamaterials to Metadevices, *Nature Materials*, vol. 11, pp. 917–924, 2012.

Y. Liu and X. Zhang, Metamaterials: A New Frontier of Science and Technology, *Chemical Society Reviews*, vol. 40, pp. 2494–2507, 2011.

M. Wegener and S. Linden, Shaping Optical Space with Metamaterials, *Physics Today*, vol. 63, no. 10, pp. 32–36, 2010.

T. J. Cui, D. R. Smith, and R. Liu, eds., *Metamaterials: Theory, Design, and Applications*, Springer-Verlag, 2010.

W. Cai and V. Shalaev, *Optical Metamaterials: Fundamentals and Applications*, Springer-Verlag, 2009.

A. K. Sarychev and V. M. Shalaev, *Electrodynamics of Metamaterials*, World Scientific, 2007.

N. Engheta, Circuits with Light at Nanoscales: Optical Nanocircuits Inspired by Metamaterials, *Science*, vol. 317, pp. 1698–1702, 2007.

N. Engheta and R. W. Ziolkowski, eds., *Electromagnetic Metamaterials: Physics and Engineering Explorations*, Wiley–IEEE, 2006.

S. Linden, C. Enkrich, G. Dolling, M. W. Klein, J. Zhou, T. Koschny, C. M. Soukoulis, S. Burger, F. Schmidt, and M. Wegener, Photonic Metamaterials: Magnetism at Optical Frequencies, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 12, pp. 1097–1105, 2006.

G. V. Eleftheriades and K. G. Balmain, eds., *Negative-Refraction Metamaterials: Fundamental Principles and Applications*, Wiley–IEEE, 2005.

D. R. Smith, J. B. Pendry, and M. C. K. Wiltshire, Metamaterials and Negative Refractive Index, *Science*, vol. 305, 788–792, 2004.

Optical Antennas and Metasurfaces

N. Meinzer, W. L. Barnes, and I. R. Hooper, Plasmonic Meta-Atoms and Metasurfaces, *Nature Photonics*, vol. 8, pp. 889–898, 2014.

N. Yu and F. Capasso, Flat Optics with Designer Metasurfaces, *Nature Materials*, vol. 13, pp. 139–150, 2014.

M. Agio and A. Alù, eds., *Optical Antennas*, Cambridge University Press, 2013.

A. V. Kildishev, A. Boltasseva, and V. M. Shalaev, Planar Photonics with Metasurfaces, *Science*, vol. 339, 1232009, 2013.

A. A. Maradudin, *Structured Surfaces as Optical Metamaterials*, Cambridge University Press, 2011.

N. Yu, P. Genevet, M. A. Kats, F. Aieta, J.-P. Tetienne, F. Capasso, and Z. Gaburro, Light Propagation with Phase Discontinuities: Generalized Laws of Reflection and Refraction, *Science*, vol. 334, pp. 333–337, 2011.

R. Gordon, A. G. Brolo, D. Sinton, and K. L. Kavanagh, Resonant Optical Transmission Through Hole-Arrays in Metal Films: Physics and Applications, *Laser & Photonics Reviews*, vol. 4, pp. 311–335, 2010.

P. Bharadwaj, B. Deutsch, and L. Novotny, Optical Antennas, *Advances in Optics and Photonics*, vol. 1, pp. 438–483, 2009.

J. Bravo-Abad, L. Martín-Moreno, and F. J. Garcia-Vidal, Resonant Transmission of Light Through Subwavelength Holes in Thick Metal Films, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 12, pp. 1221–1227, 2006.

F. J. Garcia-Vidal, L. Martín-Moreno, and J. B. Pendry, Surfaces with Holes in Them: New Plasmonic Metamaterials, *Journal of Optics A: Pure and Applied Optics*, vol. 7, pp. S97–S101, 2005.

T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio, and P. A. Wolff, Extraordinary Optical Transmission Through Sub-Wavelength Hole Arrays, *Nature*, vol. 391, pp. 667–669, 1998.

Transformation Optics

J. B. Pendry, Y. Luo, and R. Zhao, Transforming the Optical Landscape, *Science*, vol. 348, pp. 521–524, 2015.

D. H. Werner and D.-H. Kwon, *Transformation Electromagnetics and Metamaterials: Fundamental Principles and Applications*, Springer-Verlag, 2014.

A. B. Shvartsburg and A. A. Maradudin, *Waves in Gradient Metamaterials*, World Scientific/Imperial College Press, 2013.

M. McCall, Transformation Optics and Cloaking, *Contemporary Physics*, vol. 54, pp. 273–286, 2013.

J. B. Pendry, A. Aubry, D. R. Smith, and S. A. Maier, Transformation Optics and Subwavelength Control of Light, *Science*, vol. 227, pp. 549–552, 2012.

U. Leonhardt and T. G. Philbin, Transformation Optics and the Geometry of Light, in E. Wolf, ed., *Progress in Optics*, Elsevier, 2009, vol. 53, pp. 69–152.

D. Schurig, J. B. Pendry, and D. R. Smith, Calculation of Material Properties and Ray Tracing in Transformation Media, *Optics Express*, vol. 14, pp. 9794–9804, 2006.

J. B. Pendry, D. Schurig, and D. R. Smith, Controlling Electromagnetic Fields, *Science*, vol. 312, pp. 1780–1782, 2006.

A. Alù and N. Engheta, Achieving Transparency with Plasmonic and Metamaterial Coatings, *Physical Review E*, vol. 72, 016623, 2005.

C. Gomez-Reino, M. V. Perez, and C. Bao, *Gradient-Index Optics: Fundamentals and Applications*, Springer-Verlag, 2002, paperback ed. 2010.

PROBLEMS

8.1-1 **SPP at Boundary Between DPS Medium and Lossy SNG Medium.** Consider a DPS-SNG boundary with $\mu_1 = \mu_2 = \mu_o$ and ϵ_1 real and positive, and with $\epsilon_2 = \epsilon_2' + j\epsilon_2''$ complex with ϵ_2' real and negative. Given that $|\epsilon_2''| \ll \epsilon_2'$, demonstrate that the plasmon wavelength λ_o/n_b and the propagation length d_b may be calculated with the help of the following approximate formulas:

$$n_\mathrm{b} \approx \sqrt{\frac{\epsilon_\mathrm{b}}{\epsilon_o}}\,, \qquad \epsilon_\mathrm{b} \approx \frac{\epsilon_1\epsilon_2'}{\epsilon_1+\epsilon_2'}\,, \qquad d_\mathrm{b} \approx \frac{\lambda_o}{2\pi}\frac{1}{n_\mathrm{b}^3}\frac{\epsilon_2'^2}{\epsilon_o\epsilon_2''}\,.$$

8.1-2 **NIM Slab as a Near-Field Imaging System.** Demonstrate the near-field imaging capability of a slab of refractive index $n = -2$ in air ($n = 1$) by constructing a graph similar to that displayed in Fig. 8.1-6(b), and by determining the imaging equation. Find the reflection and transmission coefficients for a wave normally incident on the boundaries of the slab. Sketch the amplitude of an evanescent wave transmitted through the slab by constructing a profile similar to that portrayed in Fig. 8.1-6(d).

8.1-3 **Subwavelength-Resolution Near-Field Imaging with a Lossy NIM Slab.** Consider a slab of refractive index $n = -2$ in air ($n = 1$). An evanescent wave entering the slab from air is amplified by the slab material, as demonstrated in Prob. 8.1-2. If the material is lossy, there will also be attenuation. If the attenuation coefficient $\gamma = 0.1k_o$, where $k_o = 2\pi/\lambda_o$ is the wavenumber in free space, determine the spatial angular frequency k_x (in units of k_o) at which the amplification will be smaller than the attenuation? What is the corresponding resolution in units of wavelength? Assume that $k_y = 0$.

*8.1-4 **Type-II Hyperbolic Medium.** For the hyperbolic medium described in Sec. 8.1C, ϵ_1 and ϵ_2 are positive, while ϵ_3 is negative. This is called a *Type-I hyperbolic medium*. Find the **k** surface for a *Type-II hyperbolic medium*, in which ϵ_1 and ϵ_2 are negative, while ϵ_3 is positive. Show that a Type-II hyperbolic medium also supports propagating waves with very short wavelengths, but can be highly reflective.

8.2-1 **Group Velocity in a Metal.** For a medium described by the simplified Drude model with effective permittivity (8.2-18), show that the product of the phase velocity and the group velocity is c_o^2.

8.2-2 **Prism Coupler for Exciting a SPP Wave.** A SPP wave is to be generated at the interface between air and a 60-nm-thick layer of Ag by making use of a SiN prism (refractive index $n = 2.0$) that abuts the opposite face of the Ag layer (see Fig. 8.2-5). At a free-space wavelength $\lambda_o = 700$ nm, the relative permittivity of Ag is $-20 + j\,1.3$. Calculate the angle of incidence of the light inside the prism required to couple to the Ag-air SPP. Assume that the dispersion relation of the SPP wave at the Ag-air interface is undisturbed by the presence of the prism.

8.2-3 **Silver Nanosphere in Glass.** Consider a silver nanosphere of radius $a = 10$ nm embedded in glass ($n = 1.45$). Calculate and plot the scattering efficiency Q_s, the absorption efficiency Q_a, and the normalized internal field E_i/E_0 as functions of the free-space wavelength λ_o over the 350–1000-nm wavelength range. Identify the resonance frequency and determine the scattering and absorption coefficients, α_s and α_a, respectively, at resonance. Use the fitted modified-Drude relative-permittivity function for bulk Ag:

$$\frac{\epsilon_\mathrm{s}}{\epsilon_o} = 1 + \frac{\omega_p^2}{-\omega^2 + j\omega\zeta} + \frac{2.2\,\omega_L^2}{\omega_L^2 - \omega^2 + j\omega\zeta_L},$$

where $\zeta = 0.00229\,\omega_p$, $\omega_L = 0.575\,\omega_p$, and $\zeta_L = 0.124\,\omega_p$. The quantity ω_p corresponds to a free-space plasma wavelength $\lambda_p = 135.2$ nm. This model incorporates an interband absorption (bound-electron) contribution to the complex permittivity.

8.3-1 **Negative-Permittivity Metamaterial: Silver Nanospheres in Water.** Use the Maxwell-Garnett mixing rule (5.6-20) to calculate and plot the real and imaginary parts of the effective permittivity ϵ_e of water with inclusions of silver nanospheres [see Fig. 8.3-2(b)], as a function of the free-space wavelength λ_o, over the wavelength range from 250 to 1000 nm. Use the relative-permittivity function of Ag defined in Prob. 8.2-3 and take the refractive index of water to be constant at $n = 1.33$. Assume that the volume fraction $\mathrm{f} = 3\%$. Identify wavelength ranges within which the real part of ϵ_e is negative. Investigate the effects of changing both f and n.

*8.3-2 **Layered Metal–Dielectric Hyperbolic Metamaterial.** Planar layers of metal are alternately stacked with layers of dielectric material. The anisotropic medium has an effective dielectric tensor with components $\epsilon_1 = \epsilon_2$ in the plane of the layers and ϵ_3 in the orthogonal direction. Use an effective-medium approximation,

$$\epsilon_1 = \epsilon_2 = \frac{\epsilon_\mathrm{m} d_\mathrm{m} + \epsilon_\mathrm{d} d_\mathrm{d}}{d_\mathrm{m} + d_\mathrm{d}}, \qquad 1/\epsilon_3 = \frac{d_\mathrm{m}/\epsilon_\mathrm{m} + d_\mathrm{d}/\epsilon_\mathrm{d}}{d_\mathrm{m} + d_\mathrm{d}},$$

where ϵ_m and ϵ_d are the permittivities of the metal and dielectric layers, respectively, and d_m and d_d are their thicknesses. With the help of the simplified Drude-model expression for ϵ_m, show that this structure can behave as a hyperbolic material. Identify which of the components, ϵ_1 or ϵ_3, is negative. Sketch the **k** surface.

CHAPTER

9

GUIDED-WAVE OPTICS

9.1 PLANAR-MIRROR WAVEGUIDES 355

9.2 PLANAR DIELECTRIC WAVEGUIDES 363

A. Waveguide Modes

B. Field Distributions

C. Dispersion Relation and Group Velocities

9.3 TWO-DIMENSIONAL WAVEGUIDES 372

9.4 OPTICAL COUPLING IN WAVEGUIDES 376

A. Input Couplers

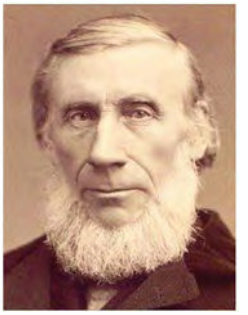
B. Coupled Waveguides

*C. Waveguide Arrays

9.5 PHOTONIC-CRYSTAL WAVEGUIDES 385

9.6 PLASMONIC WAVEGUIDES 386

Jean-Daniel Colladon (1802–1893)



John Tyndall (1820–1893)

Total internal reflection, the basis of guided-wave optics, was first demonstrated in water jets in the mid-1800s by the Swiss physicist Jean-Daniel Colladon and by the Irish-born physicist John Tyndall.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Traditional optical instruments and systems make use of bulk optics, in which light is transmitted between different locations in the form of beams that are collimated, relayed, focused, and scanned by mirrors, lenses, and prisms. The beams diffract and broaden as they propagate although they can be refocused by the use of lenses and mirrors. However, the bulk optical components that comprise such systems are often large and unwieldy, and objects in the paths of the beams can obstruct or scatter them.

In many circumstances it is advantageous to transmit optical beams through dielectric conduits rather than through free space. The technology for achieving this is known as **guided-wave optics**. It was initially developed to provide long-distance light transmission without the necessity of using relay lenses. This technology now has many important applications. A few examples are: carrying light over long distances for optical fiber communications, biomedical imaging where light must reach awkward locations, and connecting components within miniaturized optical and optoelectronic devices and systems.

The underlying principle of optical confinement is straightforward. A medium of refractive index n_1, embedded in a medium of lower refractive index $n_2 < n_1$, acts as a light "trap" within which optical rays remain confined by multiple total internal reflections at the boundaries. Because this effect facilitates the confinement of light generated inside a medium of high refractive index [see Exercise (1.2-6)], it can be exploited in making light conduits — guides that transport light from one location to another. An **optical waveguide** is a light conduit consisting of a slab, strip, or cylinder of dielectric material embedded in another dielectric material of lower refractive index (Fig. 9.0-1). The light is transported through the inner medium without radiating into the surrounding medium. The most widely used of these waveguides is the optical fiber, comprising two concentric cylinders of low-loss dielectric material such as glass (see Chapter 10). Other forms of optical waveguides make use of photonic crystals (Chapter 7) and metal–dielectric structures (Chapter 8).

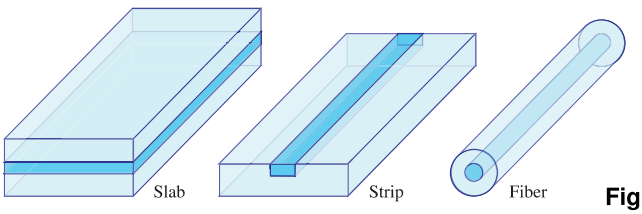


Figure 9.0-1 Optical waveguides.

Integrated photonics, also known as **integrated optics**, is the technology of combining, on a single substrate ("**chip**"), various optical devices and components useful for generating, focusing, splitting, combining, isolating, polarizing, coupling, switching, modulating, and detecting light. Optical waveguides provide the links among these components. Such chips, called **photonic integrated circuits (PICs)**, are optical versions of electronic integrated circuits. An example of a transceiver (transmitter/receiver) chip is schematized in Fig. 9.0-2. Integrated photonics serves to miniaturize optics in much the same way that silicon integrated circuits have miniaturized electronics.

This Chapter

The basic theory of optical waveguides is presented in this and the following chapter. In this chapter, we consider rectangular waveguides, which are used extensively in integrated photonics. Chapter 10 deals with cylindrical waveguides, i.e., optical fibers.

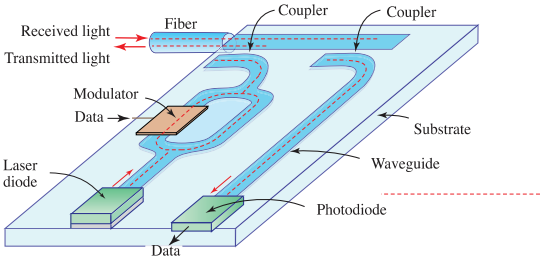


Figure 9.0-2 Schematic of a photonic integrated circuit that serves as an elementary optical transceiver (transmitter/receiver). Received light enters via a waveguide and is directed to a photodiode where it is detected. Light from a laser diode is guided, modulated, and coupled into a fiber for transmission.

If reflectors are placed at the two ends of a short waveguide, the result is a structure that traps and stores light — an optical resonator. These devices, which are essential to the operation of lasers, are described in Chapter 11. Various integrated-photonic components and devices (such as laser diodes, detectors, modulators, interconnects, and switches) are considered in the chapters that deal specifically with those components and devices. Photonic integrated circuits based on silicon photonics are addressed in Sec. 25.1E. Optical fiber communication systems are discussed in detail in Chapter 25.

9.1 PLANAR-MIRROR WAVEGUIDES

We begin by examining wave propagation in a waveguide comprising two parallel infinite planar *mirrors* separated by a distance d (Fig. 9.1-1). The mirrors are assumed to reflect light without loss. A ray of light, say in the y–z plane, making an angle θ with the mirrors reflects and bounces between them without loss of energy. The ray is thus guided along the z direction.

This waveguide appears to provide a perfect conduit for light rays. It is *not* used in practical applications, however, principally because of the difficulty and cost of fabricating low-loss mirrors. Nevertheless, we study this simple example in detail because it provides a valuable pedagogical introduction to the dielectric waveguide, which we examine in Sec. 9.2, and to the optical resonator, which is the subject of Chapter 11.

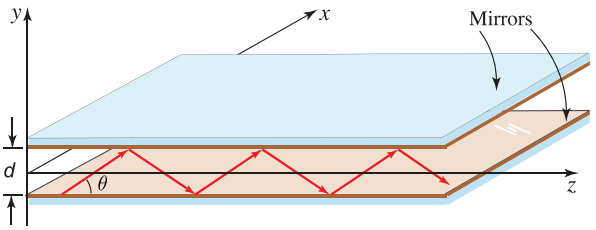


Figure 9.1-1 Planar-mirror waveguide.

Waveguide Modes

The ray-optics picture of light being guided by multiple reflections cannot explain a number of important effects that require the use of electromagnetic theory. A simple approach for carrying out an electromagnetic analysis is to associate with each optical ray a transverse electromagnetic (TEM) plane wave. The total electromagnetic field is then the sum of these plane waves.

Consider a monochromatic TEM plane wave of wavelength $\lambda = \lambda_o/n$, wavenumber $k = nk_o$, and phase velocity $c = c_o/n$, where n is the refractive index of the medium between the mirrors. The wave is polarized in the x direction and its wavevector lies in

the y–z plane at an angle θ with the z axis (Fig. 9.1-1). Like the optical ray, the wave reflects from the upper mirror, travels at an angle $-\theta$, reflects from the lower mirror, and travels once more at an angle θ, and so on. Since the electric-field vector is parallel to the mirror, each reflection is accompanied by a phase shift π for a perfect mirror, but the amplitude and polarization are not changed. The π phase shift ensures that the sum of each wave and its own reflection vanishes so that the total field is zero at the mirrors. At each point within the waveguide we have TEM waves traveling upward at an angle θ and others traveling downward at an angle $-\theta$; all waves are polarized in the x direction.

We now impose a self-consistency condition by requiring that as the wave reflects twice, it reproduces itself [see Fig. 9.1-2(a)], so that we have only two distinct plane waves. Fields that satisfy this condition are called the modes (or eigenfunctions) of the waveguide (see Appendix C). *Modes are fields that maintain the same transverse distribution and polarization at all locations along the waveguide axis*. We shall see that self-consistency guarantees this shape invariance. In connection with Fig. 9.1-2, the phase shift $\Delta\varphi$ encountered by the original wave in traveling from A to B must be equal to, or differ by an integer multiple of 2π, from that encountered when the wave reflects, travels from A to C, and reflects once more. Accounting for a phase shift of π at each reflection, we have $\Delta\varphi = 2\pi\overline{AC}/\lambda - 2\pi - 2\pi\overline{AB}/\lambda = 2\pi q$, where $q = 0, 1, 2, \ldots$, so that $2\pi(\overline{AC} - \overline{AB})/\lambda = 2\pi(q+1)$. The geometry portrayed in Fig. 9.1-2(a), together with the identity $\cos(2x) = 1 - 2\sin^2 x$, provides $\overline{AC} - \overline{AB} = 2d\sin\theta$, where d is the distance between the mirrors. Thus, $2\pi(2d\sin\theta)/\lambda = 2\pi(q+1)$ so that

$$\frac{2\pi}{\lambda}\, 2d\sin\theta = 2\pi\, m, \qquad m = 1, 2, \ldots. \tag{9.1-1}$$

where $m = q + 1$. The self-consistency condition is therefore satisfied only for certain bounce angles $\theta = \theta_m$ satisfying

$$\boxed{\sin\theta_m = m\,\frac{\lambda}{2d}, \qquad m = 1, 2, \ldots.} \tag{9.1-2}$$

Bounce Angles

Each integer m corresponds to a bounce angle θ_m, and the corresponding field is called the mth mode. The $m = 1$ mode has the smallest angle, $\theta_1 = \sin^{-1}(\lambda/2d)$; modes with larger m are composed of more oblique plane-wave components.

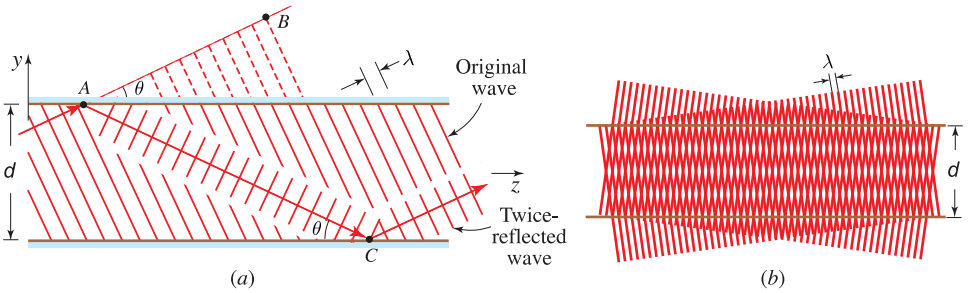


Figure 9.1-2 (a) Condition of self-consistency: as a wave reflects twice it duplicates itself. (b) At angles for which self-consistency is satisfied, the two waves interfere and create a pattern that does not change with z.

When the self-consistency condition is satisfied, the phases of the upward and downward plane waves at points on the z axis differ by half the round-trip phase shift $q\pi$, $q = 0, 1, \ldots$, or $(m-1)\pi$, $m = 1, 2, \ldots$, so that they add for odd m and subtract for even m.

Since the y component of the propagation constant is given by $k_y = nk_o \sin\theta$, it is quantized to the values $k_{ym} = nk_o \sin\theta_m = (2\pi/\lambda)\sin\theta_m$. Using (9.1-2), we obtain

$$\boxed{k_{ym} = m\frac{\pi}{d},} \qquad m = 1, 2, 3 \ldots, \tag{9.1-3}$$

Wavevector Transverse Component

so that the k_{ym} are spaced by π/d. Equation (9.1-3) states that the phase shift encountered when a wave travels a distance $2d$ (one round trip) in the y direction, with propagation constant k_{ym}, must be a multiple of 2π.

Propagation Constants

A guided wave is composed of two distinct plane waves traveling in the y–z plane at angles $\pm\theta$ with the z axis. Their wavevectors have components $(0, k_y, k_z)$ and $(0, -k_y, k_z)$. Their sum or difference therefore varies with z as $\exp(-jk_z z)$, so that the propagation constant of the guided wave is $\beta \equiv k_z = k\cos\theta$. Thus, β is quantized to the values $\beta_m = k\cos\theta_m$, from which $\beta_m^2 = k^2(1 - \sin^2\theta_m)$. Using (9.1-2), we obtain

$$\boxed{\beta_m^2 = k^2 - \frac{m^2\pi^2}{d^2}.} \tag{9.1-4}$$

Propagation Constants

Higher-order (more oblique) modes travel with smaller propagation constants. The values of θ_m, k_{ym}, and β_m for the different modes are illustrated in Fig. 9.1-3.

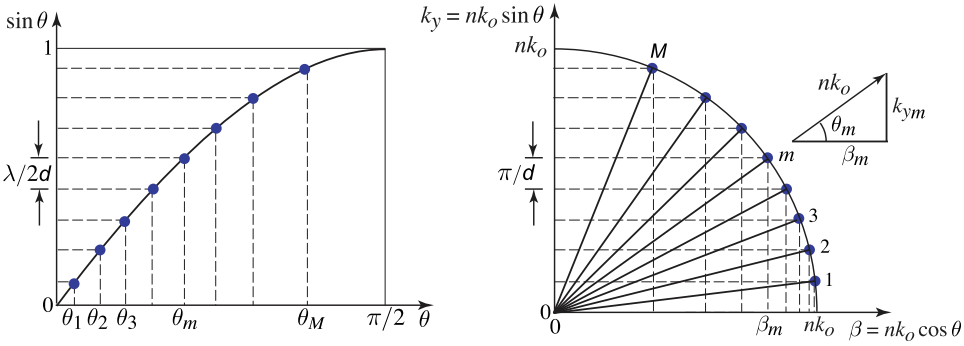


Figure 9.1-3 The bounce angles θ_m and the wavevector components of the modes of a planar-mirror waveguide (indicated by dots). The transverse components $k_{ym} = k\sin\theta_m$ are spaced uniformly at multiples of π/d, but the bounce angles θ_m and the propagation constants β_m are not equally spaced. Mode $m = 1$ has the smallest bounce angle and the largest propagation constant.

Field Distributions

The complex amplitude of the total field in the waveguide is the superposition of the two bouncing TEM plane waves. If $A_m \exp(-jk_{ym}y - j\beta_m z)$ is the upward wave, then $e^{j(m-1)\pi}A_m \exp(+jk_{ym}y - j\beta_m z)$ must be the downward wave [at $y = 0$, the two waves differ by a phase shift $(m-1)\pi$]. There are therefore symmetric modes, for which the two plane-wave components are added, and anti-symmetric modes, for which they are subtracted. The total field turns out to be $E_x(y,z) = 2A_m \cos(k_{ym}y)\exp(-j\beta_m z)$ for odd modes and $2jA_m \sin(k_{ym}y)\exp(-j\beta_m z)$ for even modes.

Using (9.1-3) we write the complex amplitude of the electric field in the form

$$E_x(y,z) = \mathrm{a}_m u_m(y)\exp(-j\beta_m z), \tag{9.1-5}$$

where

$$u_m(y) = \begin{cases} \sqrt{\dfrac{2}{d}}\cos\left(m\pi\dfrac{y}{d}\right), & m = 1,3,5,\ldots \\ \sqrt{\dfrac{2}{d}}\sin\left(m\pi\dfrac{y}{d}\right), & m = 2,4,6,\ldots, \end{cases} \tag{9.1-6}$$

with $\mathrm{a}_m = \sqrt{2d}A_m$ and $j\sqrt{2d}A_m$, for odd m and even m, respectively. The functions $u_m(y)$ have been normalized to satisfy

$$\int_{-d/2}^{d/2} u_m^2(y)\,dy = 1. \tag{9.1-7}$$

Thus, a_m is the amplitude of mode m. It can be shown that the functions $u_m(y)$ also satisfy

$$\int_{-d/2}^{d/2} u_m(y)\,u_l(y)\,dy = 0, \qquad l \neq m, \tag{9.1-8}$$

i.e., they are orthogonal in the $[-d/2, d/2]$ interval.

The transverse distributions $u_m(y)$ are plotted in Fig. 9.1-4. Each mode can be viewed as a standing wave in the y direction, traveling in the z direction. Modes of large m vary in the transverse plane at a greater rate k_y and travel with a smaller propagation constant β. The field vanishes at $y = \pm d/2$ for all modes, so that the boundary conditions at the surface of the mirrors are always satisfied.

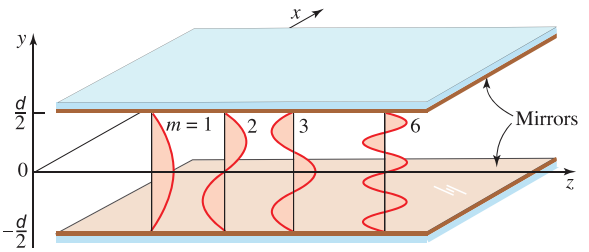


Figure 9.1-4 Field distributions of the modes of a planar-mirror waveguide.

Since we assumed at the outset that the bouncing TEM plane wave is polarized in the x direction, the total electric field is also in the x direction and the guided wave is a transverse-electric (TE) wave. Transverse magnetic (TM) waves are analyzed in a similar fashion as will be seen subsequently.

EXERCISE 9.1-1

Optical Power. Show that the optical power flow in the z direction associated with the TE mode $E_x(y,z) = \mathfrak{a}_m u_m(y) \exp(-j\beta_m z)$ is $(|\mathfrak{a}_m|^2/2\eta)\cos\theta_m$, where $\eta = \eta_o/n$ and $\eta_o = \sqrt{\mu_o/\epsilon_o}$ is the impedance of free space.

Number of Modes

Since $\sin\theta_m = m\lambda/2d$, $m = 1, 2, \ldots$, and taking $\sin\theta_m < 1$, the maximum allowed value of m is the greatest integer smaller than $1/(\lambda/2d)$,

$$M \doteq \frac{2d}{\lambda}. \tag{9.1-9}$$

Number of Modes

The symbol $\doteq$ denotes that $2d/\lambda$ is reduced to the nearest integer. As examples, when $2d/\lambda = 0.9$, 1, and 1.1, we have $M = 0$, 0, and 1, respectively. Thus, M is the number of modes of the waveguide. Light can be transmitted through the waveguide in one, two, or many modes. The actual number of modes that carry optical power depends on the source of excitation, but the maximum number is M.

The number of modes increases with increasing ratio of the mirror separation to the wavelength. Under conditions such that $2d/\lambda \leq 1$, corresponding to $d \leq \lambda/2$, M is seen to be 0, which indicates that the self-consistency condition cannot be met and the waveguide cannot support any modes. The wavelength $\lambda_c = 2d$ is called the **cutoff wavelength** of the waveguide. It is the longest wavelength that can be guided by the structure. It corresponds to the **cutoff frequency**

$$\nu_c = \frac{c}{2d}, \tag{9.1-10}$$

Cutoff Frequency

or the cutoff angular frequency $\omega_c = 2\pi\nu_c = \pi c/d$, the lowest frequency of light that can be guided by the waveguide. If $1 < 2d/\lambda \leq 2$ (i.e., $d \leq \lambda < 2d$ or $\nu_c \leq \nu < 2\nu_c$), only one mode is allowed. The structure is then said to be a **single-mode waveguide**. If $d = 5$ μm, for example, the waveguide has a cutoff wavelength $\lambda_c = 10$ μm; it supports a single mode for 5 μm $\leq \lambda < 10$ μm, and more modes for $\lambda < 5$ μm. Equation (9.1-9) can also be written in terms of the frequency ν, $M \doteq \nu/\nu_c = \omega/\omega_c$, so that the number of modes increases by unity when the angular frequency ω is incremented by ω_c, as illustrated in Fig. 9.1-5(a).

Dispersion Relation

The relation between the propagation constant β and the angular frequency ω is an important characteristic of the waveguide, known as the **dispersion relation**. For a homogeneous medium, the dispersion relation is simply $\omega = c\beta$. For mode m of a planar-mirror waveguide, β_m and ω are related by (9.1-4) so that

$$\beta_m^2 = (\omega/c)^2 - m^2\pi^2/d^2. \tag{9.1-11}$$

This relation may be written in terms of the cutoff angular frequency $\omega_c = 2\pi\nu_c = \pi c/d$ as

$$\beta_m = \frac{\omega}{c}\sqrt{1 - m^2\frac{\omega_c^2}{\omega^2}}\,. \tag{9.1-12}$$

Dispersion Relation

As shown in Fig. 9.1-5(b) for $m = 1, 2, \ldots$, the propagation constant β for mode m is zero at angular frequency $\omega = m\omega_c$, increases monotonically with angular frequency, and ultimately approaches the linear relation $\beta = \omega/c$ for sufficiently large values of β.

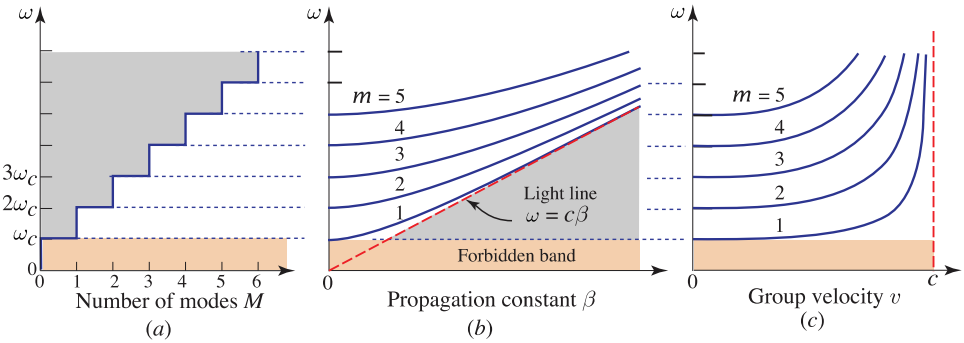


Figure 9.1-5 (a) Number of modes M as a function of angular frequency ω. Modes are not permitted for angular frequencies below the cutoff, $\omega_c = \pi c/d$. M increments by unity as ω increases by ω_c. (b) Dispersion relation. A forbidden band exists for angular frequencies below ω_c. (c) Group velocities of the modes as a function of angular frequency.

Group Velocities

In a medium with a given ω-β dispersion relation, a pulse of light (wavepacket) that has an angular frequency centered at ω travels with a velocity $v = d\omega/d\beta$, known as the group velocity (see Sec. 5.7). Taking the derivative of (9.1-12) and assuming that c is independent of ω (i.e., ignoring dispersion in the waveguide material), we obtain $2\beta_m\, d\beta_m/d\omega = 2\omega/c^2$, so that $d\omega/d\beta_m = c^2\beta_m/\omega = c^2 k\cos\theta_m/\omega = c\cos\theta_m$, from which the group velocity of mode m is

$$v_m = c\cos\theta_m = c\sqrt{1 - m^2\frac{\omega_c^2}{\omega^2}}\,. \tag{9.1-13}$$

Group Velocity

It follows that more oblique modes travel with smaller group velocities since they are delayed by the longer paths of the zigzagging process. The dependence of the group velocity on angular frequency is illustrated in Fig. 9.1-5(c), which shows that for each mode, the group velocity increases monotonically from 0 to c as the angular frequency increases above the mode cutoff frequency.

Equation (9.1-13) may also be obtained geometrically by examining the plane wave as it bounces between the mirrors and determining the distance advanced in the z

direction and the time taken by the zigzagging process. For the trip from the bottom mirror to the top mirror (Fig. 9.1-6) we have

$$v = \frac{\text{distance}}{\text{time}} = \frac{d\cot\theta}{d\csc\theta/c} = c\cos\theta. \tag{9.1-14}$$

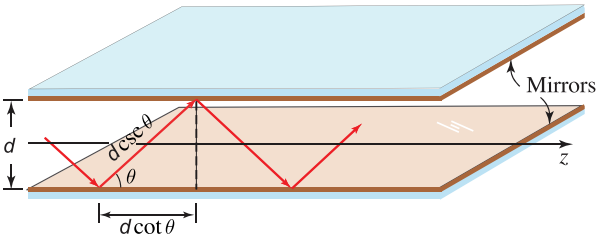


Figure 9.1-6 A plane wave bouncing at an angle θ advances in the z direction by a distance $d\cot\theta$ in a time $d\csc\theta/c$. The velocity is $c\cos\theta$.

TM Modes

Only TE modes (electric field in the x direction) have been considered to this point. TM modes (magnetic field in the x direction) can also be supported by the mirror waveguide. They can be studied by means of a TEM plane wave with the magnetic field in the x direction, traveling at an angle θ and reflecting from the two mirrors (Fig. 9.1-7). The electric-field complex amplitude then has components in the y and z directions. Since the z component is parallel to the mirror, it behaves precisely like the x component of the TE mode (i.e., it undergoes a phase shift π at each reflection and vanishes at the mirrors). When the self-consistency condition is applied to this component the result is mathematically identical to that of the TE case. The angles θ, the transverse wavevector components k_y, and the propagation constants β of the TM modes associated with this component are identical to those of the TE modes. There are $M \doteq 2d/\lambda$ TM modes (and a total of $2M$ modes) supported by the waveguide.

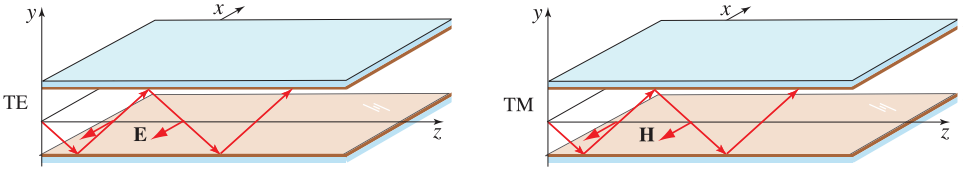


Figure 9.1-7 TE and TM polarized guided waves.

The z component of the electric-field complex amplitude of mode m, as previously, is the sum of an upward plane wave $A_m\exp(-jk_{ym}y)\exp(-j\beta_m z)$ and a downward plane wave $e^{j(m-1)\pi}A_m\exp(jk_{ym}y)\exp(-j\beta_m z)$, with equal amplitudes and phase shift $(m-1)\pi$, so that

$$E_z(y,z) = \begin{cases} a_m\sqrt{\dfrac{2}{d}}\cos\left(m\pi\dfrac{y}{d}\right)\exp(-j\beta_m z), & m = 1,3,5,\ldots \\ a_m\sqrt{\dfrac{2}{d}}\sin\left(m\pi\dfrac{y}{d}\right)\exp(-j\beta_m z), & m = 2,4,6,\ldots, \end{cases} \tag{9.1-15}$$

where $a_m = \sqrt{2d}A_m$ and $j\sqrt{2d}A_m$ for odd and even m, respectively. Since the electric-field vector of a TEM plane wave is normal to its direction of propagation,

it makes an angle $\pi/2 + \theta_m$ with the z axis for the upward wave, and $\pi/2 - \theta_m$ for the downward wave.

The y components of the electric field of these waves are

$$A_m \cot\theta_m \exp(-jk_{ym}y)\exp(-j\beta_m z) \text{ and } e^{jm\pi}A_m \cot\theta_m \exp(jk_{ym}y)\exp(-j\beta_m z), \tag{9.1-16}$$

so that

$$E_y(y,z) = \begin{cases} a_m\sqrt{\dfrac{2}{d}}\cot\theta_m \cos\left(m\pi\dfrac{y}{d}\right)\exp(-j\beta_m z), & m = 1,3,5,\ldots \\ a_m\sqrt{\dfrac{2}{d}}\cot\theta_m \sin\left(m\pi\dfrac{y}{d}\right)\exp(-j\beta_m z), & m = 2,4,6,\ldots . \end{cases} \tag{9.1-17}$$

Satisfaction of the boundary conditions is assured because $E_z(y,z)$ vanishes at the mirrors. The magnetic field component $H_x(y,z)$ may be similarly determined by noting that the ratio of the electric to the magnetic fields of a TEM wave is the impedance of the medium η. The resultant fields $E_y(y,z)$, $E_z(y,z)$, and $H_x(y,z)$ do, of course, satisfy Maxwell's equations.

Multimode Fields

For light to be guided by the mirrors, it is not necessary that it have the distribution of one of the modes. In fact, a field satisfying the boundary conditions (vanishing at the mirrors) but otherwise having an arbitrary distribution in the transverse plane *can* be guided by the waveguide. The optical power, however, is then divided among the modes. Since different modes travel with different propagation constants and different group velocities, the transverse distribution of the field will alter as it travels through the waveguide. Fig. 9.1-8 illustrates how the transverse distribution of a single mode is invariant to propagation, whereas the multimode distribution varies with z (the illustration is for the *intensity* distribution).

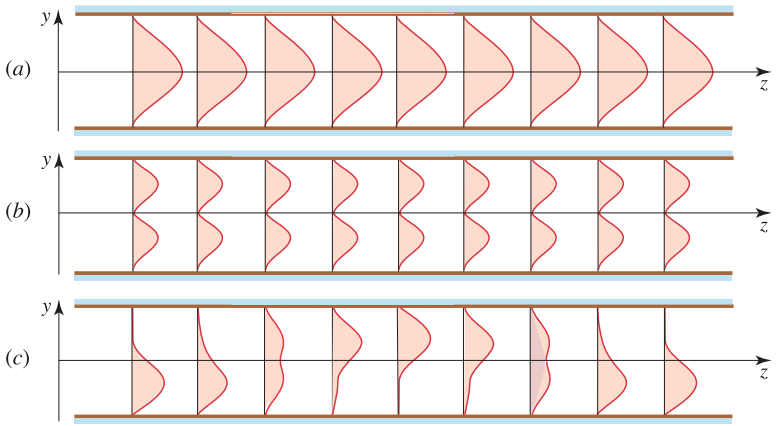


Figure 9.1-8 Variation of the intensity distribution in the transverse direction y at different axial distances z. (a) The electric-field complex amplitude in mode 1 is $E(y,z) = u_1(y)\exp(-j\beta_1 z)$, where $u_1(y) = \sqrt{2/d}\cos(\pi y/d)$. The intensity does not vary with z. (b) The complex amplitude in mode 2 is $E(y,z) = u_2(y)\exp(-j\beta_2 z)$, where $u_2(y) = \sqrt{2/d}\sin(2\pi y/d)$. The intensity does not vary with z. (c) The complex amplitude in a mixture of modes 1 and 2, $E(y,z) = u_1(y)\exp(-j\beta_1 z) + u_2(y)\exp(-j\beta_2 z)$. Since $\beta_1 \neq \beta_2$, the intensity distribution changes with z.

An arbitrary field polarized in the x direction and satisfying the boundary conditions can be written as a weighted superposition of the TE modes,

$$E_x(y,z) = \sum_{m=0}^{M} \mathcal{A}_m u_m(y) \exp(-j\beta_m z), \tag{9.1-18}$$

where $\mathcal{A}_m$, the superposition weights, are the amplitudes of the different modes.

EXERCISE 9.1-2

Optical Power in a Multimode Field. Show that the optical power flow in the z direction associated with the multimode field in (9.1-18) is the sum of the powers $(|\mathcal{A}_m|^2/2\eta)\cos\theta_m$ carried by each of the modes.

9.2 PLANAR DIELECTRIC WAVEGUIDES

A planar dielectric waveguide is a slab of dielectric material surrounded by media of lower refractive indices. The light is guided inside the slab by total internal reflection. In thin-film devices the slab is called the "film" and the upper and lower media are called the "cover" and the "substrate," respectively. The inner medium and outer media may also be called the "core" and the "cladding" of the waveguide, respectively. In this section we study the propagation of light in a symmetric planar dielectric waveguide made of a slab of width d and refractive index n_1 surrounded by a cladding of smaller refractive index n_2, as illustrated in Fig. 9.2-1. All materials are assumed to be lossless.

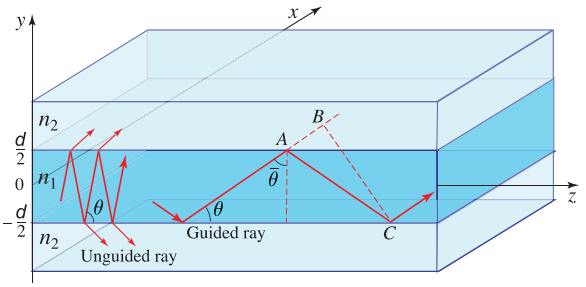


Figure 9.2-1 Planar dielectric (slab) waveguide. Rays making an angle $\theta < \bar{\theta}_c = \cos^{-1}(n_2/n_1)$ are guided by total internal reflection.

Light rays making angles θ with the z axis, in the y–z plane, undergo multiple total internal reflections at the slab boundaries, provided that θ is smaller than the complement of the critical angle $\bar{\theta}_c = \pi/2 - \sin^{-1}(n_2/n_1) = \cos^{-1}(n_2/n_1)$ [see (1.2-5) and Figs. 6.2-3 and 6.2-5]. They travel in the z direction by bouncing between the slab surfaces without loss of power. Rays making larger angles refract, losing a portion of their power at each reflection, and eventually vanish.

To determine the waveguide modes, a formal approach may be pursued by developing solutions to Maxwell's equations in the inner and outer media with the appropriate boundary conditions imposed (see Prob. 9.2-6). We shall instead write the solution in terms of TEM plane waves bouncing between the surfaces of the slab. By imposing the

self-consistency condition, we determine the bounce angles of the waveguide modes from which the propagation constants, field distributions, and group velocities are determined. The analysis is analogous to that used in the previous section for the planar-mirror waveguide.

A. Waveguide Modes

Assume that the field in the slab is in the form of a monochromatic TEM plane wave of wavelength $\lambda = \lambda_o/n_1$ bouncing back and forth at an angle θ smaller than the complementary critical angle $\bar{\theta}_c$. The wave travels with a phase velocity $c_1 = c_o/n_1$, has a wavenumber $n_1 k_o$, and has wavevector components $k_x = 0$, $k_y = n_1 k_o \sin\theta$, and $k_z = n_1 k_o \cos\theta$. To determine the modes we impose the self-consistency condition that a wave reproduces itself after each round trip.

In one round trip, the twice-reflected wave lags behind the original wave by a distance $\overline{AC} - \overline{AB} = 2d \sin\theta$, as in Fig. 9.1-2. There is also a phase φ_r introduced by each internal reflection at the dielectric boundary (see Sec. 6.2). For self-consistency, the phase shift between the two waves must be zero or a multiple of 2π,

$$\frac{2\pi}{\lambda} 2d \sin\theta - 2\varphi_r = 2\pi m, \qquad m = 0, 1, 2, \ldots \tag{9.2-1}$$

or

$$2k_y d - 2\varphi_r = 2\pi m. \tag{9.2-2}$$

The only difference between this condition and the corresponding condition in the mirror waveguide, (9.1-1) and (9.1-3), is that the phase shift π introduced by the mirror is replaced here by the phase shift φ_r introduced at the dielectric boundary.

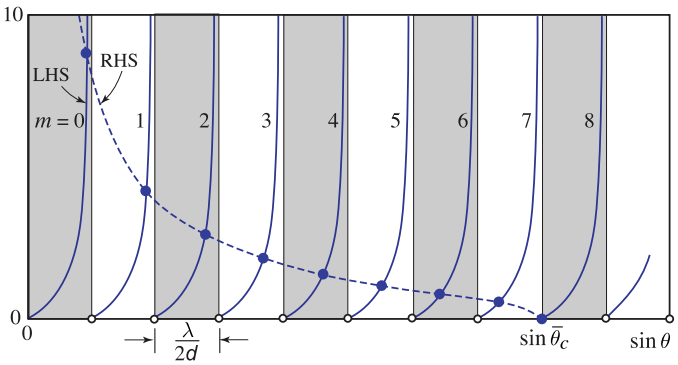


Figure 9.2-2 Graphical solution of (9.2-4) to determine the bounce angles θ_m of the modes of a planar dielectric waveguide. The right-hand side (RHS) and left-hand side (LHS) of (9.2-4) are plotted versus $\sin\theta$. The intersection points, marked by filled circles, determine $\sin\theta_m$. Each branch of the tan or cot function on the left-hand side corresponds to a mode. In this plot $\sin\bar{\theta}_c = 8(\lambda/2d)$ and the number of modes is $M = 9$. The open circles mark $\sin\theta_m = m\lambda/2d$, which provide the bounce angles of the modes of a planar-mirror waveguide of the same dimensions.

The reflection phase shift φ_r is a function of the angle θ. It also depends on the polarization of the incident wave, TE or TM. In the TE case (the electric field is in the

x direction), substituting $\theta_1 = \pi/2 - \theta$ and $\theta_c = \pi/2 - \bar{\theta}_c$ in (6.2-11) gives

$$\tan\frac{\varphi_r}{2} = \sqrt{\frac{\sin^2\bar{\theta}_c}{\sin^2\theta} - 1}\,, \tag{9.2-3}$$

so that φ_r varies from π to 0 as θ varies from 0 to $\bar{\theta}_c$. Rewriting (9.2-1) in the form $\tan(\pi d \sin\theta/\lambda - m\pi/2) = \tan(\varphi_r/2)$ and using (9.2-3), we obtain

$$\boxed{\tan\left(\pi\frac{d}{\lambda}\sin\theta - m\frac{\pi}{2}\right) = \sqrt{\frac{\sin^2\bar{\theta}_c}{\sin^2\theta} - 1}\,.} \tag{9.2-4}$$

Self-Consistency Condition (TE Modes)

This is a transcendental equation in one variable, $\sin\theta$. Its solutions yield the bounce angles θ_m of the modes. A graphic solution is instructive. The right- and left-hand sides of (9.2-4) are plotted in Fig. 9.2-2 as functions of $\sin\theta$. Solutions are given by the intersection points. The right-hand side, $\tan(\varphi_r/2)$, is a monotonic decreasing function of $\sin\theta$ that reaches 0 when $\sin\theta = \sin\bar{\theta}_c$. The left-hand side generates two families of curves, $\tan[(\pi d/\lambda)\sin\theta]$ and $\cot[(\pi d/\lambda)\sin\theta]$, when m is even and odd, respectively. The intersection points determine the angles θ_m of the modes. The bounce angles of the modes of a *mirror* waveguide of mirror separation d may be obtained from this diagram by using $\varphi_r = \pi$ or, equivalently, $\tan(\varphi_r/2) = \infty$. For comparison, these angles are marked by open circles.

The angles θ_m lie between 0 and $\bar{\theta}_c$. They correspond to wavevectors with components $(0,\ n_1k_o\sin\theta_m,\ n_1k_o\cos\theta_m)$. The z components are the propagation constants

$$\boxed{\beta_m = n_1k_o\cos\theta_m.} \tag{9.2-5}$$

Propagation Constants

Since $\cos\theta_m$ lies between 1 and $\cos\bar{\theta}_c = n_2/n_1$, β_m lies between n_2k_o and n_1k_o, as illustrated in Fig. 9.2-3.

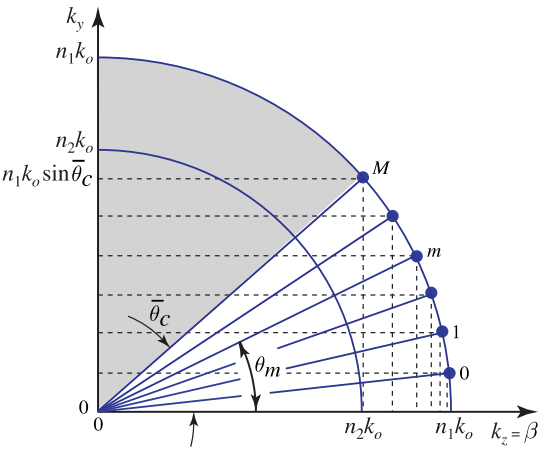


Figure 9.2-3 The bounce angles θ_m and the corresponding components k_z and k_y of the wavevector of the waveguide modes are indicated by dots. The angles θ_m lie between 0 and $\bar{\theta}_c$, and the propagation constants β_m lie between n_2k_o and n_1k_o. These results should be compared with those shown in Fig. 9.1-3 for the planar-mirror waveguide.

The bounce angles θ_m and the propagation constants β_m of TM modes can be found by using the same equation (9.2-1), but with the phase shift φ_r given by (6.2-13). Similar results are obtained.

Number of Modes

To determine the number of TE modes supported by the dielectric waveguide we examine the diagram in Fig. 9.2-2. The abscissa is divided into equal intervals of width $\lambda/2d$, each of which contains a mode marked by a filled circle. This extends over angles for which $\sin\theta \leq \sin\overline{\theta}_c$. The number of TE modes is therefore the smallest integer greater than $\sin\overline{\theta}_c/(\lambda/2d)$, so that

$$M \doteq \frac{\sin\overline{\theta}_c}{\lambda/2d}. \tag{9.2-6}$$

The symbol $\doteq$ denotes that $\sin\overline{\theta}_c/(\lambda/2d)$ is increased to the nearest integer. For example, if $\sin\overline{\theta}_c/(\lambda/2d) = 0.9$, 1, or 1.1, then $M = 1$, 2, and 2, respectively. Substituting $\cos\overline{\theta}_c = n_2/n_1$ into (9.2-6), we obtain

$$\boxed{M \doteq \frac{2d}{\lambda_o}\,\mathrm{NA},} \tag{9.2-7}$$

Number of TE Modes

where

$$\boxed{\mathrm{NA} = \sqrt{n_1^2 - n_2^2}} \tag{9.2-8}$$

Numerical Aperture

is the numerical aperture of the waveguide (the NA is the sine of the angle of acceptance of rays from air into the slab; see Exercise 1.2-5). If $d/\lambda_o = 10$, $n_1 = 1.47$, and $n_2 = 1.46$, for example, then $\overline{\theta}_c = 6.7°$, NA $= 0.171$, and $M = 4$ TE modes. A similar expression can be obtained for the TM modes.

When $\lambda/2d > \sin\overline{\theta}_c$ or $(2d/\lambda_o)\mathrm{NA} < 1$, only one mode is allowed. The waveguide is then a **single-mode waveguide**. This occurs when the slab is sufficiently thin or the wavelength is sufficiently long. Unlike the mirror waveguide, the dielectric waveguide has no absolute cutoff wavelength (or cutoff frequency). In a dielectric waveguide there is at least one TE mode, since the fundamental mode $m = 0$ is always allowed. Each of the modes $m = 1, 2, \ldots$ has its own cutoff wavelength, however.

Stated in terms of frequency, the condition for single-mode operation is that $\nu < \nu_c$, or $\omega < \omega_c$, where the mode cutoff frequency is

$$\boxed{\nu_c = \omega_c/2\pi = \frac{1}{\mathrm{NA}}\,\frac{c_o}{2d}.} \tag{9.2-9}$$

Mode Cutoff Frequency

The number of modes is then $M \doteq \nu/\nu_c = \omega/\omega_c$, which is the relation illustrated in Fig. 9.2-4. M is incremented by unity as ω increases by ω_c. Identical expressions for the number of TM modes are obtained via a similar derivation.

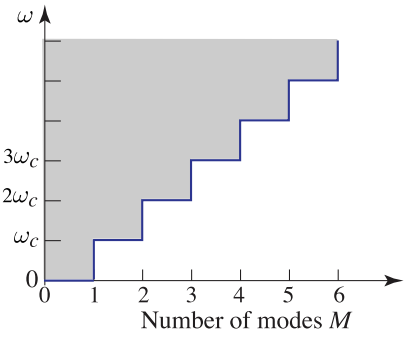


Figure 9.2-4 Number of TE modes as a function of frequency. Compare with Fig. 9.1-5(*a*) for the planar-mirror waveguide. There is no forbidden band in the case at hand.

EXAMPLE 9.2-1. ***Number of Modes in an AlGaAs Waveguide.*** Consider a waveguide fabricated by sandwiching a layer of $Al_xGa_{1-x}As$ between two layers of $Al_yGa_{1-y}As$. The refractive index of this ternary semiconductor depends on the relative proportions of Al and Ga. Assume that x and y are chosen such that $n_1 = 3.50$ and $n_1 - n_2 = 0.05$ at an operating wavelength of $\lambda_o = 0.9\ \mu$m. If the core has width $d = 10\ \mu$m, in accordance with (9.2-7) and (9.2-8) there will be $M = 14$ supported TE modes. Only a single mode is allowed when $d < 0.76\ \mu$m.

B. Field Distributions

We now determine the field distributions of the TE modes.

Internal Field

The field inside the slab is composed of two TEM plane waves traveling at angles θ_m and $-\theta_m$ with the z axis with wavevector components $(0, \pm n_1 k_o \sin\theta_m, n_1 k_o \cos\theta_m)$. They have the same amplitude and phase shift $m\pi$ (half that of a round trip) at the center of the slab. The electric-field complex amplitude is therefore $E_x(y,z) = \mathfrak{a}_m u_m(y) \exp(-j\beta_m z)$, where $\beta_m = n_1 k_o \cos\theta_m$ is the propagation constant, $\mathfrak{a}_m$ is a constant,

$$u_m(y) \propto \begin{cases} \cos\left(2\pi \dfrac{\sin\theta_m}{\lambda}\, y\right), & m = 0, 2, 4, \ldots \\ \sin\left(2\pi \dfrac{\sin\theta_m}{\lambda}\, y\right), & m = 1, 3, 5, \ldots, \end{cases} \qquad -\frac{d}{2} \le y \le \frac{d}{2}, \tag{9.2-10}$$

and $\lambda = \lambda_o / n_1$. Note that although the field is harmonic, it does not vanish at the slab boundary. As m increases, $\sin\theta_m$ increases, so that higher-order modes vary more rapidly with y.

External Field

The external field must match the internal field at all boundary points $y = \pm d/2$. It is therefore clear that it must vary with z as $\exp(-j\beta_m z)$. Substituting the field $E_x(y,z) = \mathfrak{a}_m u_m(y) \exp(-j\beta_m z)$ into the Helmholtz equation $(\nabla^2 + n_2^2 k_o^2) E_x(y,z) = 0$ leads to

$$\frac{d^2 u_m}{dy^2} - \gamma_m^2 u_m = 0, \tag{9.2-11}$$

where

$$\gamma_m^2 = \beta_m^2 - n_2^2 k_o^2. \tag{9.2-12}$$

Since $\beta_m > n_2 k_o$ for guided modes (See Fig. 9.2-3), $\gamma_m^2 > 0$, so that (9.2-11) is satisfied by the exponential functions $\exp(-\gamma_m y)$ and $\exp(\gamma_m y)$. Since the field must decay away from the slab, we choose $\exp(-\gamma_m y)$ in the upper medium and $\exp(\gamma_m y)$ in the lower medium

$$u_m(y) \propto \begin{cases} \exp(-\gamma_m y), & y > d/2 \\ \exp(\gamma_m y), & y < -d/2. \end{cases} \tag{9.2-13}$$

The decay rate γ_m is the field extinction coefficient. The wave is said to be an **evanescent wave**. Substituting $\beta_m = n_1 k_o \cos\theta_m$ and $\cos\bar{\theta}_c = n_2/n_1$ into (9.2-12), we obtain

$$\boxed{\gamma_m = n_2 k_o \sqrt{\frac{\cos^2\theta_m}{\cos^2\bar{\theta}_c} - 1}\,.} \tag{9.2-14}$$

Extinction Coefficient

As the mode number m increases, θ_m increases, and γ_m decreases. Higher-order modes therefore penetrate deeper into the cover and substrate.

To determine the proportionality constants in (9.2-10) and (9.2-13), we match the internal and external fields at $y = d/2$ and use the normalization

$$\int_{-\infty}^{\infty} u_m^2(y)\, dy = 1. \tag{9.2-15}$$

This gives an expression for $u_m(y)$ valid for all y. These functions are illustrated in Fig. 9.2-5. As in the mirror waveguide, all of the $u_m(y)$ are orthogonal, i.e.,

$$\int_{-\infty}^{\infty} u_m(y)\, u_l(y)\, dy = 0, \qquad l \neq m. \tag{9.2-16}$$

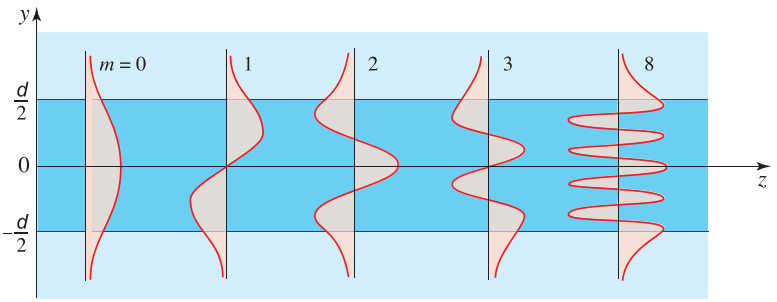


Figure 9.2-5 Field distributions for TE guided modes in a dielectric waveguide. These results should be compared with those shown in Fig. 9.1-4 for the planar-mirror waveguide.

An arbitrary TE field in the dielectric waveguide can be written as a superposition of these modes:

$$E_x(y,z) = \sum_m \mathrm{a}_m u_m(y) \exp(-j\beta_m z), \tag{9.2-17}$$

where a_m is the amplitude of mode m.

EXERCISE 9.2-1

Confinement Factor. The power confinement factor is the ratio of power in the slab to the total power

$$\Gamma_m = \frac{\int_0^{d/2} u_m^2(y)\,dy}{\int_0^{\infty} u_m^2(y)\,dy}. \tag{9.2-18}$$

Derive an expression for Γ_m as a function of the angle θ_m and the ratio d/λ. Demonstrate that the lowest-order mode (smallest θ_m) has the highest power confinement factor.

The field distributions of the TM modes may be similarly determined (Fig. 9.2-6). Since it is parallel to the slab boundary, the z component of the electric field behaves similarly to the x component of the TE electric field. The analysis may start by determining $E_z(y,z)$. Using the properties of the constituent TEM waves, the other components $E_y(y,z)$ and $H_x(y,z)$ may readily be determined, as was done for mirror waveguides. Alternatively, Maxwell's equations may be used to determine these fields.

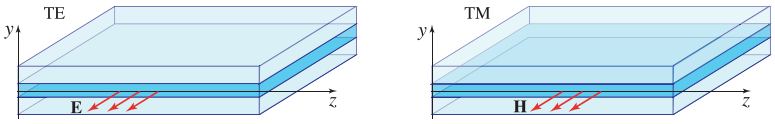
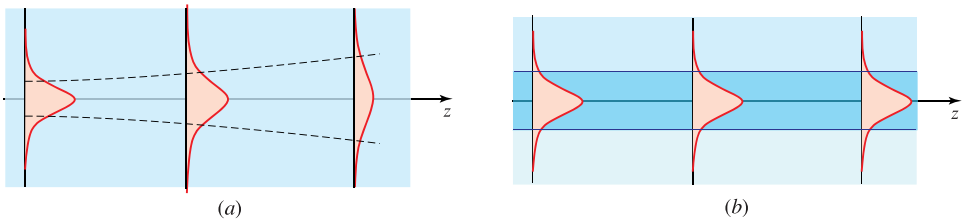


Figure 9.2-6 TE and TM modes in a planar dielectric waveguide.

The field distribution of the lowest-order TE mode ($m = 0$) is similar in shape to that of the Gaussian beam (see Chapter 3). However, unlike the Gaussian beam, guided light does not spread in the transverse direction as it propagates in the axial direction (see Fig. 9.2-7). In a waveguide, the tendency of light to diffract is compensated by the guiding action of the medium.



(*a*) (*b*)

Figure 9.2-7 (*a*) Gaussian beam in a homogeneous medium. (*b*) Guided mode in a dielectric waveguide.

C. Dispersion Relation and Group Velocities

The dispersion relation (ω versus β) is obtained by writing the self-consistency equation (9.2-2) in terms of β and ω. Since $k_y^2 = (\omega/c_1)^2 - \beta^2$, (9.2-2) gives

$$2d\sqrt{\frac{\omega^2}{c_1^2} - \beta^2} = 2\varphi_r + 2\pi m. \tag{9.2-19}$$

Since $\cos\theta = \beta/(\omega/c_1)$ and $\cos\bar{\theta}_c = n_2/n_1 = c_1/c_2$, (9.2-3) becomes

$$\tan^2\frac{\varphi_r}{2} = \frac{\beta^2 - \omega^2/c_2^2}{\omega^2/c_1^2 - \beta^2}. \tag{9.2-20}$$

Substituting (9.2-20) into (9.2-19) we obtain

$$\tan^2\left(\frac{d}{2}\sqrt{\frac{\omega^2}{c_1^2} - \beta^2} - m\frac{\pi}{2}\right) = \frac{\beta^2 - \omega^2/c_2^2}{\omega^2/c_1^2 - \beta^2}. \tag{9.2-21}$$

Dispersion Relation (TE Modes)

This relation may be plotted by rewriting it in parametric form,

$$\frac{\omega}{\omega_c} = \frac{\sqrt{n_1^2 - n_2^2}}{\sqrt{n_1^2 - n^2}}\left(m + \frac{2}{\pi}\tan^{-1}\sqrt{\frac{n^2 - n_2^2}{n_1^2 - n^2}}\right), \quad \beta = n\omega/c_o, \tag{9.2-22}$$

in terms of the **effective refractive index** n defined in (9.2-22), where $\omega_c/2\pi = c_o/2d\text{NA}$ is the mode-cutoff angular frequency. As shown in the schematic plot in Fig. 9.2-8(a), the dispersion relations for the different modes lie between the lines $\omega = c_2\beta$ and $\omega = c_1\beta$, the **light lines** representing propagation in homogeneous media with the refractive indices of the surrounding medium and the slab, respectively. As the frequency increases above the mode cutoff frequency, the dispersion relation moves from the light line of the surrounding medium toward the light line of the slab, i.e., the effective refractive index n increases from n_2 to n_1. This effect is indicative of a stronger confinement of waves of shorter wavelength in the medium of higher refractive index.

The group velocity is obtained from the dispersion relation by determining the slope $v = d\omega/d\beta$ for each of the guided modes. The dependence of the group velocity on the angular frequency is illustrated schematically in Fig. 9.2-8(b). As the angular frequency increases above the mode cutoff frequency for each mode, the group velocity decreases from its maximum value c_2, reaches a minimum value slightly below c_1, and then asymptotically returns back toward c_1. The group velocities of the allowed modes thus range from c_2 to a value slightly below c_1.

In propagating through a multimode waveguide, optical pulses spread in time since the modes have different velocities, an effect called **modal dispersion**. In a single-mode waveguide, an optical pulse spreads as a result of the dependence of the group velocity on frequency. This effect is called **group velocity dispersion** (GVD). As shown in Sec. 5.7, GVD occurs in homogeneous materials by virtue of the frequency dependence of the refractive index of the material. Moreover, GVD occurs in waveguides even in the absence of material dispersion. It is then a consequence of the frequency dependence of the propagation coefficients, which are determined by the dependence of wave confinement on wavelength. As illustrated in Fig. 9.2-8(b), each

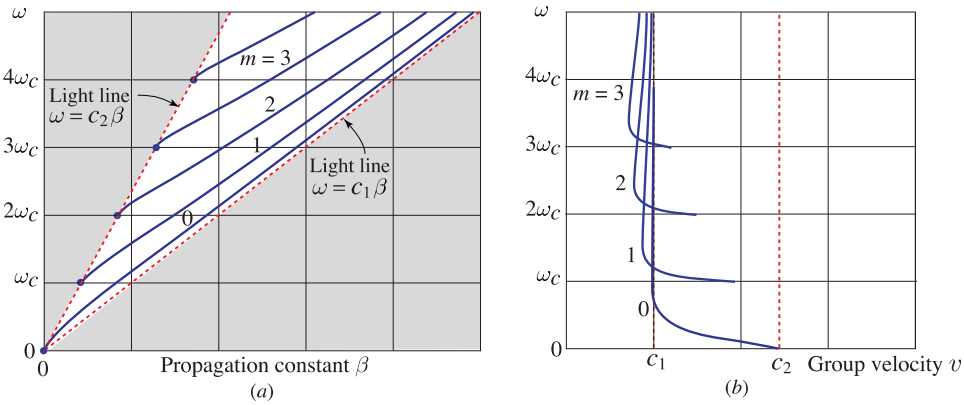


Figure 9.2-8 Schematic representations of (a) the dispersion relation for the different TE modes, $m = 0, 1, 2, \ldots$; and (b) the frequency dependence of the group velocity, which is the derivative of the dispersion relation, $v = d\omega/d\beta$.

mode has a particular angular frequency at which the group velocity changes slowly with frequency (the point at which v reaches its minimum value so that its derivative with respect to ω is zero). At this frequency, the GVD coefficient is zero and pulse spreading is negligible.

An approximate expression for the group velocity may be obtained by taking the total derivative of (9.2-19) with respect to β,

$$\frac{2d}{2k_y}\left(\frac{2\omega}{c_1^2}\frac{d\omega}{d\beta} - 2\beta\right) = 2\frac{\partial\varphi_r}{\partial\beta} + 2\frac{\partial\varphi_r}{\partial\omega}\frac{d\omega}{d\beta}. \tag{9.2-23}$$

Substituting $d\omega/d\beta = v$, $k_y/(\omega/c_1) = \sin\theta$, and $k_y/\beta = \tan\theta$, and introducing the new parameters

$$\Delta z = \frac{\partial\varphi_r}{\partial\beta}, \qquad \Delta\tau = -\frac{\partial\varphi_r}{\partial\omega}, \tag{9.2-24}$$

we obtain

$$v = \frac{d\cot\theta + \Delta z}{d\csc\theta/c_1 + \Delta\tau}. \tag{9.2-25}$$

As we recall from (9.1-14) and Fig. 9.1-6 for the planar-mirror waveguide, $d\cot\theta$ is the distance traveled in the z direction as a ray travels once between the two boundaries. This takes a time $d\csc\theta/c_1$. The ratio $d\cot\theta/(d\csc\theta)/c_1 = c_1\cos\theta$ yields the group velocity for the mirror waveguide. The expression (9.2-25) for the group velocity in a dielectric waveguide indicates that the ray travels an additional distance $\Delta z = \partial\varphi_r/\partial\beta$, a trip that lasts a time $\Delta\tau = -\partial\varphi_r/\partial\omega$. We can think of this as an effective penetration of the ray into the cladding, or as an effective lateral shift of the ray, as shown in Fig. 9.2-9. The penetration of a ray undergoing total internal reflection is known as the **Goos–Hänchen effect** (see Prob. 6.2-6). Using (9.2-24) it can be shown that $\Delta z/\Delta\tau = \omega/\beta = c_1/\cos\theta$.

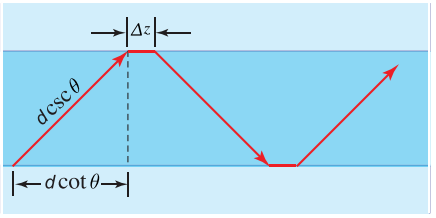


Figure 9.2-9 A ray model that replaces the reflection phase shift with an additional distance Δz traversed at velocity $c_1/\cos\theta$.

EXERCISE 9.2-2

The Asymmetric Planar Waveguide. Examine the TE field in an asymmetric planar waveguide consisting of a dielectric slab of width d and refractive index n_1 placed on a substrate of lower refractive index n_2 and covered with a medium of refractive index $n_3 < n_2 < n_1$, as illustrated in Fig. 9.2-10.

(a) Determine an expression for the maximum inclination angle θ of plane waves undergoing total internal reflection, and the corresponding numerical aperture NA of the waveguide.

(b) Write an expression for the self-consistency condition, similar to (9.2-4).

(c) Determine an approximate expression for the number of modes M (valid when M is very large).

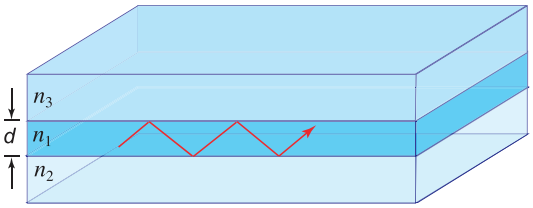


Figure 9.2-10 Asymmetric planar waveguide.

9.3 TWO-DIMENSIONAL WAVEGUIDES

The planar-mirror waveguide and the planar dielectric waveguide studied in the preceding two sections confine light in one transverse direction (the y direction) while guiding it along the z direction. Two-dimensional waveguides confine light in the two transverse directions (the x and y directions). The principle of operation and the underlying modal structure of two-dimensional waveguides is basically the same as planar waveguides; only the mathematical description is lengthier. This section is a brief description of the nature of modes in two-dimensional waveguides. Details can be found in specialized books. Chapter 10 is devoted to an important example of two-dimensional waveguides, the cylindrical dielectric waveguide used in optical fibers.

Rectangular Mirror Waveguide

The simplest generalization of the planar waveguide is the rectangular waveguide (Fig. 9.3-1). If the walls of the waveguide are mirrors, then, as in the planar case, light is guided by multiple reflections at all angles. For simplicity, we assume that the cross section of the waveguide is a square of width d. If a plane wave of wavevector (k_x, k_y, k_z) and its multiple reflections are to exist self-consistently inside the wave-

guide, it must satisfy the conditions:

$$\begin{aligned} 2k_x d &= 2\pi m_x, \qquad m_x = 1, 2, \ldots \\ 2k_y d &= 2\pi m_y, \qquad m_y = 1, 2, \ldots, \end{aligned} \tag{9.3-1}$$

which are obvious generalizations of (9.1-3).

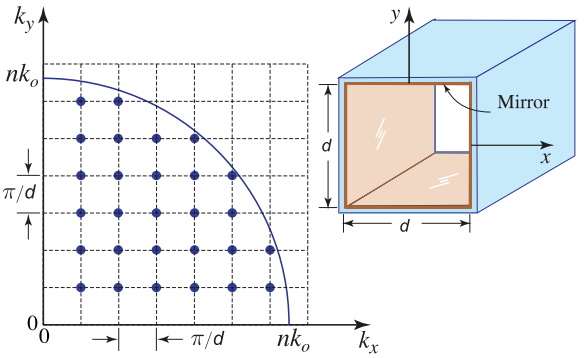


Figure 9.3-1 Modes of a rectangular mirror waveguide are characterized by a finite number of discrete values of k_x and k_y represented by dots.

The propagation constant $\beta = k_z$ can be determined from k_x and k_y by using the relation $k_x^2 + k_y^2 + \beta^2 = n^2 k_o^2$. The three components of the wavevector therefore have discrete values, yielding a finite number of modes. Each mode is identified by two indices m_x and m_y (instead of one index m). All positive integer values of m_x and m_y are allowed as long as $k_x^2 + k_y^2 \leq n^2 k_o^2$, as illustrated in Fig. 9.3-1.

The number of modes M can be easily determined by counting the number of dots within a quarter circle of radius nk_o in the k_y versus k_x diagram (Fig. 9.3-1). If this number is large, it may be approximated by the ratio of the area $\pi(nk_o)^2/4$ to the area of a unit cell $(\pi/d)^2$:

$$M \approx \frac{\pi}{4}\left(\frac{2d}{\lambda}\right)^2. \tag{9.3-2}$$

Since there are two polarizations per mode, the total number of modes is actually $2M$. Comparing this to the number of modes in a one-dimensional mirror waveguide, $M \approx 2d/\lambda$, we see that increase of the dimensionality yields approximately the square of the number of modes. The number of modes is a measure of the degrees of freedom. When we add a second dimension we simply multiply the number of degrees of freedom.

The field distributions associated with these modes are generalizations of those in the planar case. Patterns such as those in Fig. 9.1-4 are obtained in each of the x and y directions depending on the mode indices m_x and m_y.

Rectangular Dielectric Waveguide

A dielectric cylinder of refractive index n_1 with square cross section of width d is embedded in a medium of slightly lower refractive index n_2. The waveguide nodes can be determined using a similar theory. Components of the wavevector (k_x, k_y, k_z) must satisfy the condition $k_x^2 + k_y^2 \leq n_1^2 k_o^2 \sin^2 \overline{\theta}_c$, where $\overline{\theta}_c = \cos^{-1}(n_2/n_1)$, so that k_x and k_y lie in the area shown in Fig. 9.3-2. The values of k_x and k_y for the different modes can be obtained from a self-consistency condition in which the phase shifts at the dielectric boundary are included, as was done in the planar case.

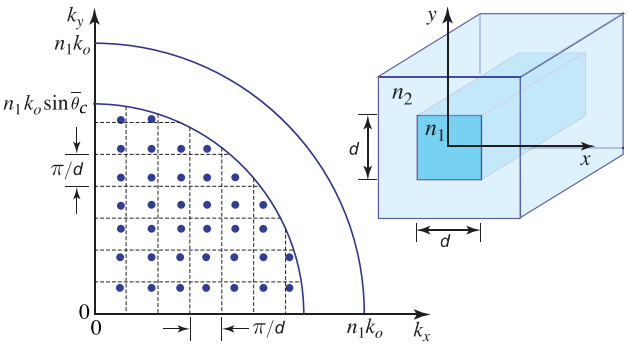


Figure 9.3-2 Geometry of a rectangular dielectric waveguide. The values of k_x and k_y for the waveguide modes are marked by dots.

Unlike the mirror waveguide, k_x and k_y of the modes are not uniformly spaced. However, two consecutive values of k_x (or k_y) are separated by an *average* value of π/d (the same as for the mirror waveguide). The number of modes can therefore be approximated by counting the number of dots in the inner circle in the k_y versus k_x diagram of Fig. 9.3-2, assuming an average spacing of π/d. The result is $M \approx (\pi/4)(n_1 k_o \sin\bar{\theta}_c)^2/(\pi/d)^2$, from which

$$M \approx \frac{\pi}{4}\left(\frac{2d}{\lambda_o}\right)^2 (\text{NA})^2, \tag{9.3-3}$$

Number of TE Modes

where $\text{NA} = \sqrt{n_1^2 - n_2^2}$ is the numerical aperture. The approximation is satisfactory when M is large. There is also an identical number M of TM modes. The number of modes is roughly the square of that for the planar dielectric waveguide (9.2-7).

Geometries for Channel Waveguides

As illustrated in Fig. 9.3-3, channel waveguides can take many forms. Representative examples include immersed-strip (or buried channel), embedded-strip, ridge, rib, and strip-loaded geometries. Exact analysis for many of these geometries can be rather complex, but approximations serve well. The reader is referred to specialized texts for details pertaining to this topic.

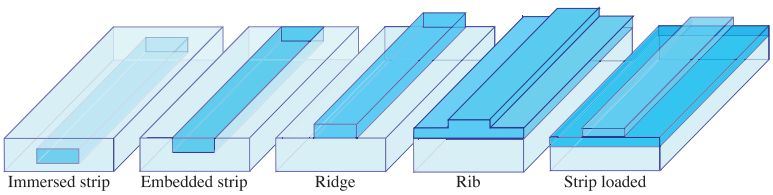


Figure 9.3-3 Various waveguide geometries. The darker the shading, the higher the refractive index.

Waveguides may also be fabricated in a variety of configurations, as illustrated in Fig. 9.3-4 for the embedded-strip geometry. S-bends are used to offset the propagation axis. The Y-branch plays the role of a beamsplitter or beam combiner. A pair of Y-branches may be used to construct a Mach–Zehnder interferometer. Two waveguides in close proximity, or intersecting with each other, can exchange power and be used as directional couplers, as will become apparent in Sec. 9.4B.

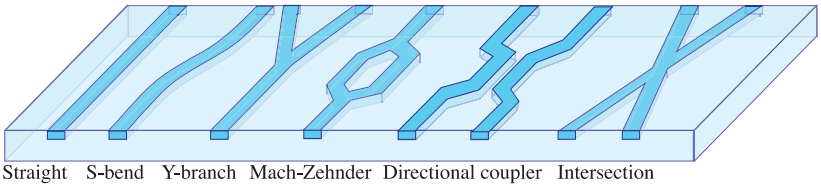


Figure 9.3-4 Different waveguide configurations, in this case for the embedded-strip geometry.

Materials

The earliest optical waveguides were fabricated from electro-optic materials, principally lithium niobate (**$LiNbO_3$**). As shown in Fig. 9.3-5(a), an embedded-strip waveguide using this material may be fabricated by indiffusing titanium (Ti) in a lithium niobate substrate to increase its refractive index in the region of the strip.

Semiconductors are also commonly used. A **GaAs** rib waveguide may be fabricated by using layers of GaAs and AlGaAs, which has lower refractive index [Fig. 9.3-5(b)]. Another semiconductor material of substantial importance in optical waveguides is **InP**. Its refractive index may be controlled by making use of n-type and p-type dopants, or by using the quaternary semiconductor InGaAsP with various mixing ratios. The ridge waveguide illustrated in Fig. 9.3-5(c) offers strong optical confinement because it is surrounded on three sides by lower index materials — air on two sides and InGaAsP of a different composition on the third.

Waveguides may also be fabricated from **silicon-on-insulator (SOI)**, usually **silica-on-silicon** (SiO_2/Si), by making use of standard silicon and oxide etching tools. Since the refractive index of Si is ≈ 3.5, and that of silica is < 1.5, this combination of materials exhibits a large refractive-index difference Δn. A typical SOI structure takes the form of a Si rib waveguide atop a layer of silica, which serves as a lower cladding, supported by a silicon substrate [Fig. 9.3-5(d)]. Silicon processing and fabrication has been extraordinarily well developed by the microelectronics industry, and compatibility with **complementary metal-oxide-semiconductor (CMOS)** fabrication technology offers an important advantage. This approach lies in the domain of **silicon photonics** (Sec. 25.1E).

Glass waveguides fabricated by ion exchange, as well as polymer waveguides, are also emerging as viable technologies.

The ability to modulate the refractive index is an important requirement for materials used in integrated-photonic devices such as light modulators and switches, as will become evident in Chapters 21 and 24.

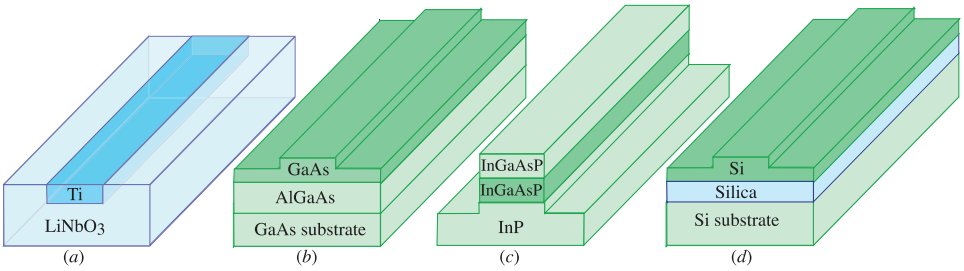


Figure 9.3-5 (a) Ti:$LiNbO_3$ embedded-strip waveguide. (b) Rib waveguide with GaAs core, AlGaAs lower cladding, and GaAs substrate. (c) InGaAsP ridge waveguide with air and lower-index InGaAsP cladding. (d) SOI rib waveguide with Si core, silica lower cladding, and Si substrate compatible with CMOS electronics technology.

9.4 OPTICAL COUPLING IN WAVEGUIDES

A. Input Couplers

Mode Excitation

As indicated in previous sections, light propagates in a waveguide in the form of modes. The complex amplitude of the optical field is generally a superposition of these modes,

$$E(y,z) = \sum_m \mathrm{a}_m u_m(y) \exp(-j\beta_m z), \tag{9.4-1}$$

where a_m is the amplitude, $u_m(y)$ is the transverse distribution (assumed to be real), and β_m is the propagation constant of mode m.

The amplitudes of the different modes depend on the nature of the light source used to excite the waveguide. If the source has a distribution that is a perfect match to a specific mode, only that mode will be excited. In general, a source of arbitrary distribution $s(y)$ excites different modes at different levels. The fraction of power transferred from the source to mode m depends on the degree of similarity between $s(y)$ and $u_m(y)$. To establish this, we write $s(y)$ as an expansion (a weighted superposition) of the orthogonal functions $u_m(y)$,

$$s(y) = \sum_m \mathrm{a}_m u_m(y), \tag{9.4-2}$$

where the coefficient a_l, which represents the amplitude of the excited mode l, is

$$\mathrm{a}_l = \int_{-\infty}^{\infty} s(y)\, u_l(y)\, dy. \tag{9.4-3}$$

This expression can be derived by multiplying both sides of (9.4-2) by $u_l(y)$, integrating with respect to y, and using the orthogonality relation $\int_{-\infty}^{\infty} u_l(y)\, u_m(y)\, dy = 0$ for $l \neq m$ along with the normalization condition. The coefficient a_l represents the degree of similarity (or correlation) between the source distribution $s(y)$ and the mode distribution $u_l(y)$.

Input Couplers

Light may be coupled into a waveguide by directly focusing it at one end (Fig. 9.4-1). To excite a given mode, the transverse distribution of the incident light $s(y)$ should match that of the mode. The polarization of the incident light must also match that of the desired mode. Because of the small dimensions of the waveguide slab, focusing and alignment are usually difficult and coupling using this method is inefficient.

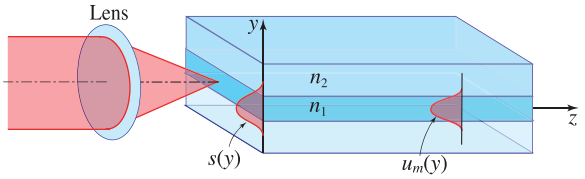


Figure 9.4-1 Coupling an optical beam into an optical waveguide.

In a multimode waveguide, the amount of coupling can be assessed by using a ray-optics approach (Fig. 9.4-2). The guided rays within the waveguide are confined

to an angle $\overline{\theta}_c = \cos^{-1}(n_2/n_1)$. Because of refraction at the input to the waveguide, this corresponds to an external angle θ_a satisfying $\text{NA} = \sin\theta_a = n_1 \sin\overline{\theta}_c = n_1\sqrt{1-(n_2/n_1)^2} = \sqrt{n_1^2 - n_2^2}$, where NA is the numerical aperture of the waveguide (see Exercise 1.2-5). For maximum coupling efficiency the incident light should be focused within the angle θ_a.

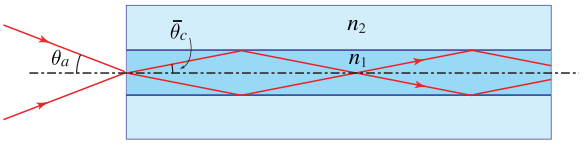


Figure 9.4-2 Focusing rays into a multimode waveguide.

Light may also be coupled from a semiconductor source (a light-emitting diode or a laser diode) into a waveguide by simply aligning the ends of the source and the waveguide, leaving a small space that is selected for maximum coupling (Fig. 9.4-3). In light-emitting diodes, light originates from a semiconductor junction region and is emitted in all directions. In a laser diode, the emitted light is confined in a waveguide of its own (light-emitting diodes and laser diodes are described in Chapter 18). Other methods of coupling light into waveguides include the use of prisms, diffraction gratings, and other waveguides, as discussed below.

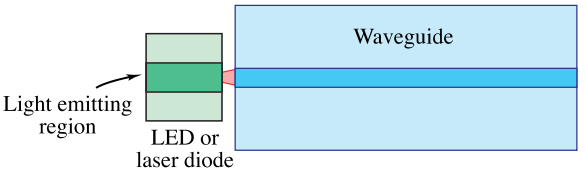


Figure 9.4-3 End butt coupling from a light-emitting diode or laser diode into a waveguide.

Prism and Grating Side Couplers

Can optical power be coupled into a guided mode of a waveguide by use of a source wave entering from the side at some angle θ_i in the cladding, as shown in Fig. 9.4-4(a)? The condition for such coupling is that the axial component of the wavevector of the incident wave, $n_2 k_o \cos\theta_i$, equals the propagation constant β_m of the guided mode. Since $\beta_m > n_2 k_o$ (see Fig. 9.4-4), it is not possible to achieve the required phase-matching condition $\beta_m = n_2 k_o \cos\theta_i$. The axial component of the wavevector of the incident wave is simply too small. However, the problem may be alleviated by use of a prism or a grating.

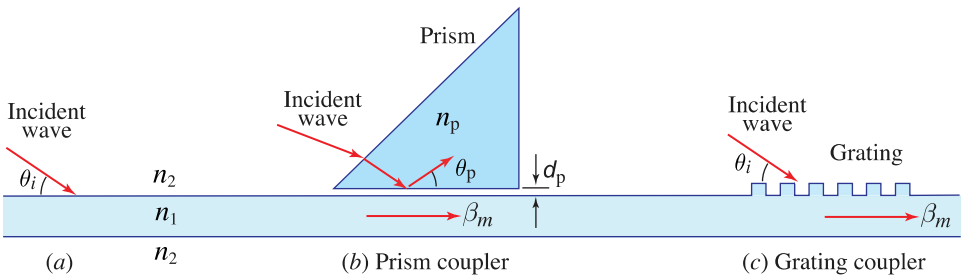


Figure 9.4-4 Prism and grating side couplers.

As illustrated in Fig. 9.4-4(b), a prism of refractive index $n_\mathrm{p} > n_2$ is placed at a small distance d_p from the waveguide slab. The incident wave is refracted into the prism where it undergoes total internal reflection at an angle θ_p. The incident and reflected waves form a wave traveling in the z direction with propagation constant $\beta_\mathrm{p} = n_\mathrm{p}\, k_o \cos\theta_\mathrm{p}$. The transverse field distribution extends into the space separating the prism and the slab as an exponentially decaying evanescent wave. If the distance d_p is sufficiently small, the wave couples to a mode of the slab waveguide with a matching propagation constant $\beta_m \approx \beta_\mathrm{p} = n_\mathrm{p}\, k_o \cos\theta_\mathrm{p}$. Since $n_\mathrm{p} > n_2$, phase matching is possible, and if an appropriate interaction distance is selected, frustrated total internal reflection ensues and significant power can be coupled into the waveguide. The operation may also be reversed to make an output coupler, extracting light from the slab waveguide into free space. This is the same approach as that used to excite a surface plasmon polariton wave at a metal–dielectric boundary, as illustrated in Fig. 8.2-5.

The grating [Fig. 9.4-4(c)] addresses the phase-matching problem by modifying the wavevector of the incoming wave. A grating with period Λ modulates the incoming wave by phase factors $2\pi q/\Lambda z$, where $q = \pm 1, \pm 2, \ldots$. These are equivalent to changes of the axial component of the wavevector by factors $2\pi q/\Lambda$. The phase-matching condition can now be satisfied if $n_2\, k_o \cos\theta_i + 2\pi q/\Lambda = \beta_m$, with $q = 1$, for example. The grating may even be designed to enhance the $q = 1$ component.

B. Coupled Waveguides

If two waveguides are sufficiently close such that their fields overlap, light can be coupled from one into the other. Optical power can then be transferred between the waveguides, an effect that can be used to make optical couplers and switches. The basic principle of waveguide coupling is presented here; couplers and switches are discussed in Chapters 24 and 25.

Consider two parallel planar waveguides made of two slabs of widths d, separation $2a$, and refractive indices n_1 and n_2, embedded in a medium of refraction index n that is slightly smaller than n_1 and n_2, as illustrated in Fig. 9.4-5. Each of the waveguides is assumed to be single-mode. The separation between the waveguides is such that the optical field outside the slab of one waveguide (in the absence of the other) overlaps slightly with the slab of the other waveguide.

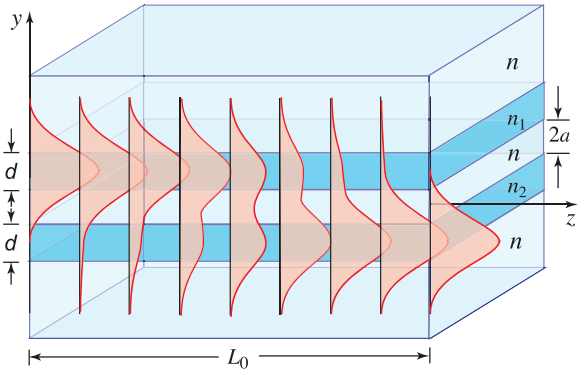


Figure 9.4-5 Coupling between two parallel planar waveguides. At $z = 0$ the light is located principally in the upper waveguide (waveguide 1); at $z = L_0/2$ the light is divided equally between the two waveguides; and at $z = L_0$ the light is located principally in the lower waveguide (waveguide 2). The distance L_0 at which the power is completely transferred from one waveguide to the other is called the **coupling length** or **transfer distance**.

The formal approach to studying the propagation of light in this structure is to write Maxwell's equations for the different regions and use the boundary conditions to determine the modes of the overall system. These modes are different from those of each of the waveguides in isolation. An exact analysis is not easy and is beyond the scope of this book. For weak coupling, however, a simplified approximate theory, known as coupled-mode theory, is often satisfactory.

Coupled-mode theory assumes that the mode of each waveguide is determined as if the other waveguide were absent. In the presence of both waveguides, the modes are taken to remain approximately unchanged, say $u_1(y)\exp(-j\beta_1 z)$ and $u_2(y)\exp(-j\beta_2 z)$. Coupling is assumed to modify only the *amplitudes* of these modes without affecting either their transverse spatial distributions or their propagation constants. The amplitudes of the modes of waveguides 1 and 2 are therefore functions of z, $\mathfrak{a}_1(z)$, and $\mathfrak{a}_2(z)$. The theory is directed toward determining $\mathfrak{a}_1(z)$ and $\mathfrak{a}_2(z)$ under appropriate boundary conditions.

Coupling can be regarded as a scattering effect. The field of waveguide 1 is scattered from waveguide 2, creating a source of light that changes the amplitude of the field in waveguide 2. The field of waveguide 2 has a similar effect on waveguide 1. An analysis of this mutual interaction leads to two coupled differential equations that govern the variation of the amplitudes $\mathfrak{a}_1(z)$ and $\mathfrak{a}_2(z)$.

It can be shown (see the derivation at the end of this section) that the amplitudes $\mathfrak{a}_1(z)$ and $\mathfrak{a}_2(z)$ are governed by two coupled first-order differential equations

$$\frac{d\mathfrak{a}_1}{dz} = -j\mathcal{C}_{21}\exp(j\,\Delta\beta\, z)\,\mathfrak{a}_2(z) \tag{9.4-4a}$$

$$\frac{d\mathfrak{a}_2}{dz} = -j\mathcal{C}_{12}\exp(-j\,\Delta\beta\, z)\,\mathfrak{a}_1(z), \tag{9.4-4b}$$

Coupled-Mode Equations

where

$$\Delta\beta = \beta_1 - \beta_2 \tag{9.4-5}$$

is the phase mismatch per unit length and

$$\mathcal{C}_{21} = \tfrac{1}{2}\left(n_2^2 - n^2\right)\frac{k_o^2}{\beta_1}\int_a^{a+d} u_1(y)\,u_2(y)\,dy,$$

$$\mathcal{C}_{12} = \tfrac{1}{2}\left(n_1^2 - n^2\right)\frac{k_o^2}{\beta_2}\int_{-a-d}^{-a} u_2(y)\,u_1(y)\,dy \tag{9.4-6}$$

are coupling coefficients. We see from (9.4-4) that the rate of variation of $\mathfrak{a}_1$ is proportional to $\mathfrak{a}_2$, and *vice versa*. The coefficient of proportionality is the product of the coupling coefficient and the phase mismatch factor $\exp(j\,\Delta\beta\, z)$.

The coupled-mode equations in (9.4-4) may be solved by beginning with harmonic trial solutions of the form $\mathfrak{a}_1(z) = \mathfrak{b}_1\exp(j\gamma z)\exp(j\Delta\beta\, z/2)$ and $\mathfrak{a}_2(z) = \mathfrak{b}_2\exp(j\gamma z)\exp(-j\Delta\beta\, z/2)$, where $\mathfrak{b}_1$ and $\mathfrak{b}_2$ are constants. These solution satisfy (9.4-4) provided that the following condition is satisfied:

$$\gamma = \pm\sqrt{\left(\frac{\Delta\beta}{2}\right)^2 + \mathcal{C}^2}, \quad \mathcal{C} = \sqrt{\mathcal{C}_{12}\mathcal{C}_{21}}\,. \tag{9.4-7}$$

Since γ has two possible values, we modify the trial solutions to be superpositions of $\exp(j\gamma z)$ and $\exp(-j\gamma z)$, or of $\sin(\gamma z)$ and $\cos(\gamma z)$, where γ is the positive value of the square root in (9.4-7). The weights of the superposition are established from the boundary values $\mathfrak{a}_1(0)$ and $\mathfrak{a}_2(0)$. The final outcome is

$$\mathfrak{a}_1(z) = A(z)\mathfrak{a}_1(0) + B(z)\mathfrak{a}_2(0) \tag{9.4-8a}$$

$$\mathfrak{a}_2(z) = C(z)\mathfrak{a}_1(0) + D(z)\mathfrak{a}_2(0), \tag{9.4-8b}$$

where

$$A(z) = D^*(z) = \exp\left(\frac{j\,\Delta\beta\, z}{2}\right)\left(\cos\gamma z - j\frac{\Delta\beta}{2\gamma}\sin\gamma z\right) \tag{9.4-9a}$$

$$B(z) = \frac{\mathcal{C}_{21}}{j\gamma}\exp\left(j\frac{\Delta\beta\, z}{2}\right)\sin\gamma z \tag{9.4-9b}$$

$$C(z) = \frac{\mathcal{C}_{12}}{j\gamma}\exp\left(-j\frac{\Delta\beta\, z}{2}\right)\sin\gamma z \tag{9.4-9c}$$

are elements of a transmission matrix $\mathbf{T}$ that relates the output and input fields.

If we assume that no light enters waveguide 2 so that $a_2(0) = 0$, then the optical powers $P_1(z) \propto |a_1(z)|^2$ and $P_2(z) \propto |a_2(z)|^2$ are

$$P_1(z) = P_1(0)\left[\cos^2\gamma z + \left(\frac{\Delta\beta}{2\gamma}\right)^2\sin^2\gamma z\right] \tag{9.4-10a}$$

$$P_2(z) = P_1(0)\frac{|\mathcal{C}_{21}|^2}{\gamma^2}\sin^2\gamma z. \tag{9.4-10b}$$

Thus, power is exchanged periodically between the two waveguides, as illustrated in Fig. 9.4-6(a). The period is π/γ.

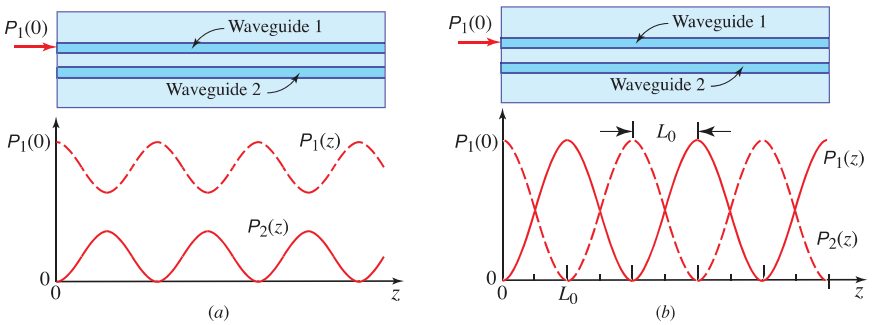


Figure 9.4-6 Periodic exchange of power between waveguides 1 and 2: (a) Phase-mismatched case; (b) phase-matched case.

When the waveguides are identical, i.e., $n_1 = n_2$, $\beta_1 = \beta_2$, and $\Delta\beta = 0$, the two guided waves are said to be phase matched. In this case, $\gamma = \mathcal{C}$, $\mathcal{C}_{12} = \mathcal{C}_{21} = \mathcal{C}$, and the transmission matrix takes the simpler form

$$\mathbf{T} = \begin{bmatrix} A(z) & B(z) \\ C(z) & D(z) \end{bmatrix} = \begin{bmatrix} \cos\mathcal{C}z & -j\sin\mathcal{C}z \\ -j\sin\mathcal{C}z & \cos\mathcal{C}z \end{bmatrix}. \tag{9.4-11}$$

Equations (9.4-10) then simplify to

$$P_1(z) = P_1(0)\cos^2\mathcal{C}z \tag{9.4-12a}$$

$$P_2(z) = P_1(0)\sin^2\mathcal{C}z. \tag{9.4-12b}$$

The exchange of power between the waveguides can then be complete, as illustrated in Fig. 9.4-6(b).

We thus have a device capable of coupling any desired fraction of optical power from one waveguide into another. At a distance $z = L_0 = \pi/2\mathcal{C}$, called the **coupling length** or the **transfer distance**, the power is transferred completely from waveguide 1 into waveguide 2 [Fig. 9.4-7(a)]. At a distance $L_0/2$, half the power is transferred, so that the device acts as a 3-dB coupler, i.e., a 50/50 beamsplitter [Fig. 9.4-7(b)].

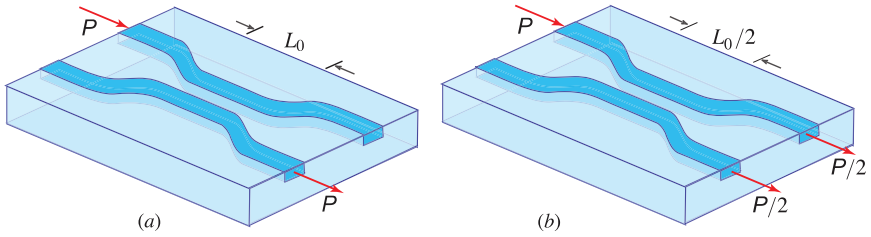


Figure 9.4-7 Optical couplers: (a) switching power from one waveguide to another; (b) a 3-dB coupler.

Switching by Control of Phase Mismatch

A waveguide coupler of fixed length, $L_0 = \pi/2\mathcal{C}$ for example, changes its power-transfer ratio if a small phase mismatch $\Delta\beta$ is introduced. Using (9.4-10b) and (9.4-7), the power-transfer ratio $\mathcal{T} = P_2(L_0)/P_1(0)$ may be written as a function of $\Delta\beta$,

$$\mathcal{T} = \frac{\pi^2}{4}\,\text{sinc}^2\left[\frac{1}{2}\sqrt{1+\left(\frac{\Delta\beta\, L_0}{\pi}\right)^2}\right], \tag{9.4-13}$$

Power-Transfer Ratio

where $\text{sinc}(x) = \sin(\pi x)/(\pi x)$. Figure 9.4-8 illustrates the dependence of the power-transfer ratio $\mathcal{T}$ on the mismatch parameter $\Delta\beta\, L_0$. The ratio achieves a maximum value of unity at $\Delta\beta\, L_0 = 0$, decreases with increasing $\Delta\beta\, L_0$, and then vanishes when $\Delta\beta\, L_0 = \sqrt{3}\,\pi$.

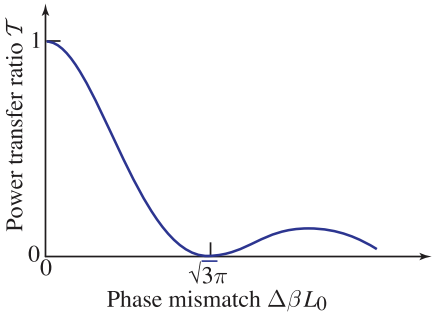


Figure 9.4-8 Dependence of the power transfer ratio $\mathcal{T} = P_2(L_0)/P_1(0)$ on the phase-mismatch parameter $\Delta\beta\, L_0$. The waveguide length is chosen such that for $\Delta\beta = 0$ (the phase-matched case), maximum power is transferred to waveguide 2, i.e., $\mathcal{T} = 1$.

The dependence of the transferred power on the phase mismatch can be utilized in making electrically activated directional couplers. If the mismatch $\Delta\beta\, L_0$ is switched between 0 and $\sqrt{3}\,\pi$, the light is switched from waveguide 2 to waveguide 1. Electrical control of $\Delta\beta$ can be achieved if the material of the waveguides is electro-optic (i.e.,

if its refractive index can be altered by applying an electric field). Such devices will be examined in Secs. 21.1D and 24.3B in connection with electro-optic switches.

□ *__Derivation of the Coupled Wave Equations.__ We proceed to derive the differential equations (9.4-4) that govern the amplitudes $\mathfrak{a}_1(z)$ and $\mathfrak{a}_2(z)$ of the coupled modes. When the two waveguides are not interacting they carry optical fields whose complex amplitudes are of the form

$$E_1(y,z) = \mathfrak{a}_1 u_1(y)\ \exp(-j\beta_1 z) \tag{9.4-14a}$$

$$E_2(y,z) = \mathfrak{a}_2 u_2(y)\ \exp(-j\beta_2 z). \tag{9.4-14b}$$

The amplitudes $\mathfrak{a}_1$ and $\mathfrak{a}_2$ are then constant. In the presence of coupling, we assume that the amplitudes $\mathfrak{a}_1$ and $\mathfrak{a}_2$ become functions of z but the transverse functions $u_1(y)$ and $u_2(y)$, and the propagation constants β_1 and β_2, are not altered. The amplitudes $\mathfrak{a}_1$ and $\mathfrak{a}_2$ are assumed to be slowly varying functions of z in comparison with the distance β^{-1} (the inverse of the propagation constant, β_1 or β_2), which is of the order of magnitude of the wavelength of light.

The presence of waveguide 2 is regarded as a perturbation of the medium outside waveguide 1 in the form of a slab of refractive index $n_2 - n$ and width d at a distance $2a$. The excess refractive index $(n_2 - n)$ and the field E_2 correspond to an excess polarization density $P = (\epsilon_2 - \epsilon)E_2 = \epsilon_o(n_2^2 - n^2)E_2$, which creates a source of optical radiation into waveguide 1 [see (5.2-25)] $\mathcal{S}_1 = -\mu_o \partial^2 \mathcal{P}/\partial t^2$ with complex amplitude

$$\begin{aligned} S_1 &= \mu_o \omega^2 P = \mu_o \omega^2 \epsilon_o \left(n_2^2 - n^2\right) E_2 = \left(n_2^2 - n^2\right) k_o^2 E_2 \\ &= \left(k_2^2 - k^2\right) E_2. \end{aligned} \tag{9.4-15}$$

Here ϵ_2 and ϵ are the electric permittivities associated with the refractive indices n_2 and n, respectively, and $k_2 = n_2 k_o$. This source is present only in the slab of waveguide 2.

To determine the effect of such a source on the field in waveguide 1, we write the Helmholtz equation in the presence of a source as

$$\nabla^2 E_1 + k_1^2 E_1 = -S_1 = -\left(k_2^2 - k^2\right) E_2. \tag{9.4-16a}$$

We similarly write the Helmholtz equation for the wave in waveguide 2 with a source generated as a result of the field in waveguide 1,

$$\nabla^2 E_2 + k_2^2 E_2 = -S_2 = -\left(k_1^2 - k^2\right) E_1, \tag{9.4-16b}$$

where $k_1 = n_1 k_o$. Equations (9.4-16) are two coupled partial different equations that we solve to determine E_1 and E_2. This type of perturbation analysis is valid only for weakly coupled waveguides.

We now write $E_1(y,z) = \mathfrak{a}_1(z)\, e_1(y,z)$ and $E_2(y,z) = \mathfrak{a}_2(z)\, e_2(y,z)$, where $e_1(y,z) = u_1(y)\exp(-j\beta_1 z)$ and $e_2(y,z) = u_2(y)\exp(-j\beta_2 z)$ and note that e_1 and e_2 must satisfy the Helmholtz equations,

$$\nabla^2 e_1 + k_1^2 e_1 = 0 \tag{9.4-17a}$$

$$\nabla^2 e_2 + k_2^2 e_2 = 0, \tag{9.4-17b}$$

where $k_1 = n_1 k_o$ and $k_2 = n_2 k_o$ for points inside the slabs of waveguides 1 and 2, respectively, and $k_1 = k_2 = n k_o$ elsewhere. Substituting $E_1 = \mathfrak{a}_1 e_1$ into (9.4-16a), we obtain

$$\frac{d^2 \mathfrak{a}_1}{dz^2} e_1 + 2\frac{d\mathfrak{a}_1}{dz}\frac{de_1}{dz} = -\left(k_2^2 - k^2\right) \mathfrak{a}_2 e_2. \tag{9.4-18}$$

Noting that $\mathfrak{a}_1$ varies slowly, whereas e_1 varies rapidly with z, we neglect the first term of (9.4-18) in comparison with the second. The ratio between these terms is $[(d\Psi/dz)e_1]/[2\Psi de_1/dz] = [(d\Psi/dz)e_1]/[2\Psi(-j\beta_1 e_1)] = j(d\Psi/\Psi)/2\beta_1\, dz$ where $\Psi = d\mathfrak{a}_1/dz$. The approximation is valid if $d\Psi/\Psi \ll \beta_1\, dz$, i.e., if the variation in $\mathfrak{a}_1(z)$ is slow in comparison with the length β_1^{-1}.

We proceed by substituting $e_1 = u_1 \exp(-j\beta_1 z)$ and $e_2 = u_2 \exp(-j\beta_2 z)$ into (9.4-18). Neglecting the first term leads to

$$2\frac{d\mathfrak{a}_1}{dz}(-j\beta_1)\, u_1(y)\, e^{-j\beta_1 z} = -\left(k_2^2 - k^2\right) \mathfrak{a}_2\, u_2(y)\, e^{-j\beta_2 z}. \tag{9.4-19}$$

Multiplying both sides of (9.4-19) by $u_1(y)$, integrating with respect to y, and using the fact that $u_1^2(y)$ is normalized so that its integral is unity, we finally obtain

$$\frac{d\mathrm{a}_1}{dz}\, e^{-j\beta_1 z} = -j\mathcal{C}_{21}\,\mathrm{a}_2(z)\, e^{-j\beta_2 z}, \tag{9.4-20}$$

where $\mathcal{C}_{21}$ is given by (9.4-6). A similar equation is obtained by repeating the procedure for waveguide 2. These equations yield the coupled differential equations (9.4-4). ■

*C. Waveguide Arrays

The foregoing analysis of light propagation in a pair of weakly coupled waveguides, as presented in Sec. 9.4B, may be generalized to light propagation in waveguide arrays. Consider an array of N identical parallel slab waveguides separated by equal distances, under the assumption that the coupling is sufficiently weak so that only next-neighbor coupling is significant.

If $\mathrm{a}_n(z)$ represents the complex amplitude of light in the nth waveguide, then the set of N coupled-mode equations can be written as

$$\frac{d\mathrm{a}_n}{dz} = -j\mathcal{C}\,(\mathrm{a}_{n+1} + \mathrm{a}_{n-1}), \quad n = 1, \ldots, N, \tag{9.4-21}$$

where $\mathcal{C}$ is the coupling coefficient and $\mathrm{a}_0 = \mathrm{a}_{N+1} = 0$. If the amplitudes are represented by a vector $\mathbf{a}$ of dimension N, comprising the elements $\{\mathrm{a}_n\}$, then (9.4-21) may be expressed in matrix form as $d\mathbf{a}/dz = -j\mathbf{H}\mathbf{a}$, where $\mathbf{H}$ is an $N \times N$ matrix whose elements are $H_{nm} = \mathcal{C}$ for $m = n \pm 1$, and zero otherwise. The solution to this equation is $\mathbf{a}(z) = \mathbf{T}\,\mathbf{a}(0)$, where $\mathbf{T} = \exp(-jz\mathbf{H})$ is the transmission matrix.

The transmission of light through such an array is best described in terms of modes (see Appendix C). The modes of the waveguide array, known as **supermodes**, are to be distinguished from the modes of individual isolated waveguides. The supermodes are determined by diagonalizing the matrix $\mathbf{H}$. This $N \times N$ matrix has N eigenvalues λ_r, and corresponding eigenvectors $\mathbf{b}_r$ comprising the elements $\{\mathrm{b}_{rn}\}$, given by

$$\lambda_r = 2\mathcal{C}\cos\left(\frac{r\pi}{N+1}\right), \quad \mathrm{b}_{rn} = \sqrt{\frac{2}{N+1}}\,\sin\left(\frac{r\pi n}{N+1}\right), \quad r = 1, 2, \ldots, N. \tag{9.4-22}$$

The related transmission matrix $\mathbf{T} = \exp(-jz\mathbf{H})$ has eigenvalues $\exp(-j\lambda_r z)$ and the same eigenvectors $\mathbf{b}_r$ corresponding to N modes. If the initial amplitudes $\{\mathrm{a}_n(0)\}$ are equal to the amplitudes $\{\mathrm{b}_{rn}\}$ of the rth mode, they evolve in accordance with the simple relation $\mathrm{a}_n(z) = \mathrm{a}_n(0)e^{-j\lambda_r z}$, independent of the other modes. The associated optical fields then propagate with a single propagation constant $\beta_r = \beta_0 + \lambda_r$, where β_0 is the propagation constant in an isolated waveguide. Since $-2\mathcal{C} \le \lambda_r \le 2\mathcal{C}$, the propagation constants of the modes lie in the range $\beta_0 - 2\mathcal{C} \le \beta_r \le \beta_0 + 2\mathcal{C}$.

An arbitrary input distribution $\{\mathrm{a}_n(0)\}$ to the waveguide array can be expressed as a superpositions of the modes, $\mathrm{a}_n(0) = \sum_{r=1}^{N} w_r \mathrm{b}_{rn}$, where $w_r = \sum_{n=1}^{N} \mathrm{a}_n(0)\mathrm{b}_{rn}$ are the superposition weights. The amplitudes at a distance z are then given by

$$\mathrm{a}_n(z) = \sum_{r=1}^{N} w_r \mathrm{b}_{rn} e^{-j\lambda_r z}. \tag{9.4-23}$$

The modal analysis enables us to determine $\mathrm{a}_n(z)$ for arbitrary $\mathrm{a}_n(0)$.

EXAMPLE 9.4-1. ***Supermodes of Two Coupled Waveguides.*** For an array of two coupled waveguides ($N = 2$), $\mathbf{H} = \begin{bmatrix} 0 & \mathcal{C} \\ \mathcal{C} & 0 \end{bmatrix}$ and the transmission matrix $\mathbf{T} = \exp(-jz\mathbf{H}) = \begin{bmatrix} \cos \mathcal{C}z & -j \sin \mathcal{C}z \\ -j \sin \mathcal{C}z & \cos \mathcal{C}z \end{bmatrix}$, which reproduces (9.4-11). The two modes have eigenvalues $\lambda_r = \pm\mathcal{C}$ and corresponding propagation constants $\beta_0 \pm \mathcal{C}$. The eigenvectors are $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, corresponding to equal or opposite excitation of the two waveguides. An input field at a single waveguide, say $\mathbf{a}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, excites both supermodes, which have different propagation constants, and the result is an exchange of power between the two waveguides.

Periodic Waveguides

In the limit of large N, the waveguide array may be regarded as a periodic medium and the theory presented in Sec. 7.2 may be readily applied. In particular, it is instructive to compare the dispersion diagram for light propagation in an isolated slab dielectric waveguide to that for light propagation in an array with a finite number of parallel slab waveguides, and to a periodic dielectric medium comprising an infinite set of such waveguides. These diagrams are presented in Fig. 9.4-9. In the single-slab waveguide [Fig. 9.4-9(a)], light travels in modes, each with a dispersion curve lying in the region between the light lines $\omega = c_1\beta$ and $\omega = c_2\beta$. At any frequency, there is at least one mode. In an array of N waveguides [Fig. 9.4-9(b)], each dispersion curve splits into N curves, representing the supermodes. The shapes of these curves are dependent on the coupling coefficient $\mathcal{C}$, which is frequency dependent in accordance with (9.4-6). In the periodic waveguide [Fig. 9.4-9(c)], the dispersion curves broaden into bands that lie between the light lines, and the bands are separated by photonic bandgaps.

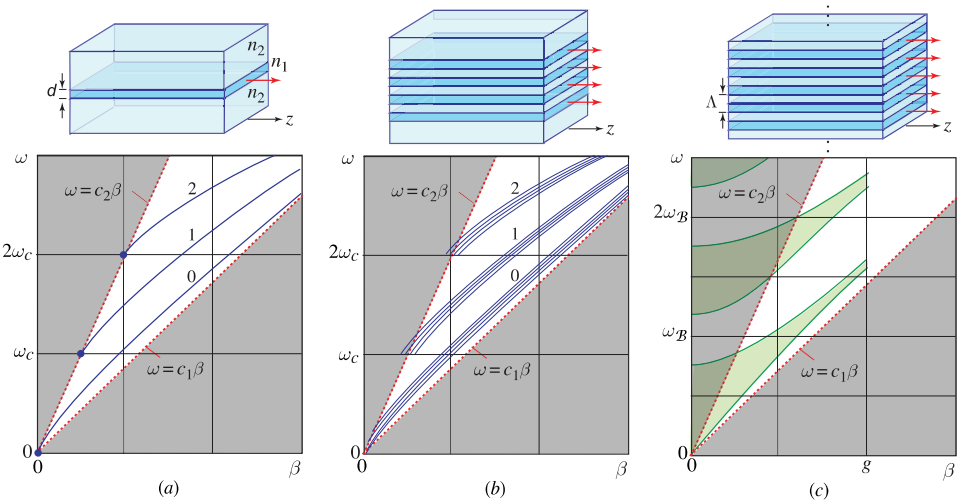


Figure 9.4-9 (a) Dispersion diagram of a slab waveguide with cutoff angular frequency $\omega_c = (\pi/d)(c_o/\mathrm{NA})$, as displayed in Fig. 9.2-8(a). (b) Dispersion diagram of the supermodes of a waveguide array. (c) Dispersion diagram of a periodic waveguide with period Λ, spatial frequency $g = 2\pi/\Lambda$, and Bragg angular frequency $\omega_B = (\pi/\Lambda)(c_o/\bar{n})$, as shown in Fig. 7.2-7. Here, the waves travel in the z direction, which is parallel to the layers, and the higher-index medium is denoted n_1. In Fig. 7.2-7, in contrast, the direction parallel to the layers is the x direction and the higher-index medium is denoted n_2.

9.5 PHOTONIC-CRYSTAL WAVEGUIDES

Bragg-Grating Waveguide

We have seen earlier in this chapter that light may be guided by bouncing between two parallel reflectors — e.g., planar mirrors as described in Sec. 9.1; or planar dielectric boundaries at which the light undergoes total internal reflection, as described in Sec. 9.2. Alternatively, Bragg grating reflectors (see Sec. 7.1C) may be used to guide light, as illustrated in Fig. 9.5-1. The Bragg grating reflector (BGR) is a stack of alternating dielectric layers that has special angle- and frequency-dependent reflectance. For a given angle, the reflectance is close to unity at frequencies within a stop band. Similarly, at a given frequency, the reflectance is close to unity within a range of angles, but omnidirectional reflection is also possible. Thus, a wave with a given frequency can be guided through the waveguide by repeated reflections within a range of bounce angles. Within this angular range, the self-consistency condition is satisfied at a discrete set of angles, each corresponding to a propagating mode. The field distribution of a propagating mode is confined principally to the slab; decaying (evanescent) tails reach into the adjacent grating layers, as illustrated in Fig. 9.5-1.

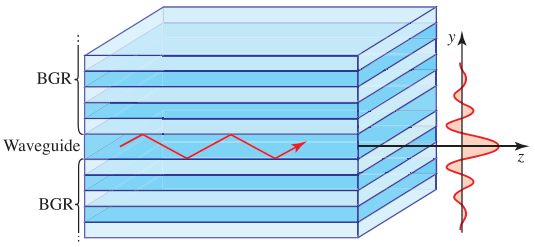


Figure 9.5-1 Planar waveguide comprising a dielectric slab sandwiched between two Bragg-grating reflectors (BGRs).

Bragg-Grating Waveguide as a Photonic Crystal with a Defect Layer

If the upper and lower gratings of a Bragg-grating waveguide are identical, and the slab thickness is comparable to the thickness of the periodic layers constituting the gratings, then the entire medium may be regarded as a 1D periodic structure, i.e., a 1D photonic crystal, but with a **defect**. For example, the device shown in Fig. 9.5-1 is periodic everywhere except for the slab, which is a layer of different thickness and/or different refractive index; the slab may therefore be viewed as a "defective" layer. As described in Sec. 7.2, a perfect photonic crystal has a dispersion relation, or energy-band diagram, containing bandgaps within which no propagating modes exist. In the presence of the "defective" layer, however, a mode whose frequency lies within the bandgap may exist, but it is confined primarily within that layer. Such a mode corresponds to a frequency in the dispersion diagram that lies within the photonic bandgap, as illustrated in Fig. 9.5-2. Such a frequency is the analog of a defect energy level that lies within the bandgap of a semiconductor crystal.

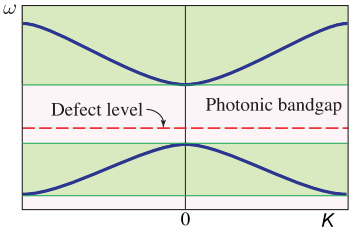


Figure 9.5-2 Dispersion diagram of a photonic crystal with a defect layer.

2D Photonic-Crystal Waveguides

Waveguides may also be created by introducing a path of defects in a 2D photonic crystal. In the example illustrated in Fig. 9.5-3(a), a 2D photonic crystal comprising a set of parallel cylindrical holes, placed in a dielectric material at the points of a periodic triangular lattice, exhibits a complete photonic bandgap for waves traveling along directions parallel to the plane of periodicity (normal to the cylindrical holes). The defect waveguide may take the form of a line of absent holes. A wave entering the waveguide at frequencies within the photonic bandgap does not leak into the surrounding periodic media so that the light is guided through the waveguide. A schematic profile of the propagating mode is illustrated in Fig. 9.5-3(a).

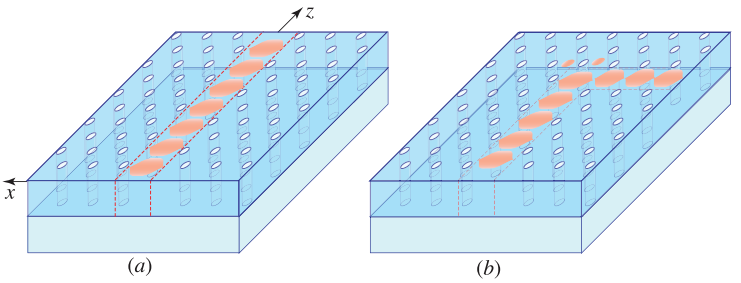


Figure 9.5-3 (a) Propagating mode in a photonic-crystal waveguide. (b) An L-shaped photonic-crystal waveguide.

Moreover, because of the omnidirectional nature of the photonic bandgap, light may be guided through photonic-crystal waveguides with sharp bends and corners without losing energy into the surrounding medium, as illustrated by the L-shaped waveguide configuration shown in Fig. 9.5-3(b). Such behavior is not possible with conventional dielectric waveguides based on total internal reflection.

9.6 PLASMONIC WAVEGUIDES

As demonstrated earlier in this chapter, it is difficult to confine a propagating wave to dimensions much smaller than its wavelength (see also Sec. 4.4D). For the perfect-mirror waveguide described in Sec. 9.1, it is demonstrated in Fig. 9.6-1(a) that a wave of wavelength λ can be guided if the mirror separation $d > \lambda/2$, but it cannot be guided if d is smaller. For the dielectric waveguide described in Sec. 9.2, it was shown that if the slab width d is reduced below $\lambda/2$, only a single guided mode is supported, and as d is further reduced, the guided wave spreads substantially outside the slab and into the surrounding dielectric medium (see Fig. 9.2-5).

Light can, however, be confined and guided at subwavelength spatial scales by making use of metallic structures. As shown in Sec. 8.2B, a metal–dielectric boundary supports a *surface* wave, called a surface plasmon polariton (SPP), that is tightly confined to the boundary, with penetration depths on both sides of the boundary that are much smaller than the wavelength [see Fig. 9.6-1(b)]. It will become apparent below that an optical wave may be guided within an ultrathin dielectric slab embedded in metal cladding if its width is much smaller than the wavelength, as illustrated in Fig. 9.6-1(c). Under these conditions, the SPP waves at the two boundaries couple with each other and combine into modes that extend through the dielectric medium. Similarly, an ultrathin metallic film can guide an optical wave of subwavelength scale [Fig. 9.6-1(d)].

Other metal–dielectric structures with more complex configurations may also be constructed to guide light through various optical circuits. As discussed in Sec. 8.2, this branch of integrated photonics is called **plasmonics**. The propagation lengths of plasmonic waveguides are limited by metallic losses.

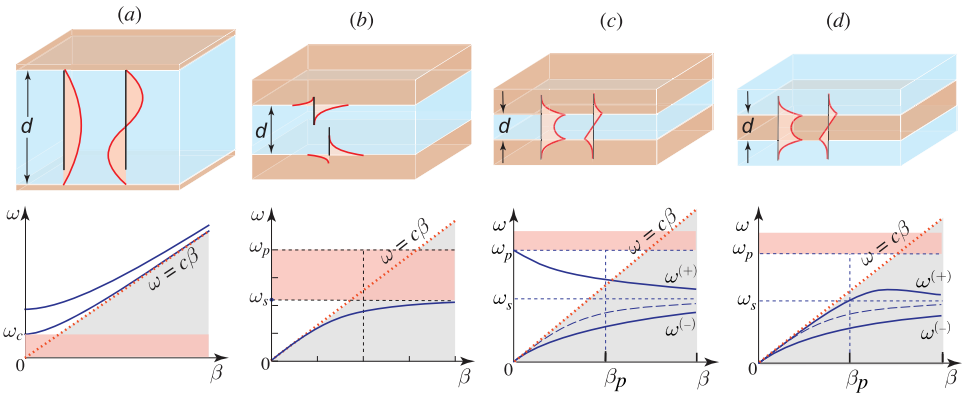


Figure 9.6-1 Configurations and dispersion relations for various optical and plasmonic waveguides. (a) A perfect-mirror waveguide supports optical guided modes if its width $d > \lambda/2$, i.e., if $\omega > \omega_c$ where $\omega_c = \pi c/d$ is the cutoff frequency. (b) A metal–dielectric–metal waveguide of width $d < \lambda/2$ does not support optical guided modes, but does support independent surface plasmon polariton (SPP) waves at the two boundaries if $\omega < \omega_s$, where $\omega_s = \omega_p/\sqrt{1+\epsilon_{r1}}$, ω_p is the plasma frequency of the metal, and $\epsilon_{r1} = \epsilon_1/\epsilon_o$ is the relative permittivity of the dielectric. (c) A metal–insulator–metal (MIM) waveguide of width $d \ll \lambda$ supports a symmetric guided mode $\omega^{(+)}$ for $\omega_s < \omega < \omega_p$, and an anti-symmetric mode $\omega^{(-)}$ for $\omega < \omega_s$. (d) A thin-metal slab of width $d \ll \lambda$, called a metal-slab waveguide, supports two guided modes, one below ω_s and the other that extends slightly above ω_s. For the plots in (c) and (d), $d = \lambda_p/10$, where $\lambda_p = 2\pi c_o/\omega_p$, and $\epsilon_{r1} = 2.25$, so that $\omega_s = 0.55\,\omega_p$. The dashed blue curves represent the dispersion relation for the single-boundary SPP wave. In (a)–(d) the dotted red lines are the dielectric-medium light lines $\omega = c\beta$.

Metal–Insulator–Metal Waveguide

A dielectric slab surrounded by metal claddings forms a metal–dielectric–metal, or **metal–insulator–metal** (**MIM**) waveguide. If the slab thickness is greater than twice the penetration depth of the SPP waves at the boundaries [see (8.2-22)], the structure supports two independent SPP waves [Fig. 9.6-1(b)]. For a thinner dielectric slab, these surface waves overlap and couple, splitting the SPP dispersion curve into two branches and thereby creating two distinct guided modes, labeled $\omega^{(-)}$ and $\omega^{(+)}$ in Fig. 9.6-1(c). These modes correspond to anti-symmetric and symmetric field distributions, respectively.

The dispersion relations for these two modes may be derived by matching the boundary conditions at the metal–dielectric interfaces, much as was done for the dielectric waveguide (Sec. 9.2). The result for the TM wave is similar to (9.2-4) for the dielectric waveguide:

$$\tanh\left(\frac{d}{2}\gamma_1\right) = -\frac{\epsilon_1}{\epsilon_2}\frac{\gamma_2}{\gamma_1}, \qquad \coth\left(\frac{d}{2}\gamma_1\right) = -\frac{\epsilon_1}{\epsilon_2}\frac{\gamma_2}{\gamma_1}, \tag{9.6-1}$$

with

$$\gamma_1 = \sqrt{\beta^2 - \omega^2\mu\epsilon_1}\,, \qquad \gamma_2 = \sqrt{\beta^2 - \omega^2\mu\epsilon_2}\,, \tag{9.6-2}$$

where ϵ_1 and ϵ_2 are the permittivities of the dielectric and metal materials, respectively. This dispersion relation is plotted in Fig. 9.6-1(c) for a metal described by the Drude

model, $\epsilon_2 = \epsilon_o(1 - \omega_p^2/\omega^2)$, where ω_p is the bulk-metal plasma frequency. Note that the upper branch of the dispersion curve crosses the light line, indicating that the wave travels at a phase velocity greater than the velocity of light c in the dielectric medium.

Since the two branches of the dispersion relation extend over the frequency range $0 < \omega < \omega_p$ a wave at any frequency $\omega < \omega_p$ can be guided in dielectric slabs that are significantly smaller than the wavelength. Modes at near-infrared wavelengths, for example, can be localized at the nanometer scale.

Metal-Slab Waveguide

Similarly, a thin metallic film of width $d \ll \lambda$ embedded in a dielectric medium can serve as a plasmonic waveguide [Fig. 9.6-1(d)]. If the film thickness is smaller than the penetration depth of the SPP waves from the boundaries into the metal, then these waves overlap and coalesce into two distinct waveguide modes. Again, the dispersion relation may be obtained by matching the boundary conditions, yielding

$$\tanh\left(\frac{d}{2}\gamma_2\right) = -\frac{\epsilon_1}{\epsilon_2}\frac{\gamma_2}{\gamma_1}, \qquad \coth\left(\frac{d}{2}\gamma_2\right) = -\frac{\epsilon_1}{\epsilon_2}\frac{\gamma_2}{\gamma_1}, \tag{9.6-3}$$

where γ_1 and γ_2 are given by (9.6-2) and, as before, ϵ_1 and ϵ_2 are the permittivities of the dielectric and metal materials, respectively. The Drude model for the metal is again characterized by $\epsilon_2 = \epsilon_o(1-\omega_p^2/\omega^2)$. Note that (9.6-3) is the same as (9.6-1) except that the subscripts 1 and 2 are interchanged on the left-hand sides of both equations. This dispersion relation is plotted in Fig. 9.6-1(d). As with the MIM waveguide considered above, the dispersion relation for the SPP at a single metal–dielectric boundary splits into two branches, corresponding to symmetric and anti-symmetric modes; in this case both lie below the light line for the bulk dielectric material $\omega = c\beta$.

⋆Periodic Metal–Dielectric Arrays

A periodic structure comprising an array of metal–dielectric slabs functions as a photonic crystal. In analogy with the all-dielectric arrays discussed in Chapter 7, and illustrated in Figs. 9.4-9(a) and (c), the dispersion curve of the single metal–dielectric boundary depicted as the dashed blue curve in Fig. 9.6-2(b) splits to create bands, as shown. The frequency ω of a mode with propagation constant β may only lie within two separated spectral bands, both of which are within the range $\omega < \omega_p$.

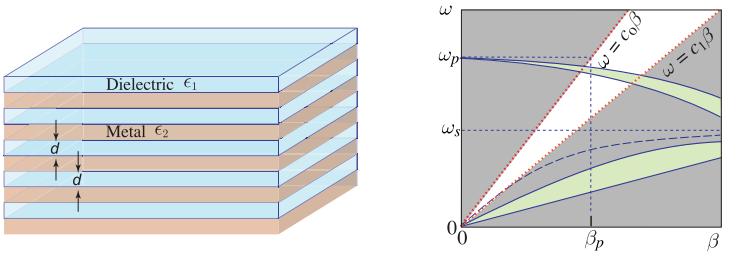


Figure 9.6-2 Metal–insulator periodic structure and its dispersion relation for light traveling in the direction of the layers. Two bands lie within the $\omega < \omega_p$ frequency range. The dashed blue curve in (b) is the dispersion relation for the single-boundary SPP wave and the dotted red lines are the free-space and dielectric-medium light lines.

READING LIST

Books and Seminal Articles

See also the reading list on layered and periodic media in Chapter 7, the reading list in Chapter 10, and the reading list on photonic integrated circuits in Chapter 25.

S. Bhadra and A. Ghatak, eds., *Guided Wave Optics and Photonic Devices*, CRC Press/Taylor & Francis, 2013.

L. N. Binh, *Guided Wave Photonics: Fundamentals and Applications with* MATLAB, CRC Press/Taylor & Francis, 2012.

D. Dai, J. Bauters, and J. E. Bowers, Passive Technologies for Future Large-Scale Photonic Integrated Circuits on Silicon: Polarization Handling, Light Non-Reciprocity and Loss Reduction, *Light: Science & Applications* (2012) **1**, e1; doi:10.1038/lsa.2012.1

R. G. Hunsperger, *Integrated Optics: Theory and Technology*, Springer-Verlag, 1982, 6th ed. 2010.

J. Bures, *Guided Optics*, Wiley–VCH, 2009.

C.-L. Chen, *Foundations for Guided Wave Optics*, Wiley, 2007.

K. Iga and Y. Kokobun, eds., *Encyclopedic Handbook of Integrated Optics*, CRC Press/Taylor & Francis, 2006.

K. Okamoto, *Fundamentals of Optical Waveguides*, Elsevier, 2nd ed. 2005.

J.-M. Liu, *Photonic Devices*, Cambridge University Press, 2005, paperback ed. 2009.

G. Lifante, *Integrated Photonics: Fundamentals*, Wiley, 2003.

C. Pollock and M. Lipson, *Integrated Photonics*, Kluwer, 2003.

A. A. Barybin and V. A. Dmitriev, *Modern Electrodynamics and Coupled-Mode Theory: Application to Guided-Wave Optics*, Rinton Press, 2002.

K. Iizuka, *Elements of Photonics, Volume 2: For Fiber and Integrated Optics*, Wiley, 2002.

M. Young, *Optics and Lasers: Including Fibers and Optical Waveguides*, Springer-Verlag, 5th ed. 2000.

A. R. Mickelson, *Guided Wave Optics*, Springer-Verlag, 1993, paperback ed. 2012.

K. J. Ebeling, *Integrated Optoelectronics: Waveguide Optics, Photonics, Semiconductors*, Springer-Verlag, 1993, paperback ed. 2011.

D. G. Hall, ed., *Selected Papers on Coupled Mode Theory in Guided-Wave Optics*, SPIE Optical Engineering Press (Milestone Series Volume 84), 1993.

D. Marcuse, *Theory of Dielectric Optical Waveguides*, Academic Press, 1974, 2nd ed. 1991.

PROBLEMS

9.1-3 **Field Distribution.**

(a) Demonstrate that a single TEM plane wave $E_x(y,z)$=$A\exp(-jk_y y)\exp(-j\beta z)$ cannot satisfy the boundary conditions, $E_x(\pm d/2, z) = 0$ at all z, for the mirror waveguide illustrated in Fig. 9.1-1.

(b) Show that the sum of two TEM plane waves written as $E_x(y,z) = A_1\exp(-jk_{y1}y)\exp(-j\beta_1 z) + A_2\exp(-jk_{y2}y)\exp(-j\beta_2 z)$ does not satisfy the boundary conditions if $A_1 = \pm A_2$, $\beta_1 = \beta_2$, and $k_{y1} = -k_{y2} = m\pi/d$ where $m = 1, 2, \ldots$.

9.1-4 **Modal Dispersion.** Light of wavelength $\lambda_o = 0.633$ μm is transmitted through a mirror waveguide of mirror separation $d = 10$ μm and $n = 1$. Determine the number of TE and TM modes. Determine the group velocities of the fastest and the slowest modes. If a narrow pulse of light is carried by all modes for a distance 1 m in the waveguide, how much does the pulse spread as a result of the differences of the group velocities?

9.2-3 **Parameters of a Dielectric Waveguide.** Light of free-space wavelength $\lambda_o = 0.87$ μm is guided by a thin planar film of width $d = 2$ μm and refractive index $n_1 = 1.6$ surrounded by a medium of refractive index $n_2 = 1.4$.

(a) Determine the critical angle θ_c and its complement $\overline{\theta}_c$, the numerical aperture NA, and the maximum acceptance angle for light originating in air ($n = 1$).

(b) Determine the number of TE modes.

(c) Determine the bounce angle θ and the group velocity v of the $m = 0$ TE mode.

9.2-4 **Effect of Cladding.** Redo Prob. 9.2-3 under the proviso that the thin film is suspended in air ($n_2 = 1$). Compare the results.

9.2-5 **Field Distribution.** The transverse distribution $u_m(y)$ of the electric-field complex amplitude of a TE mode in a slab waveguide is given by (9.2-10) and (9.2-13). Derive an expression for the ratio of the proportionality constants. Plot the distribution of the $m = 0$ TE mode for a slab waveguide with parameters $n_1 = 1.48$, $n_2 = 1.46$, $d = 0.5$ μm, and $\lambda_o = 0.85$ μm, and determine its confinement factor (percentage of power in the slab).

9.2-6 **Derivation of the Field Distributions Using Maxwell's Equations.** Assuming that the electric field in a symmetric dielectric waveguide is harmonic within the slab and exponential outside the slab and has a propagation constant β in both media, we may write $E_x(y, z) = u(y)\, e^{-j\beta z}$, where

$$u(y) = \begin{cases} A\cos(k_y y + \varphi), & -d/2 \le y \le d/2 \\ B\exp(-\gamma y), & y > d/2 \\ B\exp(\gamma y), & y < d/2. \end{cases}$$

Satisfying the Helmholtz equation requires $k_y^2 + \beta^2 = n_1^2 k_o^2$ and $-\gamma^2 + \beta^2 = n_2^2 k_o^2$. Use Maxwell's equations to derive expressions for $H_y(y, z)$ and $H_z(y, z)$. Show that the boundary conditions are satisfied if β, γ, and k_y take the values β_m, γ_m, and k_{ym} derived in the text and verify the self-consistency condition (9.2-4).

9.2-7 **Single-Mode Waveguide.** What is the largest thickness d of a planar symmetric dielectric waveguide with refractive indices $n_1 = 1.50$ and $n_2 = 1.46$ for which there is only one TE mode at $\lambda_o = 1.3$ μm? What is the number of modes if a waveguide with this thickness is used at $\lambda_o = 0.85$ μm instead?

9.2-8 **Mode Cutoff.** Show that the cutoff condition for TE mode $m > 0$ in a symmetric slab waveguide with $n_1 \approx n_2$ is approximately $\lambda_o^2 \approx 8 n_1 \Delta n\, d^2/m^2$, where $\Delta n = n_1 - n_2$.

9.2-9 **TM Mode Bounce Angles.** Derive an expression for the bounce angles of the TM modes similar to (9.2-4). Generate a plot similar to Fig. 9.2-2 for TM modes in a waveguide with $\sin\overline{\theta}_c = 0.3$ and $\lambda/2d = 0.1$. What is the number of TM modes?

9.3-1 **Modes of a Rectangular Dielectric Waveguide.** A rectangular dielectric waveguide has a square cross section of area 10^{-2} mm^2 and numerical aperture NA $= 0.1$. Use (9.3-3) to plot the number of TE modes as a function of frequency ν. Compare your results with Fig. 9.2-4.

9.4-1 **Coupling Coefficient Between Two Slabs.**

(a) Use (9.4-6) to determine the coupling coefficient between two *identical* slab waveguides of width $d = 0.5$ μm, spacing $2a = 1.0$ μm, and refractive indices $n_1 = n_2 = 1.48$, in a medium of refractive index $n = 1.46$, at $\lambda_o = 0.85$ μm. Assume that both waveguides are operating in the $m = 0$ TE mode and use the results of Prob. 9.2-5 to determine the transverse distributions.

(b) Determine the length of the waveguides that makes the device act as a 3-dB coupler.

*9.6-1 **Silver-Slab Waveguide.** Calculate and plot the dispersion relation for the symmetric and anti-symmetric modes of a silver-slab waveguide of thickness $d = 20$ nm in a host medium of frequency-independent refractive index $n_2 = 2$. Assume that the permittivity of silver is described by the Drude model (8.2-18), with a plasma frequency ω_p corresponding to a free-space wavelength of 138 nm. Determine the properties (velocity, propagation wavelength, and penetration depths) of the symmetric mode at a free-space wavelength of 400 nm. Compare these properties with those of a SPP propagating on a single interface between silver and the same host medium.

CHAPTER

10

FIBER OPTICS

10.1 GUIDED RAYS 393
- A. Step-Index Fibers
- B. Graded-Index Fibers

10.2 GUIDED WAVES 397
- A. Step-Index Fibers
- B. Single-Mode Fibers
- *C. Quasi-Plane Waves in Step-Index and Graded-Index Fibers
- D. Multicore Fibers and Fiber Couplers

10.3 ATTENUATION AND DISPERSION 415
- A. Attenuation
- B. Dispersion

10.4 HOLEY AND PHOTONIC-CRYSTAL FIBERS 426

10.5 FIBER MATERIALS 429

Working at Corning in the early 1970s, **Peter C. Schultz (born 1942)**, left, **Donald B. Keck (born 1941)**, center, and **Robert D. Maurer (born 1924)**, right, developed ultra-low-loss silica-glass optical fibers that permitted light to propagate over exceptionally long distances, thereby paving the way for worldwide optical fiber communications. Billions of kilometers of optical fiber span the globe.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Fiber optics is an enabling technology for telecommunications, data transmission, and information science. The availability of ultra-low-loss optical fibers is, in large part, responsible for the commercial viability of optical fiber communications.

An optical fiber is a cylindrical dielectric waveguide fabricated from a low-loss material such as silica glass. It has a central **core** in which the light is guided, embedded in an outer **cladding** of slightly lower refractive index (Fig. 10.0-1). Light rays in the core incident on the core–cladding boundary at angles greater than the critical angle undergo total internal reflection and are thereby guided through the core without refraction into the cladding and without loss. Rays at greater inclination to the fiber axis lose a portion of their power into the cladding at each reflection and are not guided.

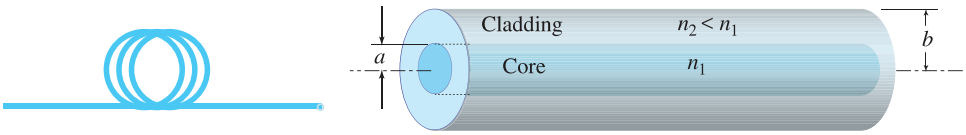


Figure 10.0-1 An optical fiber is a cylindrical dielectric waveguide with an inner core and an outer cladding of refractive index lower than that of the core.

Technological advances in the fabrication of optical fibers over the past several decades allow light to be guided through 1 km of silica-glass fiber with a loss as low as ≈ 0.15 dB ($\approx 3.4\%$) at the wavelength of maximum transparency. Because of this low loss, silica-glass optical fibers long ago replaced copper coaxial cables as the preferred transmission medium for terrestrial and sub-oceanic voice as well as for data communications. In recent years, optical fibers have transcended their monolithic silica-glass origins and have come to play a central role in the arenas of sensing, security, transportation, defense, and biomedicine. This has been facilitated by the development of photonic-crystal, specialty, multimaterial, and multifunctional optical fibers.

In this chapter we introduce the principles of light transmission in optical fibers. These principles are essentially the same as those applicable to planar dielectric waveguides (Chapter 9); the most notable distinction is that optical fibers have cylindrical geometry. In both types of waveguide, light propagates in the form of modes. Each mode travels along the axis of the waveguide with a distinct propagation constant and group velocity, maintaining its transverse spatial distribution and polarization. When the core diameter is small, only a single mode is supported and the optical fiber is said to be a **single-mode fiber**.

Optical fibers with large core diameters are **multimode fibers**. One of the difficulties associated with the propagation of light in multimode fibers arises from the differences among the group velocities of the modes. This results in a spread of travel times and leads to the broadening of a light pulse as it travels through the fiber. This effect, called **modal dispersion**, limits the rate at which adjacent pulses can be launched without resulting in pulse overlap at the far end of the fiber. Modal dispersion therefore limits the speed at which multimode optical fiber communication systems can operate.

Modal dispersion can be reduced by grading the refractive index of the fiber core from a maximum value at its center to a minimum value at the core–cladding boundary. The fiber is then called a **graded-index fiber**, or GRIN fiber, whereas conventional fibers with constant refractive indices in the core and the cladding are known as **step-index fibers**. In a graded-index fiber the travel velocity increases with radial distance from the core axis (since the refractive index decreases). Although rays of greater inclination to the fiber axis must travel farther, they travel faster. This permits the travel times of the different modes to be equalized.

Optical fibers are thus classified as step-index or graded-index, and multimode or single-mode, as illustrated in Fig. 10.0-2.

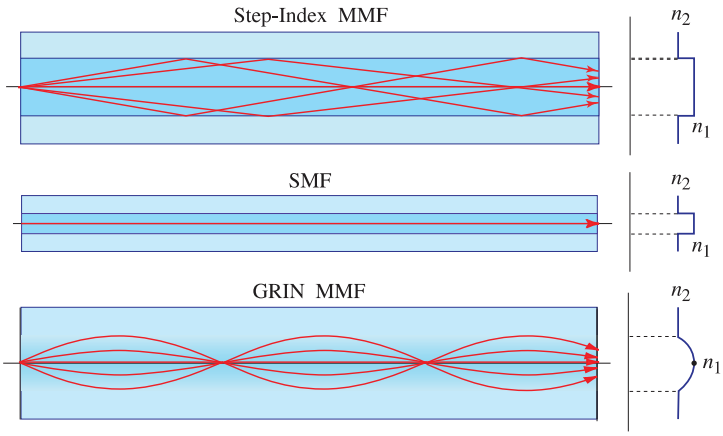


Figure 10.0-2 Geometry, refractive-index profile, and typical rays in a step-index multimode fiber (MMF), a single-mode fiber (SMF), and a graded-index multimode fiber (GRIN MMF).

This Chapter

The chapter begins with ray-optics descriptions of step-index and graded-index fibers (Sec. 10.1). An electromagnetic-optics approach, which highlights the nature of optical modes and single-mode propagation, follows in Sec. 10.2. In the simplified approximate approach set forth in Sec. 10.2C, the field is treated as a quasi-plane wave in analogy with the bouncing plane-wave construct for planar dielectric waveguides considered in Sec. 9.2. The optical properties of the fiber material (often fused silica), including attenuation and material dispersion as well as modal, waveguide, polarization-mode, and nonlinear dispersion, are presented in Sec. 10.3. Holey and photonic-crystal fibers, which have more complex refractive-index profiles, along with unusual dispersion characteristics, are introduced in Sec. 10.4. Finally, multimaterial and multifunctional fibers, including mid-infrared and specialty fibers, are considered in Sec. 10.5. We return to a discussion of fiber optics in Chapters 23 and 25, which are devoted to ultrafast optics and optical fiber communications, respectively.

10.1 GUIDED RAYS

A. Step-Index Fibers

A step-index fiber is a cylindrical dielectric waveguide specified by the refractive indices of its core and cladding, n_1 and n_2, respectively, and their radii a and b (see Fig. 10.0-1). Examples of standard core-to-cladding diameter ratios (in units of $\mu\text{m}/\mu\text{m}$) are $2a/2b = 8/125, 50/125, 62.5/125, 85/125$, and $100/140$. The refractive indices of the core and cladding differ only slightly, so that the fractional refractive-index change is small:

$$\Delta \equiv \frac{n_1^2 - n_2^2}{2n_1^2} \approx \frac{n_1 - n_2}{n_1} \ll 1. \tag{10.1-1}$$

Most fibers used in currently implemented optical fiber communication systems are made of fused silica glass (SiO_2) of high chemical purity. Slight changes in the refractive index are effected by adding low concentrations of doping materials (e.g., titanium,

germanium, boron). The refractive index n_1 ranges from 1.44 to 1.46, depending on the wavelength, and Δ typically lies between 0.001 and 0.02.

An optical ray in a step-index fiber is guided by total internal reflections within the fiber core if its angle of incidence at the core–cladding boundary is greater than the critical angle $\theta_c = \sin^{-1}(n_2/n_1)$, and remains so as the ray bounces.

Meridional Rays

Meridional rays, which are rays confined to planes that pass through the fiber axis, have a particularly simple guiding condition, as illustrated in Fig. 10.1-1. These rays intersect the fiber axis and reflect in the same plane without changing their angle of incidence, behaving as if they were in a planar waveguide. Meridional rays are guided if the angle θ they make with the fiber axis is smaller than the complement of the critical angle, i.e., if $\theta < \overline{\theta}_c = \pi/2 - \theta_c = \cos^{-1}(n_2/n_1)$. Since $n_1 \approx n_2$, $\overline{\theta}_c$ is usually small and the guided rays are approximately paraxial.

Figure 10.1-1 The trajectory of a meridional ray lies in a plane that passes through the fiber axis. The ray is guided if $\theta < \overline{\theta}_c = \cos^{-1}(n_2/n_1)$.

Skewed Rays

An arbitrary ray is identified by its plane of incidence, which is a plane parallel to the fiber axis through which the ray passes, and by the angle with that axis, as illustrated in Fig. 10.1-2. The plane of incidence intersects the core–cladding cylindrical boundary at an angle ϕ with respect to the normal to the boundary and lies at a distance R from the fiber axis. The ray is identified by its angle θ with the fiber axis and by the angle ϕ of its plane. When $\phi \neq 0$ ($R \neq 0$) the ray is said to be skewed. For meridional rays $\phi = 0$ and $R = 0$.

A skewed ray reflects repeatedly into planes that make the same angle ϕ with the core–cladding boundary; it follows a helical trajectory confined within a cylindrical shell of inner and outer radii R and a, respectively, as illustrated in Fig. 10.1-2. The projection of the trajectory onto the transverse (x–y) plane is a regular polygon that is not necessarily closed. The condition for a skewed ray to always undergo total internal reflection is that its angle with the z axis be smaller than the complementary critical angle, i.e., $\theta < \overline{\theta}_c$.

Figure 10.1-2 A skewed ray lies in a plane offset from the fiber axis by a distance R. The ray is identified by the angles θ and ϕ. It follows a helical trajectory confined within a cylindrical shell with inner and outer radii R and a, respectively. The projection of the ray on the transverse plane is a regular polygon that is not necessarily closed.

Numerical Aperture

A ray incident from air into the fiber becomes a guided ray if, upon refraction into the core, it makes an angle θ with the fiber axis that is smaller than $\overline{\theta}_c$. As shown in Fig. 10.1-3(a), if Snell's law is applied at the air–core boundary, the angle θ_a in air corresponding to the angle $\overline{\theta}_c$ in the core is obtained from $1 \cdot \sin\theta_a = n_1 \sin\overline{\theta}_c$, which leads to $\sin\theta_a = n_1\sqrt{1-\cos^2\overline{\theta}_c} = n_1\sqrt{1-(n_2/n_1)^2} = \sqrt{n_1^2-n_2^2}$ (see Exercise 1.2-5). The **acceptance angle** of the fiber is therefore

$$\theta_a = \sin^{-1}\text{NA}, \tag{10.1-2}$$

where the numerical aperture (NA) of the fiber is given by

$$\boxed{\text{NA} = \sqrt{n_1^2 - n_2^2} \approx n_1\sqrt{2\Delta}} \tag{10.1-3}$$

Numerical Aperture

since $n_1 - n_2 = n_1\Delta$ and $n_1 + n_2 \approx 2n_1$.

The acceptance angle θ_a of the fiber determines the cone of external rays that are guided by the fiber. Rays incident at angles greater than θ_a are refracted into the fiber but are guided only for a short distance since they do not undergo total internal reflection. The numerical aperture therefore describes the light-gathering capacity of the fiber, as illustrated in Fig. 10.1-3(b).

When the guided rays arrive at the terminus of the fiber, they are refracted back into a cone of angle θ_a. The acceptance angle is thus a crucial design parameter for coupling light into and out of a fiber.

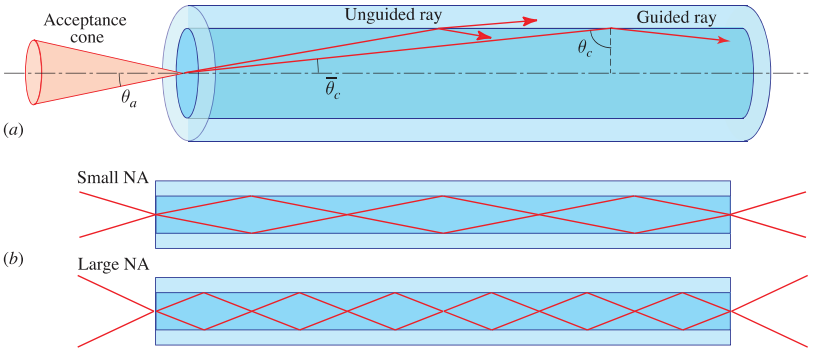


Figure 10.1-3 (a) The acceptance angle θ_a of a fiber. Rays within the acceptance cone are guided by total internal reflection. The numerical aperture $\text{NA} = \sin\theta_a$. The angles θ_a and $\overline{\theta}_c$ are typically quite small; they are exaggerated here for clarity. (b) The light-gathering capacity of a large NA fiber is greater than that of a small NA fiber.

EXAMPLE 10.1-1. ***Cladded and Uncladded Fibers.*** In a silica-glass fiber with $n_1 = 1.46$ and $\Delta = (n_1 - n_2)/n_1 = 0.01$, the complementary critical angle $\overline{\theta}_c = \cos^{-1}(n_2/n_1) = 8.1°$, and the acceptance angle $\theta_a = 11.9°$, corresponding to a numerical aperture $\text{NA} = 0.206$. By comparison, a fiber with silica-glass core ($n_1 = 1.46$) and a cladding with a much smaller refractive index $n_2 = 1.064$ has $\overline{\theta}_c = 43.2°$, $\theta_a = 90°$, and $\text{NA} = 1$. Rays incident from *all* directions are guided since they reflect within a cone of angle $\overline{\theta}_c = 43.2°$ inside the core. Likewise, for an uncladded fiber ($n_2 = 1$), $\overline{\theta}_c = 46.8°$, and rays incident from air at any angle are also refracted into guided rays. Although its light-gathering capacity is high, the uncladded fiber is generally not suitable for use as an optical waveguide because of the large number of modes it supports, as will be explained subsequently.

B. Graded-Index Fibers

Index grading is an ingenious method for reducing the pulse spreading caused by differences in the group velocities of the modes in a multimode fiber. The core of a graded-index (GRIN) fiber has a refractive index that varies; it is highest in the center of the fiber and decreases gradually to its lowest value where the core meets the cladding. The phase velocity of light is therefore minimum at the center and increases gradually with radial distance. Rays of the most axial mode thus travel the shortest distance, but they do so at the smallest phase velocity. Rays of the most oblique mode zigzag at a greater angle and travel a longer distance, but mostly in a medium where the phase velocity is high. The disparities in distances are thus compensated by opposite disparities in the phase velocities. As a consequence, the differences in the travel times associated with a light pulse are reduced. In this section we examine the propagation of light in GRIN fibers.

The core refractive index of a GRIN fiber is a function $n(r)$ of the radial position r. As illustrated in Fig. 10.1-4, the largest value of $n(r)$ is at the core center, $n(0) = n_1$, while the smallest value occurs at the core radius, $n(a) = n_2$. The cladding refractive index is maintained constant at n_2.

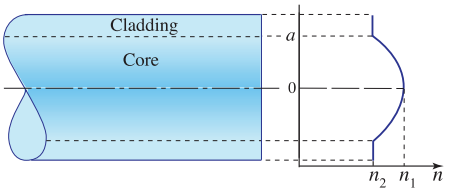


Figure 10.1-4 Geometry and refractive-index profile of a graded-index optical fiber.

A versatile refractive-index profile that exhibits this generic behavior is described by the power-law function

$$n^2(r) = n_1^2 \left[1 - 2\left(\frac{r}{a}\right)^p \Delta\right], \qquad r \le a, \tag{10.1-4}$$

where

$$\Delta = \frac{n_1^2 - n_2^2}{2n_1^2} \approx \frac{n_1 - n_2}{n_1}. \tag{10.1-5}$$

The **grade profile parameter** p determines the steepness of the profile. As illustrated in Fig. 10.1-5, $n^2(r)$ is a linear function of r for $p = 1$ and a quadratic function for $p = 2$. The quantity $n^2(r)$ becomes increasingly steep as p becomes larger, and ultimately approaches a step function for $p \to \infty$. The step-index fiber is thus a special case of the GRIN fiber.

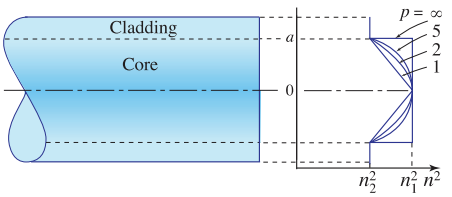


Figure 10.1-5 Power-law refractive-index profile $n^2(r)$ for various values of p.

The transmission of light rays through a GRIN medium with parabolic-index profile was discussed in Sec. 1.3. Rays in meridional planes follow oscillatory planar trajectories, whereas skewed rays follow helical trajectories. For an arbitrary refractive-index

profile, the turning points form cylindrical caustic surfaces, as illustrated in Fig. 10.1-6. Guided rays are confined within the core and do not reach the cladding.

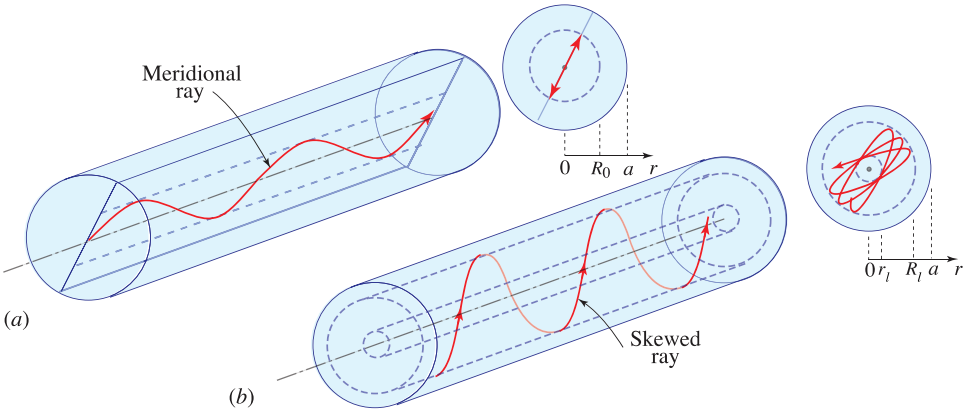


Figure 10.1-6 Guided rays in the core of a GRIN fiber. (a) A meridional ray confined to a meridional plane inside a cylinder of radius R_0. (b) A skewed ray follows a helical trajectory confined within two cylindrical shells of radii r_l and R_l. For a parabolic-index profile, the trajectory projects to a stationary ellipse, as in Fig. 1.3-7.

The numerical aperture of a GRIN optical fiber may be determined by identifying the largest angle of the incident ray that is guided within the GRIN core without reaching the cladding. For meridional rays in a GRIN fiber with parabolic profile, the numerical aperture is given by (10.1-3) (see Exercise 1.3-2).

10.2 GUIDED WAVES

We now proceed to develop an electromagnetic-optics theory of light propagation in fibers. We seek to determine the electric and magnetic fields of guided waves by using Maxwell's equations and the boundary conditions imposed by the cylindrical dielectric core and cladding. As with all waveguides, there are certain special solutions, known as modes (see Appendix C), each of which has a distinct propagation constant, a characteristic field distribution in the transverse plane, and two independent polarization states. Since an exact solution is rather difficult, a number of approximations will be used.

Helmholtz Equation

The optical fiber is a dielectric medium with refractive index $n(r)$. In a step-index fiber, $n(r) = n_1$ in the core ($r < a$) and $n(r) = n_2$ in the cladding ($r > a$). In a GRIN fiber, $n(r)$ is a continuous function in the core and has a constant value $n(r) = n_2$ in the cladding. In either case, we assume that the outer radius b of the cladding is sufficiently large so that it can be taken to be infinite when considering guided light in the core and near the core–cladding boundary.

Each of the components of the monochromatic electric and magnetic fields obeys the Helmholtz equation, $\nabla^2 U + n^2(r)k_o^2 U = 0$, where $k_o = 2\pi/\lambda_o$. This equation is obeyed exactly in each of the two regions of the step-index fiber, and is obeyed approximately within the core of the GRIN fiber if $n(r)$ varies slowly within a wavelength (see

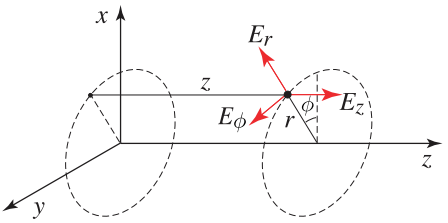


Figure 10.2-1 Cylindrical fiber coordinate system.

Sec. 5.3). In a cylindrical coordinate system (see Fig. 10.2-1) the Helmholtz equation is written as

$$\frac{\partial^2 U}{\partial r^2} + \frac{1}{r}\frac{\partial U}{\partial r} + \frac{1}{r^2}\frac{\partial^2 U}{\partial \phi^2} + \frac{\partial^2 U}{\partial z^2} + n^2 k_o^2 U = 0, \tag{10.2-1}$$

where $U = U(r, \phi, z)$. The guided modes are waves traveling in the z direction with propagation constant β, so that the z dependence of U is of the form $e^{-j\beta z}$. They are periodic in the angle ϕ with period 2π, so that they take the harmonic form $e^{-jl\phi}$, where l is an integer. Substituting

$$U(r, \phi, z) = u(r)e^{-jl\phi}e^{-j\beta z}, \qquad l = 0, \pm 1, \pm 2, \ldots \tag{10.2-2}$$

into (10.2-1) leads to an ordinary differential equation for the radial profile $u(r)$:

$$\frac{d^2 u}{dr^2} + \frac{1}{r}\frac{du}{dr} + \left(n^2(r)k_o^2 - \beta^2 - \frac{l^2}{r^2}\right)u = 0. \tag{10.2-3}$$

A. Step-Index Fibers

As we discovered in Sec. 9.2B, the wave is guided (or bound) if the propagation constant is smaller than the wavenumber in the core ($\beta < n_1 k_o$) and greater than the wavenumber in the cladding ($\beta > n_2 k_o$). It is therefore convenient to define the quantities

$$k_T^2 = n_1^2 k_o^2 - \beta^2 \tag{10.2-4a}$$

and

$$\gamma^2 = \beta^2 - n_2^2 k_o^2, \tag{10.2-4b}$$

so that, for guided waves, k_T^2 and γ^2 are positive and k_T and γ are real. Equation (10.2-3) may then be written in the core and cladding separately:

$$\frac{d^2 u}{dr^2} + \frac{1}{r}\frac{du}{dr} + \left(k_T^2 - \frac{l^2}{r^2}\right)u = 0, \qquad r < a \quad \text{(core)}, \tag{10.2-5a}$$

$$\frac{d^2 u}{dr^2} + \frac{1}{r}\frac{du}{dr} - \left(\gamma^2 + \frac{l^2}{r^2}\right)u = 0, \qquad r > a \quad \text{(cladding)}. \tag{10.2-5b}$$

Equations (10.2-5) are well-known differential equations whose solutions comprise the family of Bessel functions. Excluding functions that approach ∞ at $r = 0$ in the

core, or at $r \to \infty$ in the cladding, we obtain the bounded solutions:

$$u(r) \propto \begin{cases} J_l(k_T r), & r < a \quad \text{(core)} \\ K_l(\gamma r), & r > a \quad \text{(cladding)}, \end{cases} \tag{10.2-6}$$

where $J_l(x)$ is the Bessel function of the first kind and order l, and $K_l(x)$ is the modified Bessel function of the second kind and order l. The function $J_l(x)$ oscillates like the sine or cosine function but with a decaying amplitude. The function $K_l(x)$ decays exponentially at large x. Two representative examples of the radial distribution $u(r)$ are displayed in Fig. 10.2-2.

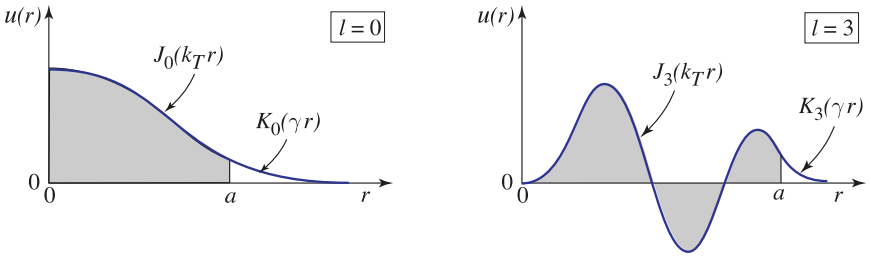


Figure 10.2-2 Examples of the radial distribution $u(r)$ provided in (10.2-6) for $l = 0$ and $l = 3$. The shaded and unshaded areas represent the fiber core and cladding, respectively. The parameters k_T and γ, and the two proportionality constants in (10.2-6), have been selected such that $u(r)$ is continuous and has a continuous derivative at $r = a$. Larger values of k_T and γ lead to a greater number of oscillations in $u(r)$.

The parameters k_T and γ determine the rate of change of $u(r)$ in the core and in the cladding, respectively. A large value of k_T means more oscillation of the radial distribution in the core. A large value of γ means more rapid decay and therefore smaller penetration of the wave into the cladding. As can be seen from (10.2-4), the sum of the squares of k_T and γ is a constant:

$$k_T^2 + \gamma^2 = \left(n_1^2 - n_2^2\right) k_o^2 = (\text{NA})^2 \cdot k_o^2, \tag{10.2-7}$$

so that as k_T increases, γ decreases and the field penetrates more deeply into the cladding. For those values of k_T that exceed $\text{NA} \cdot k_o$, the quantity γ becomes imaginary and the wave ceases to be bound to the core.

Fiber V Parameter

It is convenient to normalize k_T and γ by defining the quantities

$$X = k_T a, \qquad Y = \gamma a. \tag{10.2-8}$$

In view of (10.2-7), we have

$$X^2 + Y^2 = V^2, \tag{10.2-9}$$

where $V = \text{NA} \cdot k_o a$, from which

$$\boxed{V = 2\pi \frac{a}{\lambda_o} \text{NA}.} \tag{10.2-10}$$

V Parameter

It is important to recall that for the wave to be guided, X must be smaller than V.

As we shall see shortly, V is an important parameter that governs the number of modes of the fiber and their propagation constants. It is called the **fiber parameter** or the V **parameter**. It is directly proportional to the radius-to-wavelength ratio a/λ_o, and to the numerical aperture NA. Equation (10.2-10) is not unlike (9.2-7) for the number of TE modes in a planar dielectric waveguide.

Modes

We now consider the boundary conditions. We begin by writing the axial components of the electric- and magnetic-field complex amplitudes, E_z and H_z, in the form of (10.2-2). The condition that these components must be continuous at the core–cladding boundary $r = a$ establishes a relation between the coefficients of proportionality in (10.2-6), so that we have only one unknown for E_z and one unknown for H_z. With the help of Maxwell's equations, $j\omega\epsilon_o n^2\mathbf{E} = \nabla \times \mathbf{H}$ and $-j\omega\mu_o\mathbf{H} = \nabla \times \mathbf{E}$ [see (5.3-12) and (5.3-13), respectively], the remaining four components, E_ϕ, H_ϕ, E_r, and H_r, are determined in terms of E_z and H_z. Continuity of E_ϕ and H_ϕ at $r = a$ yields two additional equations. One equation relates the two unknown coefficients of proportionality in E_z and H_z; the other provides a condition that the propagation constant β must satisfy. This condition, called the **characteristic equation** or **dispersion relation**, is an equation for β with the ratio a/λ_o and the fiber indices n_1, n_2 as known parameters.

For each azimuthal index l, the characteristic equation has multiple solutions yielding discrete propagation constants β_{lm}, $m = 1, 2, \ldots$, each solution representing a mode. The corresponding values of k_T and γ, which govern the spatial distributions in the core and in the cladding, respectively, are determined by using (10.2-4) and are denoted k_{Tlm} and γ_{lm}. A mode is therefore described by the indices l and m, characterizing its azimuthal and radial distributions, respectively. The function $u(r)$ depends on both l and m; $l = 0$ corresponds to meridional rays. Moreover, there are two independent configurations of the $\mathbf{E}$ and $\mathbf{H}$ vectors for each mode, corresponding to the two states of polarization. The classification and labeling of these configurations are generally quite involved (details are provided in specialized books in the reading list).

Characteristic Equation (Weakly Guiding Fiber)

Most fibers are weakly guiding (i.e., $n_1 \approx n_2$ or $\Delta \ll 1$) so that the guided rays are paraxial, i.e., approximately parallel to the fiber axis. The longitudinal components of the electric and magnetic fields are then far weaker than the transverse components and the guided waves are approximately transverse electromagnetic (TEM) in nature. The linear polarization in the x and y directions then form orthogonal states of polarization. The linearly polarized (LP) mode with indices (l, m) is usually labeled as the LP_{lm} mode. The two polarizations of mode (l, m) travel with the same propagation constant and have the same spatial distribution.

For weakly guiding fibers the characteristic equation obtained using the procedure outlined earlier turns out to be approximately equivalent to the conditions that the scalar function $u(r)$ in (10.2-6) is continuous and has a continuous derivative at $r = a$. These two conditions are satisfied if

$$\frac{(k_T a)J_l'(k_T a)}{J_l(k_T a)} = \frac{(\gamma a)K_l'(\gamma a)}{K_l(\gamma a)}. \tag{10.2-11}$$

The derivatives J_l' and K_l' of the Bessel functions satisfy the identities

$$J_l'(x) = \pm J_{l\mp1}(x) \mp l\frac{J_l(x)}{x} \tag{10.2-12}$$

$$K_l'(x) = -K_{l\mp1}(x) \mp l\frac{K_l(x)}{x}. \tag{10.2-13}$$

Substituting these identities into (10.2-11) and using the normalized parameters $X = k_T a$ and $Y = \gamma a$ leads to the characteristic equation

$$X\frac{J_{l\pm1}(X)}{J_l(X)} = \pm Y\frac{K_{l\pm1}(Y)}{K_l(Y)}, \quad Y = \sqrt{V^2 - X^2}. \tag{10.2-14}$$

(10.2-14) Characteristic Equation

Given V and l, the characteristic equation contains a single unknown variable X. Note that $J_{-l}(x) = (-1)^l J_l(x)$ and $K_{-l}(x) = K_l(x)$, so that the equation remains unchanged if l is replaced by $-l$.

The characteristic equation may be solved graphically by plotting its right- and left-hand sides (RHS and LHS, respectively) versus X and finding the intersections. As illustrated in Fig. 10.2-3 for $l = 0$, the LHS has multiple branches whereas the right-hand side decreases monotonically with increasing X until it vanishes at $X = V$ ($Y = 0$). There are therefore multiple intersections in the interval $0 < X \leq V$. Each intersection point corresponds to a fiber mode with a distinct value of X. These values are denoted X_{lm}, $m = 1, 2, \ldots, M_l$ in order of increasing X. Once the X_{lm} are found, (10.2-8), (10.2-4), and (10.2-6) allow us to determine the corresponding transverse propagation constants k_{Tlm}, the decay parameters γ_{lm}, the propagation constants β_{lm}, and the radial distribution functions $u_{lm}(r)$. The graph in Fig. 10.2-3 is similar in character to that in Fig. 9.2-2, which governs the modes of a planar dielectric waveguide.

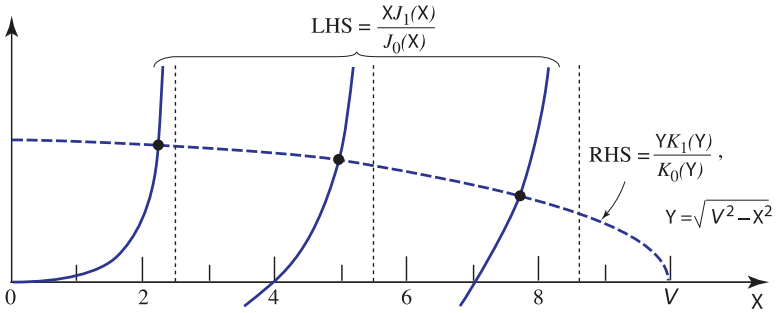


Figure 10.2-3 Graphical construction for solving the characteristic equation (10.2-14). The left- and right-hand sides are plotted as functions of X. The intersection points are the solutions. The left-hand side (LHS) has multiple branches intersecting the abscissa at the roots of $J_{l\pm1}(X)$. The right-hand side (RHS) intersects each branch once and meets the abscissa at $X = V$. The number of modes therefore equals the number of roots of $J_{l\pm1}(X)$ that are smaller than V. In this plot $l = 0$, $V = 10$, and either the $-$ or $+$ signs in (10.2-14) may be used.

Each mode has a distinct radial distribution. As examples, the two radial distributions $u(r)$ illustrated in Fig. 10.2-2 correspond to the LP_{01} mode ($l = 0$, $m = 1$) in a fiber with $V = 5$, and the LP_{34} mode ($l = 3$, $m = 4$) in a fiber with $V = 25$, respectively. Modes with $l > 0$ exist in pairs with azimuthal dependencies given by $\exp(\pm jl\phi)$, in analogy with the Laguerre–Gaussian optical beam discussed in Sec. 3.4. Since the (l, m) and $(-l, m)$ modes have the same propagation constants, the azimuthal behavior of the modes is revealed by examining the transverse intensity distribution of their equal-weight superpositions, as is understood from the

commentary associated with Fig. 3.4-2. Specifically, the complex amplitude of the sum is proportional to $u_{lm}(r)\cos l\phi\,\exp(-j\beta_{lm}z)$; the intensity, which is proportional to $u_{lm}^2(r)\cos^2 l\phi$, is illustrated in Fig. 10.2-4 for several LP_{lm} modes.

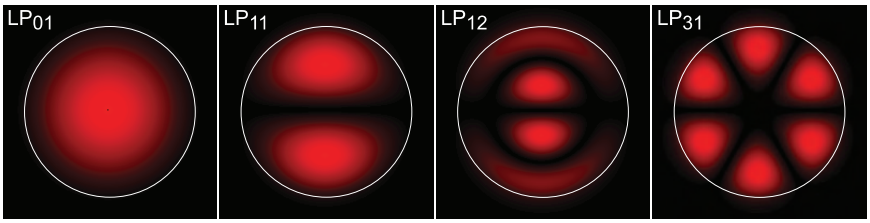


Figure 10.2-4 Intensity distributions in the transverse plane for several LP_{lm} modes for a step-index fiber with fiber parameter $V = 10$. The white circles depict core–cladding boundaries. The intensity distribution of the fundamental LP_{01} mode resembles that of the Gaussian beam displayed in Fig. 3.1-1. Each panel represents a superposition of a pair of modes with values of l that are identical but opposite in sign. The intensity distributions are proportional to $\cos^2 l\phi$ and thus display $2l$ azimuthal interference fringes.

Mode Cutoff

It is evident from the graphical construction in Fig. 10.2-3 that as V increases, the number of intersections (modes) increases since the left-hand side of the characteristic equation (10.2-14) is independent of V, whereas the right-hand side moves rightward as V increases. Considering the minus signs in the characteristic equation, branches of the left-hand side intersect the abscissa when $J_{l-1}(X) = 0$. These roots are denoted x_{lm}, $m = 1, 2, \ldots$. The number of modes M_l is therefore equal to the number of roots of $J_{l-1}(X)$ that are smaller than V. The (l, m) mode is allowed if $V > x_{lm}$. The mode reaches its cutoff point when $V = x_{lm}$. As V decreases, the $(l, m-1)$ mode also reaches its cutoff point whereupon a new root is reached, and so on. The smallest root of $J_{l-1}(X)$ is $x_{01} = 0$ for $l = 0$ and the next smallest is $x_{11} = 2.405$ for $l = 1$. The numerical values of some of these roots are provided in Table 10.2-1.

Table 10.2-1 Cutoff V parameter for low-order LP_{lm} modes.[a]

l	$m = 1$	2	3	4	5
0	0	3.832	7.016	10.174	13.324
1	2.405	5.520	8.654	11.792	14.931
2	3.832	7.016	10.174	13.324	16.471
3	5.136	8.417	11.620	14.796	17.960
4	6.380	9.761	13.015	16.224	19.409
5	7.588	11.065	14.373	17.616	20.218
6	8.772	12.339	15.700	18.980	22.218

[a]The cutoffs of the $l = 0$ modes occur at the roots of $J_{-1}(X) = -J_1(X)$. The $l = 1$ modes are cut off at the roots of $J_0(X)$, and so on.

When $V < 2.405$ all modes, with the exception of the fundamental LP_{01} mode, are cut off. The fiber then operates as a single-mode waveguide. The condition for single-mode operation is therefore

$$\boxed{V < 2.405.} \qquad (10.2\text{-}15)$$

Single-Mode Condition

Since V is proportional to the optical frequency [see (10.2-10)], the cutoff condition for the fundamental mode provided in (10.2-15) yields a corresponding cutoff frequency:

$$\nu_c = \omega_c/2\pi = \frac{1}{\text{NA}}\,\frac{c_o}{2.61a}\,. \tag{10.2-16}$$

Cutoff Frequency

By comparison, in accordance with (9.2-9), the cutoff frequency of the lowest-order mode in a dielectric slab waveguide of width d is $\nu_c = (1/\text{NA})(c_o/2d)$.

Number of Modes

A plot of the number of modes M_l as a function of V therefore takes the form of a staircase function that increases by unity at each of the roots x_{lm} of the Bessel function $J_{l-1}(X)$. A composite count of the total number of modes M (for all values of l), as a function of V, is provided in Fig. 10.2-5. Each root must be counted twice since, for each mode of azimuthal index $l > 0$, there is a corresponding mode $-l$ that is identical except for opposite polarity of the angle ϕ (corresponding to rays with helical trajectories of opposite senses), as can be seen by using the plus signs in the characteristic equation. Moreover, each mode has two states of polarization and must therefore be counted twice.

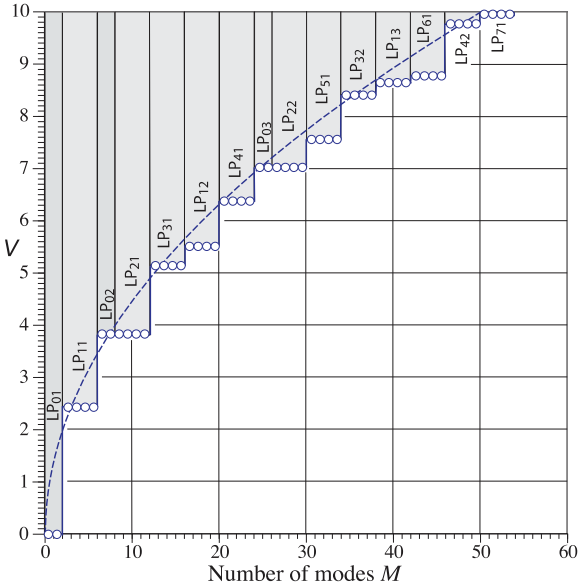


Figure 10.2-5 Total number of modes M versus fiber parameter $V = 2\pi(a/\lambda_o)\text{NA}$. Included in the count are two helical polarities for each mode with $l > 0$ as well as two polarizations per mode. For $V < 2.405$, there is only a single mode, the fundamental LP_{01} mode with two polarizations. For $V = 4$, modes LP_{11} and LP_{21} come into play, each with two helicities, along with LP_{01} and LP_{02} with zero helicities; accommodating the two polarizations in each mode leads to a total of 12 modes. The dotted curve is the relation $M = \frac{1}{2}V^2$ set forth in (10.2-18), which provides an approximate result for the number of modes when $V \gg 1$.

Though there are no explicit exact formulas for the roots of $J_l(X)$, for $X \gg l^2$ we can write $J_l(X) \approx (2/\pi X)^{1/2}\cos[X - (l + \frac{1}{2})\frac{\pi}{2}]$ in which case the roots are approximately given by $x_{lm} = (l + \frac{1}{2})\frac{\pi}{2} + (2m - 1)\frac{\pi}{2} = (l + 2m - \frac{1}{2})\frac{\pi}{2}$. This relation may be used to estimate the number of modes for fibers with large V parameter and consequently a large number of modes. When m is large the cutoff points of modes (l, m), which are the roots of $J_{l\pm1}(X)$, are

$$x_{lm} \approx \left(l + 2m - \tfrac{1}{2} \pm 1\right)\tfrac{\pi}{2} \approx (l + 2m)\tfrac{\pi}{2}, \qquad l = 0, 1, \ldots; \quad m \gg 1. \tag{10.2-17}$$

For fixed l, these roots are uniformly spaced at a distance π, in which case the number of roots M_l satisfies $l\frac{\pi}{2} + m\pi = V$, from which $M_l \approx V/\pi - l/2$. M_l then decreases

linearly with increasing l, beginning with $M_l \approx V/\pi$ for $l = 0$ and ending at $M_l = 0$ when $l = l_{\max}$, where $l_{\max} = 2V/\pi$. Accommodating the two degrees of freedom associated with positive and negative l for $l > 0$, and the two polarizations for each index (l, m), leads to a total number of modes given by $M = 2M_0 + 4\sum_{l=1}^{l_{\max}} M_l$. With the help of the relation $\sum_{l=1}^{L} l = \frac{1}{2}L(L+1)$, we obtain the approximate expression $M \approx (4/\pi^2)V^2$.

However, this expression underestimates the actual number of modes by virtue of the fact that M_l includes lower-order modes for which the separation distances are less than π, as evinced in Table 10.2-1. As illustrated in Fig. 10.2-5, a good fit to the exact number of modes is provided by the approximation

$$M \approx \tfrac{1}{2}V^2, \tag{10.2-18}$$

Number of Modes ($V \gg 1$)

an expression that obtains in the quasi-plane-wave approach, as shown in Sec. 10.2C [see (10.2-35)].

Based on (10.2-18), when V is large the approximate number of modes in the circular waveguide is given by $M \approx \frac{1}{2}V^2 = 2\pi(\pi a^2/\lambda_o^2)(\mathrm{NA})^2$. This expression is analogous to that for the number of modes in a square dielectric waveguide of cross sectional area d^2, in which case $M \approx 2\pi(d^2/\lambda_o^2)(\mathrm{NA})^2$ when both TE and TM polarizations are accommodated [see (9.3-3)].

EXAMPLE 10.2-1. ***Number of Modes.*** A silica-glass fiber with $n_1 = 1.452$ and $\Delta = 0.01$ has a numerical aperture $\mathrm{NA} = \sqrt{n_1^2 - n_2^2} \approx n_1\sqrt{2\Delta} \approx 0.205$. If $\lambda_o = 1.55\ \mu$m and the core radius $a = 20\ \mu$m, then $V = 2\pi(a/\lambda_o)\mathrm{NA} \approx 16.6$. There are therefore approximately $M \approx \frac{1}{2}V^2 \approx 138$ modes. If the cladding is stripped away so that the core is in direct contact with air, $n_2 = 1$ and $\mathrm{NA} = 1$, whereupon $V \approx 81.1$ and approximately 3,286 modes are allowed.

Propagation Constants and Group Velocities

As indicated earlier, the propagation constants can be determined by solving the characteristic equation (10.2-14) for the X_{lm} and using (10.2-4a) and (10.2-8) to obtain $\beta_{lm} = (n_1^2 k_o^2 - X_{lm}^2/a^2)^{1/2}$. Since $V = 2\pi(a/\lambda_o)\mathrm{NA}$ and $\mathrm{NA} = n_1\sqrt{2\Delta}$, we have $\beta_{lm} = n_1 k_o(1 - 2\Delta \cdot X_{lm}^2/V^2)^{1/2}$. Because $\Delta \ll 1$ and $X_{lm} < V$, we use the expansion $(1+\delta)^{1/2} \approx 1 + \frac{1}{2}\delta$ for $|\delta| \ll 1$ to obtain

$$\beta_{lm} \approx n_1 k_o \left[1 - \frac{X_{lm}^2}{V^2}\,\Delta\right]. \tag{10.2-19}$$

Propagation Constants
$0 < X_{lm} < V$

For fibers with large V, corresponding to a large number of modes, X_{lm} spans the range $0 < X_{lm} < V$ so that β_{lm} varies approximately between $n_1 k_o$ and $n_1 k_o(1-\Delta) \approx n_2 k_o$, as predicted by ray optics.

To determine the group velocity of the (l, m) mode, $v_{lm} = d\omega/d\beta_{lm}$, we express β_{lm} as an explicit function of ω by substituting $n_1 k_o = \omega/c_1$ and $V = a(\omega/c_o)\mathrm{NA}$ into (10.2-19), and then calculating $(d\beta_{lm}/d\omega)^{-1}$. Assuming that c_1 and Δ are independent of ω, and using the approximation $(1+\delta)^{-1} \approx 1 - \delta$ for $|\delta| \ll 1$, the group velocity becomes

$$v_{lm} \approx c_1 \left[1 - \frac{X_{lm}^2}{V^2}\,\Delta\right]. \tag{10.2-20}$$

Group Velocities
$0 < X_{lm} < V$

For $V \gg 1$, the group velocity varies approximately between c_1 and $c_1(1-\Delta) = c_1(n_2/n_1)$. In this case, the group velocities of the low-order modes are approximately equal to the phase velocity of the core material, whereas those of the high-order modes are smaller.

The fractional group-velocity change between the fastest and the slowest mode is roughly equal to Δ, the fractional refractive index change of the fiber. Fibers with large Δ, although endowed with a large NA and therefore large light-gathering capacity, also have a large number of modes, large modal dispersion, and consequently high pulse-spreading rates. These effects are particularly severe if the cladding is removed altogether.

B. Single-Mode Fibers

As discussed earlier, a fiber with core radius a and numerical aperture NA operates as a single-mode fiber in the fundamental LP_{01} mode if $V = 2\pi(a/\lambda_o)\text{NA} < 2.405$. Single-mode operation is therefore achieved via a small core diameter and small numerical aperture (indicating that n_2 is close to n_1), or by operating at a sufficiently low optical frequency [below the cutoff frequency $\nu_c = (1/\text{NA})(c_o/2.61a)$].

The fundamental LP_{01} mode has a bell-shaped spatial distribution (Figs. 10.2-2 and 10.2-4 for $l = 0$) similar to that of the simple Gaussian beam (Fig. 3.1-1). It provides the greatest confinement of light power within the core.

EXAMPLE 10.2-2. ***Single-Mode Operation.*** A silica-glass fiber with $n_1 = 1.447$ and $\Delta = 0.01$ (NA $= 0.205$) operates at $\lambda_o = 1.3$ μm as a single-mode fiber if $V = 2\pi(a/\lambda_o)\text{NA} < 2.405$, i.e., if the core diameter $2a < 4.86$ μm. If Δ is reduced to 0.0025, single-mode operation is maintained for a diameter $2a < 9.72$ μm.

The dependence of the effective refractive index $n = \beta/k_o$ on the V parameter for the fundamental mode is displayed in Fig. 10.2-6(a), and the corresponding dispersion relation (ω versus β) is illustrated in Fig. 10.2-6(b). As the V parameter increases, i.e., as the frequency increases or the fiber diameter increases, the effective refractive index n increases from n_2 to n_1. This is expected since the mode is more confined in the core at shorter wavelengths.

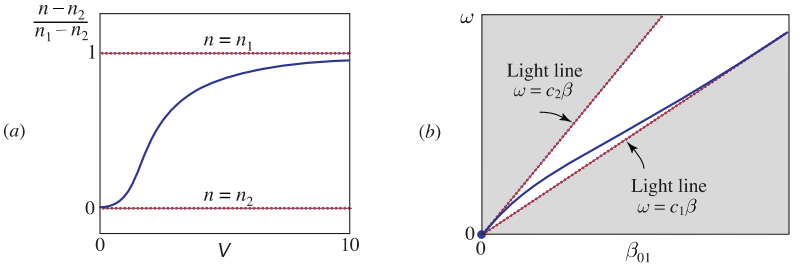


Figure 10.2-6 Schematic illustrations of the propagation characteristics of the fundamental LP_{01} mode. (a) Effective refractive index $n = \beta/k_o$ as a function of the V parameter. (b) Dispersion relation (ω versus β_{01}).

There are numerous advantages of using single-mode fibers in optical fiber communication systems. As explained earlier, the modes of a multimode fiber travel at different group velocities so that a short-duration pulse of multimode light suffers a range of delays and therefore spreads in time. Quantitative measures of modal dispersion are examined in Sec. 10.3B. In a single-mode fiber, on the other hand, there is only one mode with a single group velocity, so that a short pulse of light arrives without delay distortion. As explained in Sec. 10.3B, pulse spreading in single-mode fibers does nevertheless result from other dispersive mechanisms, but these are significantly smaller than modal dispersion.

Moreover, as shown in Sec. 10.3A, the rate of power attenuation is lower in a single-mode fiber than in a multimode fiber. This, together with the smaller rate of pulse spreading, permits substantially higher data rates to be transmitted over single-mode fibers than over multimode fibers. This topic is addressed further in Chapters 23 and 25.

Another difficulty with the use of multimode fibers stems from the random interference of the modes. As a result of uncontrollable imperfections, strains, and temperature fluctuations, each mode undergoes a random phase shift so that the sum of the complex amplitudes of the modes exhibits an intensity that is random in time and space. This randomness is known as **modal noise** or **speckle**. This effect is similar to the fading of radio signals resulting from multiple-path transmission. In a single-mode fiber there is only one path and therefore no modal noise.

Polarization-Maintaining Fibers

In a fiber with circular cross section, each mode has two independent states of polarization with the same propagation constant. Thus, the fundamental LP_{01} mode in a single-mode weakly guiding fiber may be polarized in the x or y direction; the two orthogonal polarizations have the same propagation constant and the same group velocity.

In principle, there should be no exchange of power between the two polarization components. If the power of the light source is delivered exclusively into one polarization, the power should remain in that polarization. In practice, however, slight random imperfections and uncontrollable strains in the fiber result in random power transfer between the two polarizations, as illustrated in Fig. 10.2-7.

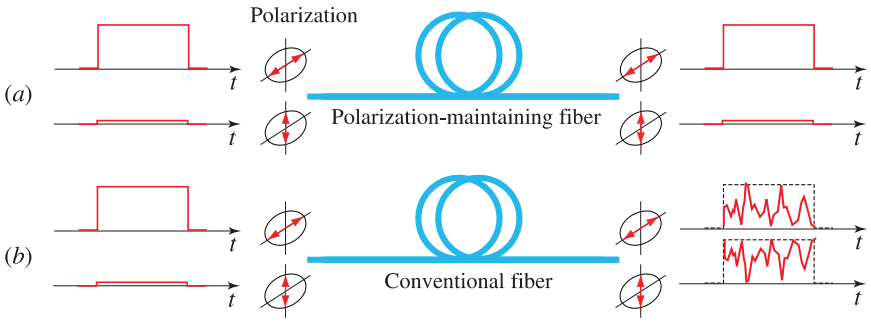


Figure 10.2-7 (a) Ideal polarization-maintaining fiber. (b) Random transfer of power between two polarizations in a conventional fiber.

Such coupling is facilitated because the two polarizations have the same propagation constant and their phases are therefore matched. Thus, linearly polarized light at the fiber input is generally transformed into elliptically polarized light at the fiber output.

In spite of the fact that the total optical power remains fixed (see Fig. 10.2-7), the ellipticity of the received light fluctuates randomly with time as a result of fluctuations in the material strain and temperature, and of the source wavelength. The randomization of the power division between the two polarization components poses no difficulty if the object is solely to transmit light power, provided that the total power is collected.

However, in many areas where fiber optics is used, e.g., in integrated-photonic devices, optical sensors based on interferometric techniques, and coherent optical communications, the fiber must transmit the complex amplitude (magnitude and phase) of a specific polarization. Polarization-maintaining fibers are required for such applications. To construct a polarization-maintaining fiber, the circular symmetry of the conventional fiber must be abandoned, for example by using fibers with elliptical cross section or stress-induced anisotropy of the refractive index. This eliminates the polarization degeneracy, thereby making the propagation constants of the two polarizations different. The introduction of such phase mismatch serves to reduce the coupling efficiency.

*C. Quasi-Plane Waves in Step-Index and Graded-Index Fibers

As shown in Sec. 10.1A, the modes of a step-index fiber are determined by writing the Helmholtz equation (10.2-1) with $n = n(r)$, solving for the spatial distributions of the field components, and using Maxwell's equations and the boundary conditions to obtain the characteristic equation. However, carrying out this procedure for a graded-index fiber is generally a difficult proposition.

In this section we rely instead on an approximate approach based on picturing the field distribution as a quasi-plane wave traveling within the core, approximately along the trajectory of an optical ray. A quasi-plane wave is a wave that is locally identical to a plane wave, but slowly changes its direction and amplitude as it travels. This approach permits us to maintain the simplicity of ray optics while at the same time retaining the phase associated with the wave, so that the self-consistency condition for determining the propagation constants of the guided modes can be used (as was done for the planar dielectric waveguide in Sec. 9.2). This approximate technique, which makes use of the WKB (Wentzel–Kramers–Brillouin) method, is applicable only for fibers with a large number of modes (large V parameter). This approach also allows us to conveniently compare the behavior of step-index and graded-index fibers.

Quasi-Plane Waves

Consider a solution of the Helmholtz equation (10.2-1) that takes the form of a quasi-plane wave (see Sec. 2.3)

$$U(\mathbf{r}) = \mathcal{a}(\mathbf{r}) \exp\left[-jk_o S(\mathbf{r})\right], \tag{10.2-21}$$

where $\mathcal{a}(\mathbf{r})$ and $S(\mathbf{r})$ are real functions of position that are slowly varying in comparison with the wavelength $\lambda_o = 2\pi/k_o$. It is known from (2.3-4) that $S(\mathbf{r})$ approximately satisfies the eikonal equation $|\nabla S|^2 \approx n^2$, and that the rays travel in the direction of the gradient ∇S. If we take $k_o S(\mathbf{r}) = k_o s(r) + l\phi + \beta z$, where $s(r)$ is a slowly varying function of r, the eikonal equation yields

$$\left(k_o \frac{ds}{dr}\right)^2 + \beta^2 + \frac{l^2}{r^2} = n^2(r)\, k_o^2. \tag{10.2-22}$$

The local spatial frequency of the wave in the radial direction is the partial derivative of the phase $k_o S(\mathbf{r})$ with respect to r,

$$k_r = k_o \frac{ds}{dr}, \tag{10.2-23}$$

so that (10.2-21) becomes

$$U(r) = \mathfrak{a}(r)\,\exp\left(-j\int_0^r k_r\,dr\right)e^{-jl\phi}\,e^{-j\beta z} \tag{10.2-24}$$

Quasi-Plane Wave

and (10.2-22) provides

$$k_r^2 = n^2(r)\,k_o^2 - \beta^2 - l^2/r^2. \tag{10.2-25}$$

Defining $k_\phi = l/r$ so that $\exp(-jl\phi) = \exp(-jk_\phi r\phi)$, and $k_z = \beta$, (10.2-25) yields $k_r^2 + k_\phi^2 + k_z^2 = n^2(r)\,k_o^2$. The quasi-plane wave therefore has a local wavevector $\mathbf{k}$ with magnitude $n(r)k_o$ and cylindrical-coordinate components (k_r, k_ϕ, k_z). Since $n(r)$ and k_ϕ are functions of r, k_r is also generally position dependent. The direction of $\mathbf{k}$ changes slowly with r (see Fig. 10.2-8), and follows a helical trajectory similar to that of the skewed ray shown earlier in Fig. 10.1-6(b).

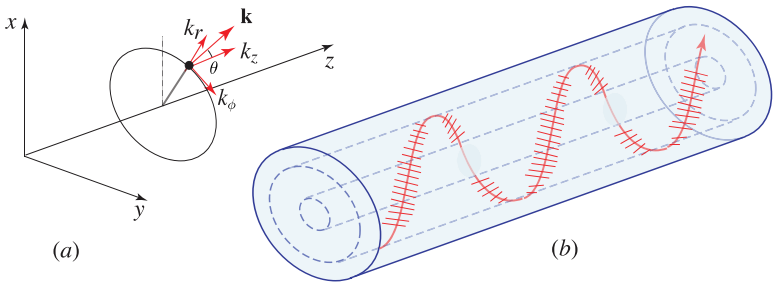


Figure 10.2-8 (a) The wavevector $\mathbf{k} = (k_r, k_\phi, k_z)$ in a cylindrical coordinate system. (b) Quasi-plane wave following the direction of a ray.

To establish the region of the core within which the wave is bound, we determine the values of r for which k_r is real, or $k_r^2 > 0$. For given values of l and β we plot $k_r^2 = [n^2(r)\,k_o^2 - l^2/r^2 - \beta^2]$ as a function of r. The term $n^2(r)\,k_o^2$ is first plotted as a function of r [thick solid curve in Fig. 10.2-9(a)]. The term l^2/r^2 is then subtracted, yielding the dashed curve. The value of β^2 is marked by the thin solid vertical line. It follows that k_r^2 is represented by the difference between the dashed curve and the thin solid line, i.e., by the shaded area. Regions where k_r^2 is positive and negative are indicated by $+$ and $-$ signs, respectively.

For the step-index fiber, we have $n(r) = n_1$ for $r < a$, and $n(r) = n_2$ for $r > a$. In this case the quasi-plane wave is guided in the core by reflecting from the core–cladding boundary at $r = a$. As illustrated in Fig. 10.2-9(a), the region of confinement is then $r_l < r < a$, where

$$n_1^2 k_o^2 - l^2/r_l^2 - \beta^2 = 0. \tag{10.2-26}$$

The wave bounces back and forth helically like the skewed ray illustrated in Fig. 10.1-2. In the cladding ($r > a$), and near the center of the core ($r < r_l$), k_r^2 is negative so that k_r is imaginary; the wave therefore decays exponentially in these regions. Note

that r_l depends on β. For large β (or large l), r_l is large so that the wave is confined to a thin cylindrical shell near the boundary of the core.

For the graded-index fiber illustrated in Fig. 10.2-9(b), k_r is real in the region $r_l < r < R_l$, where r_l and R_l are the roots of the equation

$$n^2(r)\,k_o^2 - l^2/r^2 - \beta^2 = 0. \tag{10.2-27}$$

It follows that the wave is essentially confined within a cylindrical shell of radii r_l and R_l, just as for the helical ray trajectory shown in Fig. 10.1-6(b).

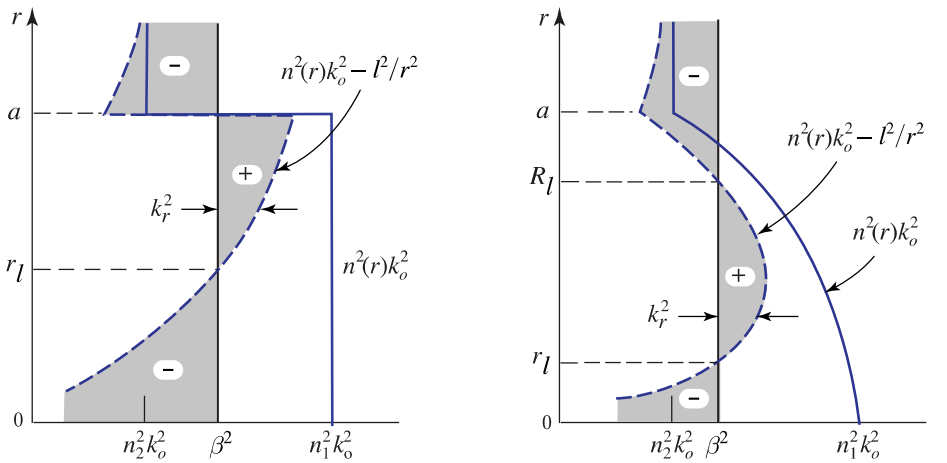


Figure 10.2-9 Dependence of $n^2(r)\,k_o^2$, $n^2(r)\,k_o^2 - l^2/r^2$, and $k_r^2 = n^2(r)\,k_o^2 - l^2/r^2 - \beta^2$ on the position r. At any r, k_r^2 is the width of the shaded area with the $+$ and $-$ signs denoting positive and negative values of k_r^2, respectively. (a) Step-index fiber: k_r^2 is positive in the region $r_l < r < a$. (b) Graded-index fiber: k_r^2 is positive in the region $r_l < r < R_l$.

Modes

The modes of the fiber are determined by imposing the self-consistency condition that the wave reproduce itself after one helical period of travel between r_l and R_l and back. The azimuthal pathlength corresponding to an angle 2π must correspond to a multiple of 2π phase shift, i.e., $k_\phi 2\pi r = 2\pi l$; $l = 0, \pm1, \pm2, \ldots$. This condition is evidently satisfied since $k_\phi = l/r$. Furthermore, since the component k_r vanishes at r_l and R_l, the phase shift encountered in traveling between these turning points must be a multiple of π, much like the case of a standing wave between two mirrors. Thus,

$$\int_{r_l}^{R_l} k_r\,dr = \pi m, \qquad m = 1, 2, \ldots, M_l, \tag{10.2-28}$$

where $R_l = a$ for the step-index fiber. This condition provides the characteristic equation from which the propagation constants β_{lm} of the modes are determined. These values are represented schematically in Fig. 10.2-10; the mode $m = 1$ has the largest value of β (approximately $n_1 k_o$) whereas the mode $m = M_l$ has the smallest value (approximately $n_2 k_o$). The WKB method, which is applicable for oscillatory solutions between turning points, yields more accurate results: the quantities l^2 and m above are replaced by $l^2 + 1/4$ and $m + 1/4$, corrections that are particularly important for small values of l and m.

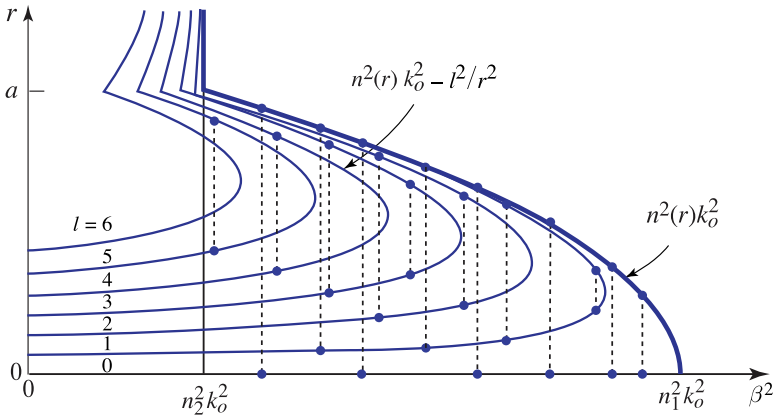


Figure 10.2-10 The propagation constants and confinement regions of the fiber modes. Each curve corresponds to an index l, which stretches from 0 to 6 in this plot. Each mode (corresponding to a certain value of m) is schematically indicated by two dots connected by a dashed vertical line. The ordinates of the dots denote the radii r_l and R_l of the cylindrical shell within which the mode is confined. Values on the abscissa are the squared propagation constants of the modes, β^2.

Number of Modes

The total number of modes can be determined by adding the number of modes M_l for $l = 0, 1, \ldots, l_{\max}$. We approach this computation using a different procedure, however. We first determine the number q_β of modes with propagation constants greater than a given value β. For each l, the number of modes $M_l(\beta)$ with propagation constant greater than β is the number of multiples of 2π the integral in (10.2-28) yields, i.e.,

$$M_l(\beta) = \frac{1}{\pi}\int_{r_l}^{R_l} k_r\, dr = \frac{1}{\pi}\int_{r_l}^{R_l} \sqrt{n^2(r)k_o^2 - l^2/r^2 - \beta^2}\, dr, \tag{10.2-29}$$

where r_l and R_l are the radii of confinement corresponding to the propagation constant β, as provided in (10.2-27). Clearly, r_l and R_l depend on β, and $R_l = a$ for the step-index fiber.

The total number of modes with propagation constant greater than β is therefore

$$q_\beta = 4\sum_{l=0}^{l_{\max}(\beta)} M_l(\beta), \tag{10.2-30}$$

where $l_{\max}(\beta)$ is the maximum value of l that yields a bound mode with propagation constants greater than β, i.e., for which the peak value of the function $n^2(r)\,k_o^2 - l^2/r^2$ is greater than β^2. The grand-total mode count M is q_β for $\beta = n_2k_o$. The factor of 4 in (10.2-30) accommodates the two possible polarities of the angle ϕ, corresponding to positive and negative helical trajectories for each (l, m), and the two possible polarizations. If the number of modes is sufficiently large, we can replace the summation in (10.2-30) by an integration, whereupon

$$q_\beta \approx 4\int_0^{l_{\max}(\beta)} M_l(\beta)\, dl. \tag{10.2-31}$$

For fibers with power-law refractive-index profiles, we insert (10.1-4) into (10.2-

29), and thence into (10.2-31). Evaluation of the integral then yields

$$q_\beta \approx M \left[\frac{1-(\beta/n_1 k_o)^2}{2\Delta}\right]^{\frac{p+2}{p}} \tag{10.2-32}$$

with

$$M \approx \frac{p}{p+2}\, n_1^2\, k_o^2\, a^2 \Delta = \frac{p}{p+2}\,\frac{V^2}{2}, \tag{10.2-33}$$

where $\Delta = (n_1 - n_2)/n_1$ and $V = 2\pi(a/\lambda_o)\mathrm{NA}$ is the fiber V parameter. Since $q_\beta \approx M$ at $\beta = n_2 k_o$, M is indeed the total number of modes.

For step-index fibers ($p \to \infty$), (10.2-32) and (10.2-33) become

$$q_\beta \approx M \left[\frac{1-(\beta/n_1 k_o)^2}{2\Delta}\right] \tag{10.2-34}$$

and

$$\boxed{M \approx \tfrac{1}{2}V^2,} \tag{10.2-35}$$

Number of Modes (Step-Index)

respectively. This expression for M is the same as that set forth in (10.2-18), which was found to be a good fit to the exact number as a function of V, as shown in Fig. 10.2-5.

Propagation Constants

The propagation constant β_q for mode q is obtained by inverting (10.2-32):

$$\beta_q \approx n_1 k_o \sqrt{1 - 2\left(\frac{q}{M}\right)^{p/(p+2)}\Delta}\ , \qquad q = 1, 2, \ldots, M, \tag{10.2-36}$$

where the index q_β has been replaced by q, and β has been replaced by β_q. Since $\Delta \ll 1$, the approximation $\sqrt{1+\delta} \approx 1 + \frac{1}{2}\delta$ (applicable for $|\delta| \ll 1$) can be applied to (10.2-36), yielding

$$\boxed{\beta_q \approx n_1 k_o \left[1 - \left(\frac{q}{M}\right)^{p/(p+2)}\Delta\right].} \tag{10.2-37}$$

Propagation Constants

The propagation constant β_q therefore decreases from $\approx n_1 k_o$ (for $q = 1$) to $n_2 k_o$ (for $q = M$), as illustrated in Fig. 10.2-11. For the step-index fiber ($p \to \infty$), (10.2-36) reduces to

$$\boxed{\beta_q \approx n_1 k_o \left(1 - \frac{q}{M}\Delta\right),} \tag{10.2-38}$$

Propagation Constants (Step-Index Fiber)

which mimics the behavior of (10.2-19).

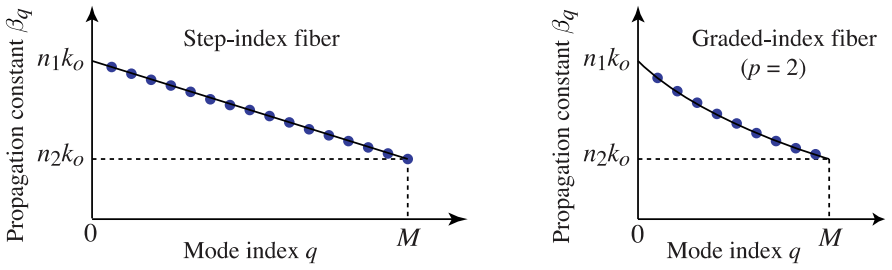


Figure 10.2-11 Dependence of the propagation constants β_q on the mode index $q = 1, 2, \ldots, M$ for (a) a step-index fiber $(p \to \infty)$ and (b) an optimal graded-index fiber $(p = 2)$.

Group Velocities

To determine the group velocity $v_q = d\omega/d\beta_q$, we write β_q as a function of ω by substituting (10.2-33) into (10.2-37), substituting $n_1 k_o = \omega/c_1$ into the result, and evaluating $v_q = (d\beta_q/d\omega)^{-1}$. With the help of the approximation $(1+\delta)^{-1} \approx 1 - \delta$ (valid for $|\delta| \ll 1$), and assuming that c_1 and Δ are independent of ω (i.e., ignoring material dispersion), we obtain

$$v_q \approx c_1 \left[1 - \frac{p-2}{p+2}\left(\frac{q}{M}\right)^{p/(p+2)} \Delta\right]. \tag{10.2-39}$$

Group Velocities

For the step-index fiber $(p \to \infty)$, (10.2-39) yields

$$v_q \approx c_1 \left(1 - \frac{q}{M}\Delta\right). \tag{10.2-40}$$

The group velocity thus varies from approximately c_1 to $c_1(1-\Delta)$, as illustrated in Fig. 10.2-12(a). The form of this equation harks back to (10.2-20).

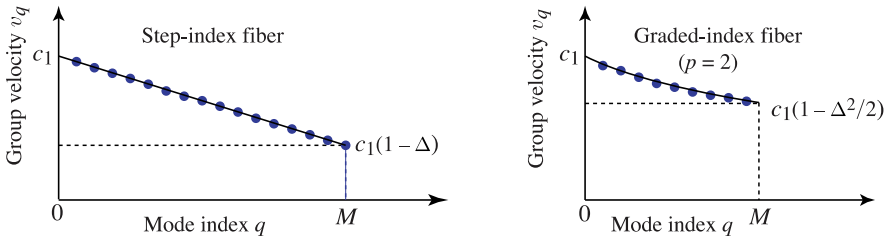


Figure 10.2-12 Group velocities v_q of the modes of (a) a step-index fiber $(p \to \infty)$ and (b) an optimal graded-index fiber $(p = 2)$.

Optimal Index Profile

Equation (10.2-39) indicates that the grade profile parameter $p = 2$ yields a group velocity $v_q \approx c_1$ for all q, so that all modes travel at approximately the same velocity c_1. This highlights the advantage of the graded-index fiber for multimode transmission.

To determine the group velocity with better accuracy, we return to the derivation of v_q from (10.2-36) for $p = 2$. Carrying the Taylor-series expansion to three terms instead of two, i.e., $\sqrt{1+\delta} \approx 1 + \frac{1}{2}\delta - \frac{1}{8}\delta^2$, gives rise to

$$v_q \approx c_1 \left(1 - \frac{q}{M}\frac{\Delta^2}{2}\right). \tag{10.2-41}$$

Group Velocities (Graded-Index, $p = 2$)

Thus, the group velocities vary from approximately c_1 at $q = 1$ to approximately $c_1(1 - \Delta^2/2)$ at $q = M$. Comparison with the results for the step-index fiber is provided in Fig. 10.2-12. The group-velocity difference for the parabolically graded fiber is $\Delta^2/2$, which is substantially smaller than the group-velocity difference Δ for the step-index fiber. Under ideal conditions, the graded-index fiber therefore reduces the group-velocity difference by a factor $\Delta/2$, thus realizing its intended purpose of equalizing the modal velocities. However, since the analysis leading to (10.2-41) is based on a number of approximations, this improvement factor is only a rough estimate — indeed it is not fully attained in practice.

The number of modes M in a graded-index fiber with grade profile parameter p is specified by (10.2-33). For $p = 2$, this becomes

$$M \approx \tfrac{1}{4}V^2. \tag{10.2-42}$$

Number of Modes (Graded-Index, $p = 2$)

Comparing this with the result for the step-index fiber provided in (10.2-35), $M \approx \frac{1}{2}V^2$, reveals that the number of modes in an optimal graded-index fiber is roughly half that in a step-index fiber with the same parameters n_1, n_2, and a.

D. Multicore Fibers and Fiber Couplers

Multicore Fibers

A **multicore fiber** (MCF) is a fiber with multiple cores embedded in a common cladding (Fig. 10.2-13). The application in which such a fiber is to be used determines the number of cores, the manner in which the cores are arranged, their diameters, and their separations (pitch). Cores with small (large) diameter function as single-mode (multimode) waveguides, respectively. Cores that are sufficiently well separated exhibit minimal inter-core crosstalk and hence behave as independent waveguides. MCFs find use in high-capacity optical networks, in inter-chip communications in datacenters, and as sensors since they respond differentially to mechanical changes (e.g., bending).

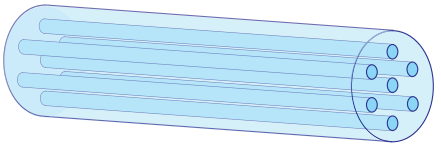


Figure 10.2-13 A multicore fiber (MCF) with seven cores.

Multicore Couplers

Optical fiber couplers are important components of optical fiber systems. **Multicore couplers** are used to connect one or more fibers at the input of the device to one or more fibers at its output. Several examples implementing various coupling configurations are illustrated in Fig. 10.2-14 (additional examples are provided in Sec. 24.1B). It should be pointed out, however, that other technologies are also useful for achieving coupling, as discussed in Sec. 24.1A.

A key consideration in the design of an optical fiber coupler is the optical loss that it introduces; the reduction of insertion loss is particularly challenging for single-mode fibers. In certain applications, it is desired to render the distribution of optical

power at the output fibers of the coupler insensitive to particular wave properties, such as wavelength or polarization. In other applications, the coupler may be deliberately designed so that coupling is governed by one of these properties, such as different wavelengths or polarizations being directed to different output fibers, as considered in Sec. 24.1.

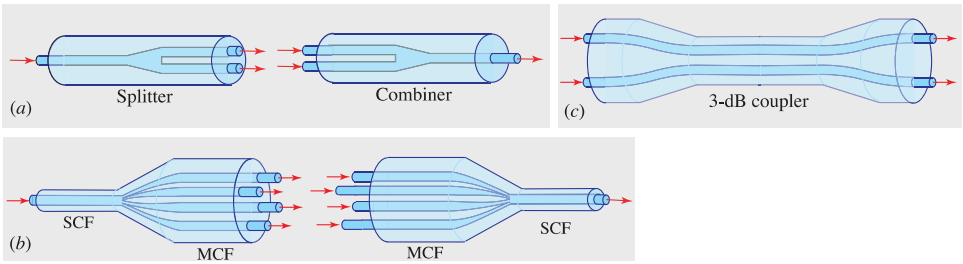


Figure 10.2-14 Multicore fiber-optic couplers. (*a*) Dual-core fiber used as a splitter or a combiner. (*b*) Fan-out and fan-in couplers using a conical fused taper to connect a single-core fiber (SCF) to a multicore fiber (MCF). (*c*) A 3-dB coupler using a dual-core fiber incorporating a biconical fused taper. Adjacent fibers are weakly coupled, which allows light to gradually migrate from one fiber to the other, much as in the integrated-photonic coupler displayed in Fig. 9.4-7(*b*).

Photonic Lantern

A **photonic lantern** is a fiber-optic coupler that connects a single multimode fiber (MMF) to multiple single-mode fibers (SMFs). As illustrated in Fig. 10.2-15, working our way from right to left, a collection of SMFs are fused into a single glass body, thereby forming a multicore fiber (MCF). This body in turn is tapered in cross section and connected to a single-core MMF at the narrow end of the taper. The device is reciprocal and can be used in either a fan-out or a fan-in configuration.

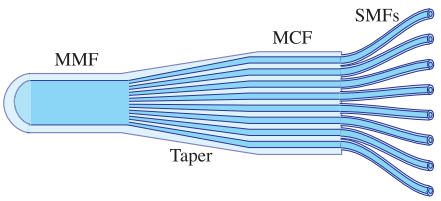


Figure 10.2-15 A photonic lantern provides an interface between the modes of one multimode fiber (MMF) and many single-mode fibers (SMFs). It can be operated in a fan-out configuration, in which light travels from left to right, or in a fan-in configuration, in which light travels from right to left.

When operated in a fan-out configuration, incoming light excites the modes of the MMF, and each mode in turn excites several, or all, of the multiple cores in the tapered region. Within the narrow portion of the tapered region, the cores are weakly coupled and thus behave as an array of coupled single-mode waveguides, wherein the light in one waveguide can migrate into neighboring waveguides (see Sec. 9.4C). In the wider portion of the tapered region, the cores are sufficiently separated so that they form a MCF comprising independent single-mode waveguides that serve to feed the outgoing SMFs.

While this structure may be used as a simple fan-out coupler that distributes the power entering the MMF among the outgoing SMFs, much like the fan-out configuration depicted in Fig. 10.2-14(*b*), it has additional capabilities. Ideally, each of the MMF modes would couple to one, and only one, of the SMFs, whereupon the device would serve as a demultiplexer separating the information carried by each of the incoming modes into distinct SMF channels (or, in the case of the reciprocal fan-in configuration, as a multiplexer). Unfortunately, however, at the entrance to the taper, and throughout its narrow region where the constituent fibers are not well separated, each of the MMF modes couples to more than one of the SMFs, thereby obviating such a one-to-one correspondence.

Nevertheless, the resultant mixing of information can be undone using computational methods. The relation between the complex amplitudes of the output waves in the SMFs and those of the modes in the input MMF is mathematically represented by a matrix with appropriate weights for such a linear superposition. This matrix is established by the geometry of the taper and the transmission through the array of coupled waveguides, as described in Sec. 9.4C for waveguide arrays. If this matrix is able to be determined, then matrix inversion can be used to computationally extract the information carried by the MMF modes from the complex amplitudes of the outgoing light in the SMFs.

Photonic lanterns find use in optical fiber communication systems in which MMF modes serve as individual communication channels (spatial-mode multiplexing is considered in Sec. 25.3B). Fan-out lanterns may also be used to distribute light carried by one MMF to several SMFs for processing by optical components designed for single-mode operation, such as single-mode fiber Bragg grating (FBG) spectral filters or optical amplifiers, before being recombined and directed to another MMF via a fan-in lantern. An example of the use of a photonic lantern in astronomical instrumentation relies on directing the multimode light delivered by a telescope to a set of single-mode fiber filters for spectral filtering.

10.3 ATTENUATION AND DISPERSION

Attenuation and dispersion limit the performance of the optical-fiber medium as a data-transmission channel. Attenuation, associated with losses of various kinds, limits the magnitude of the optical power transmitted. Dispersion, which is responsible for the temporal spread of optical pulses, limits the rate at which data-carrying pulses may be transmitted.

A. Attenuation

Attenuation Coefficient

The power of a light beam traveling through an optical fiber decreases exponentially with distance as a result of absorption and scattering. The associated **attenuation coefficient** is conventionally defined in units of decibels per kilometer (dB/km) and is denoted by the symbol α,

$$\alpha = \frac{1}{L}\, 10 \log_{10} \frac{1}{\mathcal{T}}, \tag{10.3-1}$$

where $\mathcal{T} = P(L)/P(0)$ is the **power transmission ratio** (the ratio of transmitted to incident power) for a fiber of length L km. The conversion of a ratio to dB units is illustrated in Fig. 10.3-1. An attenuation of 3 dB/km, for example, corresponds to a power transmission of $\mathcal{T} = 0.5$ through a fiber of length $L = 1$ km.

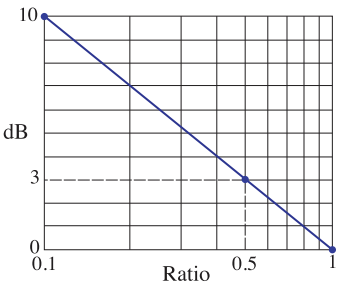


Figure 10.3-1 The dB value of a ratio. For example, 3 dB is equivalent to a ratio of 0.5; 10 dB corresponds to $\mathcal{T} = 0.1$; and 20 dB corresponds to $\mathcal{T} = 0.01$.

For light traveling through a cascade of lossy systems, the overall transmission ratio is the *product* of the constituent transmission ratios. By virtue of the logarithmic relation in (10.3-1), the overall loss in dB thus becomes the *sum* of the constituent dB losses. For a propagation distance of z km, the loss is αz dB. The associated power transmission ratio, which is obtained by inverting (10.3-1), is then

$$P(z)/P(0) = 10^{-\alpha z/10} \approx e^{-0.23\,\alpha z}, \qquad \alpha \text{ in dB/km}. \tag{10.3-2}$$

Equation (10.3-2) applies when the quantity α is specified in units of dB/km. If, instead, the attenuation coefficient α is specified in units of km^{-1}, we have

$$P(z)/P(0) = e^{-\alpha z}, \qquad \alpha \text{ in km}^{-1}, \tag{10.3-3}$$

where $\alpha \approx 0.23\,\alpha$. For components other than optical fibers, the attenuation coefficient α is usually specified in units of cm^{-1} in which case the power attenuation is described by (10.3-3) with z specified in cm.

Absorption

The attenuation coefficient α of fused silica (SiO_2) is strongly dependent on wavelength, as illustrated in Fig. 10.3-2. This material has two strong absorption bands: a mid-infrared absorption band resulting from vibrational transitions and an ultraviolet absorption band arising from electronic and molecular transitions. The tails of these bands form a window in the near infrared region of the spectrum in which there is little intrinsic absorption.

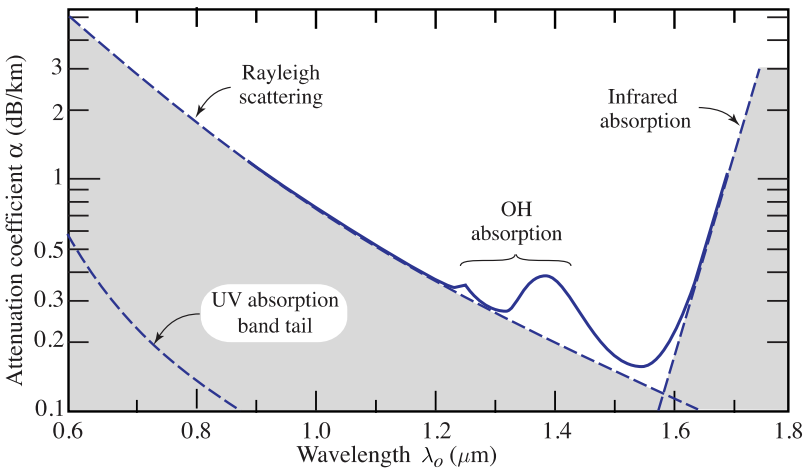


Figure 10.3-2 Attenuation coefficient α of silica glass versus free-space wavelength λ_o. There is a local minimum at 1.3 μm ($\alpha \approx 0.3$ dB/km) and an absolute minimum at 1.55 μm ($\alpha \approx 0.15$ dB/km).

Scattering

Rayleigh scattering is another intrinsic effect that contributes to the attenuation of light in glass. The random localized variations of the molecular positions in the glass itself create random inhomogeneities in the refractive index that act as tiny scattering centers. The amplitude of the scattered field is proportional to ω^2, where ω is the angular frequency of the light.[†] As discussed in Sec. 5.6B, the scattered intensity is

[†] The scattering medium creates a polarization density $\mathcal{P}$, which corresponds to a source of radiation proportional to $d^2\mathcal{P}/dt^2 = -\omega^2\mathcal{P}$; see (5.2-25).

therefore proportional to ω^4, or to $1/\lambda_o^4$, so that short wavelengths are scattered more than long wavelengths. Light in the visible region of the spectrum is therefore scattered more than light in the infrared.

The functional form of Rayleigh scattering, which decreases with wavelength as $1/\lambda_o^4$, is known as the **Rayleigh inverse fourth-power law**. In the visible region of the spectrum, Rayleigh scattering is a more significant source of loss than is the tail of the ultraviolet absorption band, as shown in Fig. 10.3-2. However, Rayleigh loss becomes negligible in comparison with infrared absorption in silica glass for wavelengths greater than $\approx 1.6\ \mu\text{m}$.

We conclude that the transparent window in silica glass is bounded by Rayleigh scattering on the short-wavelength side and by infrared absorption on the long-wavelength side (indicated by the dashed curves in Fig. 10.3-2). Near-infrared communication systems using silica-glass fibers are deliberately designed to operate in this window.

Extrinsic Effects

Aside from these intrinsic effects there are extrinsic absorption bands that result from the presence of impurities in silica glass, principally metallic ions (e.g., Fe, Cu, Cr, and Ni) and OH radicals associated with water vapor dissolved in the glass. Most metal impurities can be readily removed. Initially, OH impurities proved to be more difficult to eliminate but methods were ultimately developed to do so. Wavelengths at which glass fibers are used for optical fiber communications were traditionally selected to avoid the OH absorption bands.

Light-scattering losses may also be accentuated when dopants are added, as they often are for purposes of index grading. The attenuation coefficient for guided light in glass fibers depends on the absorption and scattering in the core and cladding materials. Each mode has a different penetration depth into the cladding, causing the rays to travel different effective distances and rendering the attenuation coefficient mode-dependent. It is generally higher for higher-order modes. Single-mode fibers therefore typically have smaller attenuation coefficients than multimode fibers (Fig. 10.3-3). Losses are also introduced by small random variations in the geometry of the fiber and by bends.

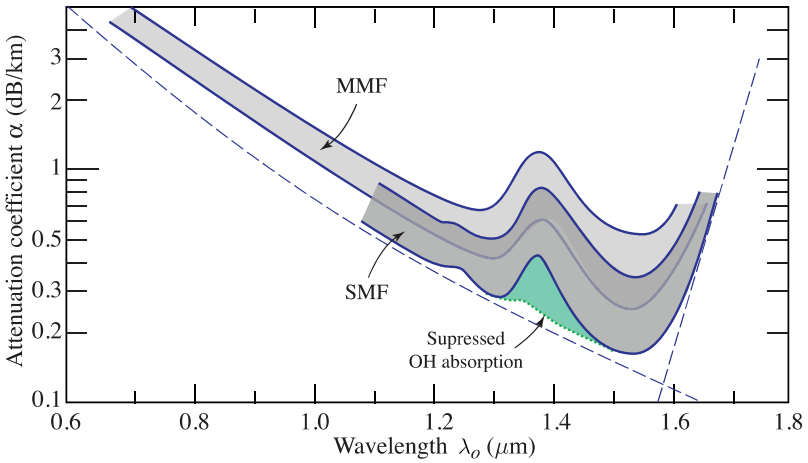


Figure 10.3-3 Ranges of attenuation coefficients for silica-glass single-mode fibers (SMF) and multimode fibers (MMF).

B. Dispersion

When a short pulse of light travels through an optical fiber, its power is "dispersed" in time so that the pulse spreads into a wider time interval. There are five principal sources of dispersion in optical fibers:

- Modal dispersion
- Material dispersion
- Waveguide dispersion
- Polarization mode dispersion
- Nonlinear dispersion

The combined contributions of these effects to the spread of pulses in time are not necessarily additive, as will be understood in the sequel.

Modal Dispersion

Modal dispersion occurs in multimode fibers as a result of the differences in the group velocities of the various modes. A single impulse of light entering an M-mode fiber at $z = 0$ disperses into M pulses whose differential delay increases as a function of z. For a fiber of length L, the time delays engendered by the different velocities are $\tau_q = L/v_q$, $q = 1, \ldots, M$, where v_q is the group velocity of mode q. If $v_{\min}$ and $v_{\max}$ are the smallest and largest group velocities, respectively, the received pulse spreads over a time interval $L/v_{\min} - L/v_{\max}$. Since the modes are usually not excited equally, the overall shape of the received pulse generally has a smooth envelope, as illustrated in Fig. 10.3-4. An estimate of the overall pulse duration (assuming a triangular envelope and using the FWHM definition of the width) is $\sigma_\tau = \frac{1}{2}(L/v_{\min} - L/v_{\max})$, which represents the modal-dispersion response time of the fiber.

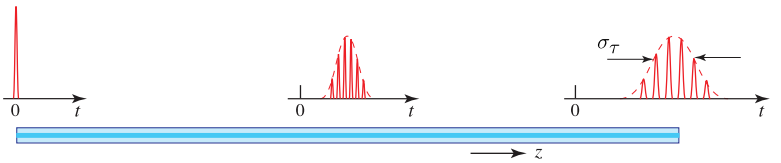


Figure 10.3-4 Pulse spreading caused by modal dispersion.

In a step-index fiber with a large number of modes, $v_{\min} \approx c_1(1-\Delta)$ and $v_{\max} \approx c_1$ [see Sec. 10.2C and Fig. 10.2-12(a)]. Since $(1-\Delta)^{-1} \approx 1+\Delta$ for $\Delta \ll 1$, the response time turns out to be a fraction $\Delta/2$ of the delay time L/c_1:

$$\sigma_\tau \approx \frac{L}{c_1} \cdot \frac{\Delta}{2}. \tag{10.3-4}$$

Response Time (Multimode Step-Index)

Modal dispersion is far smaller in graded-index (GRIN) fibers than in step-index fibers since the group velocities are equalized and the differences between the delay times of the modes, $\tau_q = L/v_q$, are reduced. It was shown in Sec. 10.2C and in Fig. 10.2-12(b) that a graded-index fiber with an optimal index profile and a large number of modes has $v_{\max} \approx c_1$ and $v_{\min} \approx c_1(1 - \Delta^2/2)$. The response time in this case is therefore a factor of $\Delta/2$ smaller than that in a step-index fiber:

$$\sigma_\tau \approx \frac{L}{c_1} \cdot \frac{\Delta^2}{4}. \tag{10.3-5}$$

Response Time (Graded-Index)

EXAMPLE 10.3-1. ***Multimode Pulse Broadening Rate.*** In a step-index fiber with $\Delta = 0.01$ and $n = 1.46$, pulses spread at a rate of approximately $\sigma_\tau/L = \Delta/2c_1 = n_1\Delta/2c_o \approx 24$ ns/km. In a 100-km fiber, therefore, an impulse spreads to a width of ≈ 2.4 μs. If the same fiber is optimally index-graded, the pulse broadening rate is approximately $n_1\Delta^2/4c_o \approx 122$ ps/km, a substantial reduction.

The pulse broadening arising from modal dispersion is proportional to the fiber length L in both step-index and GRIN fibers. Because of mode coupling, however, this dependence does not necessarily apply for fibers longer than a certain critical length. Coupling occurs among modes that have approximately the same propagation constants as a result of small imperfections in the fiber, such as random irregularities at its surface or inhomogeneities in its refractive index. This permits optical power to be exchanged between the modes. Under certain conditions, the response time σ_τ of mode-coupled fibers is proportional to L for small fiber lengths and to $\sqrt{L}$ when a critical length is exceeded, whereupon the pulses are broadened at a reduced rate.[†]

Material Dispersion

Glass is a dispersive medium, i.e., its refractive index is a function of wavelength. As discussed in Sec. 5.7, an optical pulse travels in a dispersive medium of refractive index n with a group velocity $v = c_o/N$, where $N = n - \lambda_o\, dn/d\lambda_o$. Since the pulse is a wavepacket comprising a collection of components of different wavelengths, each traveling at a different group velocity, its width spreads. The temporal duration of an optical impulse of spectral width σ_λ (nm), after traveling a distance L through a dispersive material, is $\sigma_\tau = |(d/d\lambda_o)(L/v)|\sigma_\lambda = |(d/d\lambda_o)(LN/c_o)|\sigma_\lambda$. This leads to a response time [see (5.7-2), (5.7-7), and (5.7-8)]

$$\boxed{\sigma_\tau = |D_\lambda|\sigma_\lambda L,} \qquad \begin{matrix}(10.3\text{-}6)\\ \text{Response Time}\\ \text{(Material Dispersion)}\end{matrix}$$

where the material dispersion coefficient D_λ is

$$D_\lambda = -\frac{\lambda_o}{c_o}\frac{d^2n}{d\lambda_o^2}. \qquad (10.3\text{-}7)$$

The response time increases linearly with the distance L. Usually, L is measured in km, σ_τ in ps, and σ_λ in nm, so that D_λ has units of ps/km-nm. This type of dispersion is called **material dispersion**.

The wavelength dependence of the dispersion coefficient D_λ for a silica-glass fiber is displayed in Fig. 10.3-5. At wavelengths shorter than 1.3 μm the dispersion coefficient is negative, so that wavepackets of long wavelength travel faster than those of short wavelength. At a wavelength $\lambda_o = 0.87$ μm, for example, the dispersion coefficient D_λ is approximately -80 ps/km-nm. At $\lambda_o = 1.55$ μm, on the other hand, $D_\lambda \approx +17$ ps/km-nm. At $\lambda_o \approx 1.312$ μm the dispersion coefficient vanishes, so that σ_τ in (10.3-6) vanishes. A more precise expression for σ_τ that incorporates the spread of the spectral width σ_λ about $\lambda_o = 1.312$ μm yields a very small, but nonzero, width.

[†] See, e.g., J. E. Midwinter, *Optical Fibers for Transmission*, Wiley, 1979; Krieger, reissued 1992.

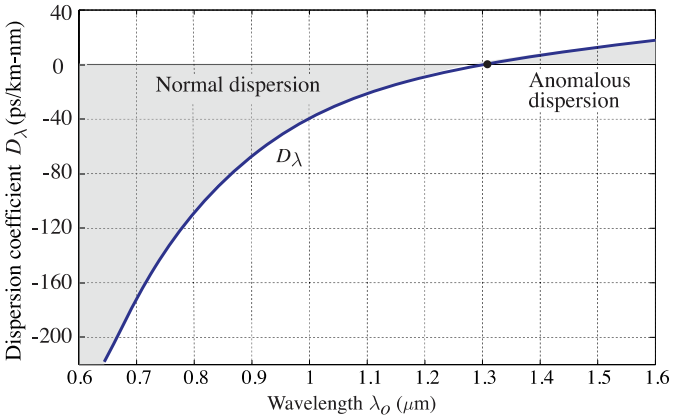


Figure 10.3-5 Observed dispersion coefficient D_λ for silica-glass fiber as a function of the wavelength λ_o. The result differs slightly from that calculated for bulk fused silica (see Fig. 5.7-5).

EXAMPLE 10.3-2. ***Pulse Broadening Associated with Material Dispersion.*** The dispersion coefficient D_λ for a silica-glass fiber is approximately -80 ps/km-nm at $\lambda_o = 0.87$ μm. For a source of spectral linewidth $\sigma_\lambda = 50$ nm (generated by an LED, for example) the pulse-spread rate in a single-mode fiber with no other sources of dispersion is $|D_\lambda|\sigma_\lambda = 4$ ns/km. An impulse of light traveling a distance $L = 100$ km in the fiber is therefore broadened to a width $\sigma_\tau = |D_\lambda|\sigma_\lambda L = 0.4$ μs. The response time of the fiber is thus $\sigma_\tau = 0.4$ μs. As another example, an impulse with narrower spectral linewidth $\sigma_\lambda = 2$ nm (generated by a laser diode, for example), operating near 1.3 μm where the dispersion coefficient is 1 ps/km-nm, spreads at a rate of only 2 ps/km. In this case, therefore, a 100-km fiber has a substantially shorter response time, $\sigma_\tau = 0.2$ ns.

Combined Material and Modal Dispersion

The effect of material dispersion on pulse broadening in multimode fibers may be determined by returning to the original equations for the propagation constants β_q of the modes and determining the group velocities $v_q = (d\beta_q/d\omega)^{-1}$ with n_1 and n_2 provided as functions of ω. Consider, for example, the propagation constants of a graded-index fiber with a large number of modes, which are given by (10.2-37) and (10.2-33). Although n_1 and n_2 are dependent on ω, it is reasonable to assume that the ratio $\Delta = (n_1 - n_2)/n_1$ is approximately independent of ω. Using this approximation and evaluating $v_q = (d\beta_q/d\omega)^{-1}$, we obtain

$$v_q \approx \frac{c_o}{N_1}\left[1 - \frac{p-2}{p+2}\left(\frac{q}{M}\right)^{p/(p+2)}\Delta\right], \tag{10.3-8}$$

where $N_1 = (d/d\omega)(\omega n_1) = n_1 - \lambda_o(dn_1/d\lambda_o)$ is the group index of the core material. Under this approximation, the earlier expression (10.2-39) for v_q remains intact, except that the refractive index n_1 is replaced with the group index N_1. For a step-index fiber $(p \to \infty)$, the group velocities of the modes vary from c_o/N_1 to $(c_o/N_1)(1-\Delta)$, so that the response time is

$$\boxed{\sigma_\tau \approx \frac{L}{(c_o/N_1)} \cdot \frac{\Delta}{2}.} \tag{10.3-9}$$

Response Time (Multimode Step-Index, Material Dispersion)

This expression should be compared with (10.3-4), which is applicable in the absence of material dispersion.

EXERCISE 10.3-1

Optimal Grade Profile Parameter. Use (10.2-37) and (10.2-33) to derive the following expression for the group velocity v_q when both n_1 and Δ are wavelength dependent:

$$v_q \approx \frac{c_o}{N_1}\left[1 - \frac{p-2-p_s}{p+2}\left(\frac{q}{M}\right)^{p/(p+2)}\Delta\right], \qquad q = 1, 2, \ldots, M \tag{10.3-10}$$

with $p_s = 2(n_1/N_1)(\omega/\Delta)\, d\Delta/d\omega$. What is the optimal value of the grade profile parameter p for minimizing modal dispersion?

Waveguide Dispersion

The group velocities of the modes in a waveguide are dependent on the wavelength even if material dispersion is negligible. This dependence, known as **waveguide dispersion**, results from the dependence of the field distribution in the fiber on the ratio of the core radius to the wavelength (a/λ_o). The relative portions of optical power in the core and cladding thus depend on λ_o. Since the phase velocities in the core and cladding differ, the group velocity of the mode is altered. Waveguide dispersion is particularly important in single-mode fibers where modal dispersion is not present, and at wavelengths for which material dispersion is small (near $\lambda_o = 1.3\ \mu\text{m}$ in silica glass), since it then dominates.

As discussed in Sec. 10.2A, the group velocity $v = (d\beta/d\omega)^{-1}$ and the propagation constant β are determined from the characteristic equation, which is governed by the fiber V parameter, $V = 2\pi(a/\lambda_o)\text{NA} = (a \cdot \text{NA}/c_o)\omega$. In the absence of material dispersion (i.e., when NA is independent of ω), V is directly proportional to ω, so that

$$\frac{1}{v} = \frac{d\beta}{d\omega} = \frac{d\beta}{dV}\frac{dV}{d\omega} = \frac{a \cdot \text{NA}}{c_o}\frac{d\beta}{dV}. \tag{10.3-11}$$

The pulse broadening associated with a source of spectral width σ_λ is related to the time delay L/v by $\sigma_\tau = |(d/d\lambda_o)(L/v)|\sigma_\lambda$. Thus,

$$\sigma_\tau = |D_w|\sigma_\lambda L, \tag{10.3-12}$$

where the waveguide dispersion coefficient D_w is given by

$$D_w = \frac{d}{d\lambda_o}\left(\frac{1}{v}\right) = -\frac{\omega}{\lambda_o}\frac{d}{d\omega}\left(\frac{1}{v}\right). \tag{10.3-13}$$

Substituting (10.3-11) into (10.3-13) leads to

$$D_w = -\left(\frac{1}{2\pi c_o}\right) V^2 \frac{d^2\beta}{dV^2}. \tag{10.3-14}$$

Thus, the group velocity is inversely proportional to $d\beta/dV$ and the waveguide dispersion coefficient is proportional to $V^2 d^2\beta/dV^2$. The dependence of β on V is displayed in Fig. 10.2-6(*a*) for the fundamental LP_{01} mode. Since β varies nonlinearly

with V, the waveguide dispersion coefficient D_w is itself a function of V and is therefore also a function of the wavelength.[†] The dependence of D_w on λ_o may be controlled by altering the radius of the core or, for graded-index fibers, the index grading profile.

Combined Material and Waveguide Dispersion (Chromatic Dispersion)

The combined effects of material dispersion and waveguide dispersion (which we refer to as **chromatic dispersion**) may be determined by including the wavelength dependence of the refractive indices, n_1 and n_2 and therefore NA, when determining $d\beta/d\omega$ from the characteristic equation. Although generally smaller than material dispersion, waveguide dispersion does shift the wavelength at which the total chromatic dispersion is minimum.

Since chromatic dispersion limits the performance of single-mode fibers, more advanced fiber designs aim at reducing this effect by using graded-index cores with refractive-index profiles selected such that the wavelength at which waveguide dispersion compensates material dispersion is shifted to the wavelength at which the fiber is to be used (Fig. 10.3-6).

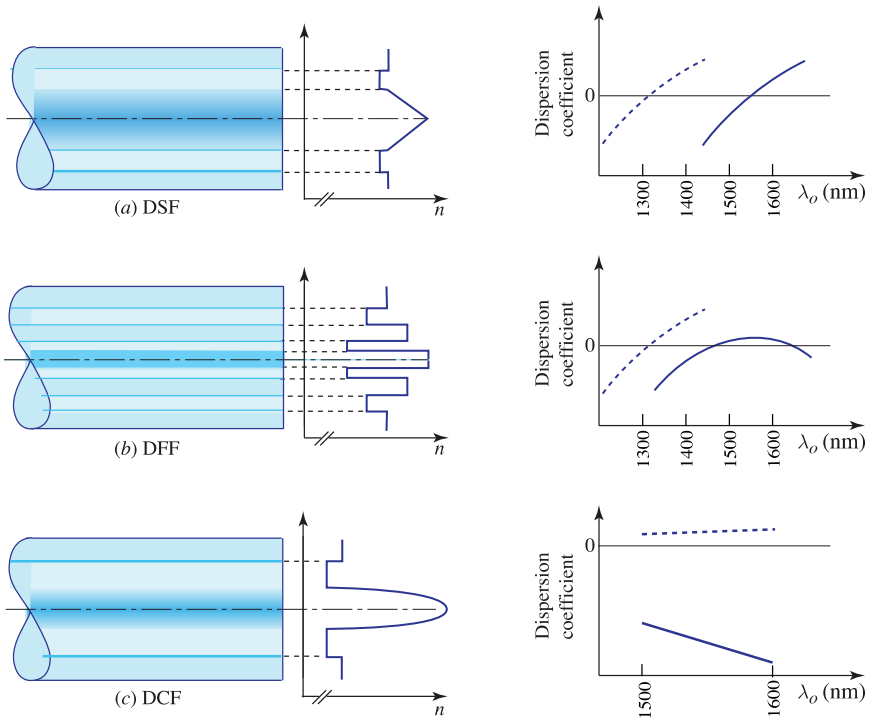


Figure 10.3-6 Refractive-index profiles with schematic wavelength dependences of the silica-glass material dispersion coefficient (dashed curves) and the combined material and waveguide dispersion coefficients (solid curves) for (a) dispersion-shifted fiber (DSF); (b) dispersion-flattened fiber (DFF); and (c) dispersion-compensating fiber (DCF).

Dispersion-shifted fibers have been successfully fabricated by using a linearly tapered core refractive index and a reduced core radius, as illustrated in Fig. 10.3-6(a). This technique can be used to shift the zero-chromatic-dispersion wavelength

[†] For further details on this topic, see the reading list, particularly the seminal articles by D. Gloge.

from 1.3 μm to 1.55 μm, where silica-glass fiber has its lowest attenuation. Other grading profiles have been developed for which the chromatic dispersion vanishes at two wavelengths and is reduced for intermediate wavelengths. These fibers, called **dispersion-flattened**, have been implemented by using a quadruple-clad layered grading, as illustrated in Fig. 10.3-6(b). Note, however, that the process of index grading itself introduces losses since dopants are used.

Fibers with other refractive index profiles may be engineered such that the combined material and waveguide dispersion coefficient is proportional to that of a conventional step-index fiber but has the opposite sign. This can be achieved over an extended wavelength band, as illustrated in Fig. 10.3-6(c). The pulse spread introduced by a conventional fiber can then be reversed by concatenating the two types of fiber. A fiber with a reversed dispersion coefficient is known as a **dispersion-compensating fiber** (**DCF**). A short segment of the DCF may be used to compensate the dispersion introduced by a long segment of conventional fiber.

Polarization Mode Dispersion (PMD)

As indicated earlier, the fundamental spatial mode (LP_{01}) of an optical fiber has two polarization modes, say linearly polarized in the x and y directions. If the fiber has perfect circular symmetry about its axis, and its material is perfectly isotropic, then the two polarization modes are degenerate, i.e., they travel with the same velocity. However, fibers exposed to real environmental conditions exhibit a small birefringence that varies randomly along their length. This is caused by slight variations in the refractive indices and fiber cross-section ellipticity. Although the effects of such inhomogeneities and anisotropies on the polarization modes, and on the dispersion of optical pulses, are generally difficult to assess, we consider these effects in terms of simple models.

Consider first a fiber modeled as a homogeneous anisotropic medium with principal axes in the x and y directions and principal refractive indices n_x and n_y. The third principal axis lies along the fiber axis (the z direction). The fiber material is assumed to be dispersive so that n_x and n_y are frequency dependent, but the principal axes are taken to be frequency independent within the spectral band of interest. If the input pulse is linearly polarized in the x direction, over a length of fiber L it will undergo a group delay $\tau_x = N_x L/c_o$; if it is linearly polarized in the y direction, the group delay will be $\tau_y = N_y L/c_o$. Here, N_x and N_y are the group indices associated with n_x and n_y (see Sec. 5.7). A pulse in a polarization state that includes both linear polarizations will undergo a **differential group delay** (**DGD**) $\delta\tau = |\tau_y - \tau_x|$ given by

$$\boxed{\delta\tau = \Delta N L/c_o,} \qquad (10.3\text{-}15)$$

Differential Group Delay

where $\Delta N = |N_y - N_x|$. Upon propagation, therefore, the pulse will split into two orthogonally polarized components whose centers will separate in time as the pulses travel (see Fig. 10.3-7). The DGD corresponds to **polarization mode dispersion** (PMD) that increases linearly with the fiber length at the rate $\Delta N/c_o$, which is usually expressed in units of ps/km.

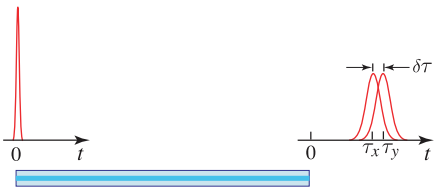


Figure 10.3-7 Differential group delay (DGD) associated with polarization mode dispersion (PMD).

Since a long fiber is typically exposed to environmental and structural factors that vary along its axis, the simple model considered above is often inadequate, however. Under these conditions, a more realistic model comprises a sequence of short homogeneous fiber segments, each with its own principal axes and principal indices. The principal axes are taken to be slightly misaligned (rotated) from one segment to the next. Such a cascaded system is generally described by a 2×2 Jones matrix $\mathbf{T}$, which is a product of the Jones matrices of the individual segments (see Sec. 6.1B). The polarization modes of the combined system are the eigenvectors of $\mathbf{T}$ and are not necessarily linearly polarized modes. If the fiber is taken to be lossless, the matrix $\mathbf{T}$ is unitary. Its eigenvalues are then phase factors $\exp(j\varphi_1)$ and $\exp(j\varphi_2)$, which may be written in the form $\exp(jn_1k_oL)$ and $\exp(jn_2k_oL)$, where n_1 and n_2 are the effective refractive indices of the two polarization modes and L is the fiber length. The propagation of light through such a length of fiber may then be determined by analyzing the input wave into components along the two polarization modes; these components travel with effective refractive indices n_1 and n_2.

Since the fiber is dispersive, $\mathbf{T}$ is frequency dependent and so too are the indices n_1 and n_2 of the modes, as well as their corresponding group indices N_1 and N_2. An input pulse with a polarization state that is the same as that of the fiber's first polarization mode travels with an effective group index N_1. Similarly, if the pulse is in the second polarization mode, it travels with an effective group index N_2. However, an input pulse with a component in each of the fiber's polarization modes suffers DGD, as provided in (10.3-15), with $\Delta N = |N_1 - N_2|$.

A statistical model describing the random variations in the magnitude and orientation of birefringence along the length of the fiber leads to an expression of the RMS value of the pulse broadening associated with DGD. This turns out to be proportional to $\sqrt{L}$ instead of L,

$$\boxed{\sigma_{\text{PMD}} = D_{\text{PMD}}\sqrt{L},} \qquad \begin{matrix}(10.3\text{-}16)\\ \text{Polarization Mode Dispersion}\end{matrix}$$

where D_{PMD} is a dispersion parameter typically ranging from 0.1 to 1 ps/$\sqrt{\text{km}}$.

Aside from DGD, higher-order dispersion effects are also present. Each of the polarization modes is spread by group velocity dispersion (GVD) with dispersion coefficients proportional to the second derivative of its refractive index (see Sec. 5.7).

Another higher-order effect relates to the coupled nature of the spectral and polarization properties of the system. Since the matrix $\mathbf{T}$ is frequency dependent, not only are the eigenvalues (i.e., the principal indices n_1 and n_2) frequency dependent, but so too are the eigenvectors (i.e., the polarization modes). If the pulse spectral width is sufficiently narrow (i.e., the pulse is not too short), we may approximately use the polarization modes at the central frequency. For ultrashort pulses, however, a more detailed analysis that includes a combined polarization and spectral description of the system is required. Polarization states may be found such that the group delays are frequency insensitive so that their associated GVD is minimal. However, these are not eigenvectors of the Jones matrix so that the input and output polarization states are not the same.[†]

[†] For further details on this topic, see C. D. Poole and R. E. Wagner, Phenomenological Approach to Polarization Dispersion in Long Single-Mode Fibers, *Electronics Letters*, vol. 22, pp. 1029-1030, 1986.

EXERCISE 10.3-2

Differential Group Delay in a Two-Segment Fiber. Consider the propagation of an optical pulse through a fiber of length 1 km comprising two segments of equal length. Each segment is a single-mode anisotropic fiber with principal group indices $N_x = 1.462$ and $N_y = 1.463$. The corresponding group velocity dispersion coefficients are $D_x = D_y = 20$ ps/km-nm. The principal axes of one segment are at an angle of 45° with respect to the other, as illustrated in Fig. 10.3-8.

(a) If the input pulse has a width of 100 ps and is linearly polarized at 45° with respect to the fiber x and y directions, sketch the temporal profile of the pulse at the output end of the fiber. Assume that the pulse source has a spectral linewidth of 50 nm.

(b) Determine the polarization modes of the full fiber and determine the temporal profile of the output pulse if the input pulse is in one of the polarization modes.

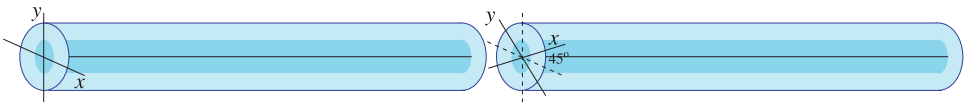


Figure 10.3-8 Two-segment birefringent fiber.

Nonlinear Dispersion

Yet another dispersion effect occurs when the intensity of light in the core of the fiber is sufficiently high, since the refractive index then becomes intensity dependent and the material exhibits nonlinear behavior. Since the phase is proportional to the refractive index, the high-intensity portions of an optical pulse undergo phase shifts different from the low-intensity portions, resulting in instantaneous frequencies shifted by different amounts. This nonlinear effect, called **self-phase modulation** (SPM), contributes to pulse dispersion. Under certain conditions, SPM can compensate the group velocity dispersion (GVD) associated with material dispersion, and the pulse can travel without altering its temporal profile. Such a guided wave is known as a soliton. Nonlinear optics is introduced in Chapter 22 and optical solitons are discussed in Chapter 23.

Summary

The propagation of pulses in optical fibers is governed by attenuation and several types of dispersion. Figure 10.3-9 provides a schematic illustration in which the profiles of pulses traveling through different types of fibers are compared.

- In a multimode fiber (MMF), modal dispersion dominates the width of the pulse received at the terminus of the fiber. It is governed by the disparity in the group delays of the individual modes.
- In a single-mode fiber (SMF), there is no modal dispersion and the transmission of optical pulses is limited by combined material and waveguide dispersion (called chromatic dispersion). The width of the output pulse is governed by group velocity dispersion (GVD).
- Material dispersion is usually much stronger than waveguide dispersion. However, at wavelengths where material dispersion is small, waveguide dispersion becomes important. Fibers with special index profiles may then be used to alter the chromatic-dispersion characteristics, creating dispersion-flattened, dispersion-shifted, and dispersion-compensating fibers.

- Pulse propagation in long single-mode fibers for which chromatic dispersion is negligible is dominated by polarization mode dispersion (PMD). Small anisotropic changes in the fiber, caused, for example, by environmental conditions, alter the polarization modes so that the input pulse travels in two polarization modes with different group indices. This differential group delay (DGD) results in a small pulse spread.
- Under certain conditions an intense pulse, called an optical soliton, can render a fiber nonlinear and travel through it without broadening. This results from a balance between material dispersion and self-phase modulation (the dependence of the refractive index on the light intensity), as discussed in Chapter 23.

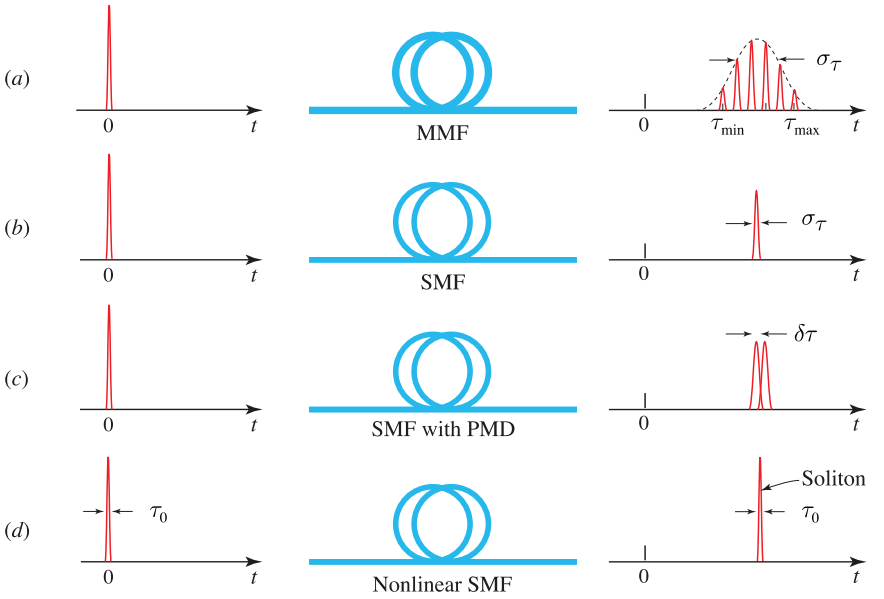


Figure 10.3-9 Broadening of a short optical pulse after transmission through different types of optical fibers. (*a*) Modal dispersion in a multimode fiber (MMF). (*b*) Material and waveguide dispersion (chromatic dispersion) in a single-mode fiber (SMF). (*c*) Polarization mode dispersion (PMD) in a SMF. (*d*) Soliton transmission in a nonlinear SMF.

10.4 HOLEY AND PHOTONIC-CRYSTAL FIBERS

A **holey fiber** is a fiber that contains multiple cylindrical air holes parallel to, and along the length of, its axis. Usually fabricated from pure silica glass, its holes are organized in a regular periodic pattern. As illustrated in Fig. 10.4-1, the core of the fiber is defined by a **defect**, or fault, in the periodic structure, such as a missing hole, a hole of a different size, or an extra hole. The holes are characterized by the spacing between their centers Λ, and their diameters d. The quantity Λ, which is also called the **pitch**, is typically in the range 1–10 μm. It is not necessary to include dopants in the glass. Holey fibers guide optical waves via one of two mechanisms: **effective-index guidance** or **photonic-bandgap guidance**, which we consider in turn below.

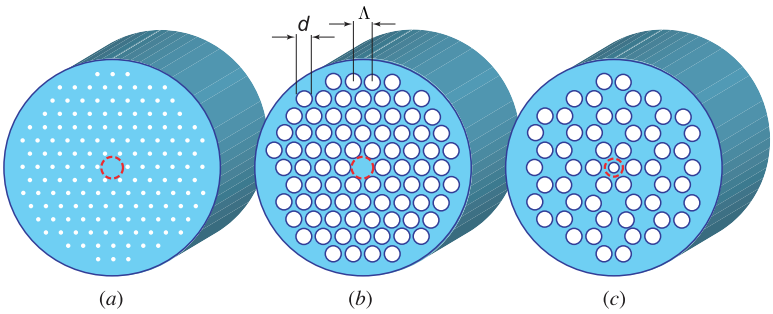


Figure 10.4-1 Various forms of holey fibers. (a) Solid core (dotted circle) surrounded by a cladding of the same material but suffused with an array of cylindrical air holes with diameters much smaller than a wavelength. The average refractive index of the cladding is lower than that of the core. (b) Photonic-crystal holey fiber with cladding that contains a periodic array of large air holes and a solid core (dotted circle). (c) Photonic-crystal holey fiber with cladding that contains a periodic array of large air holes and a core that is an air hole of a different size (dotted circle).

Effective-Index Guidance

If the hole diameter is much smaller than the wavelength of light ($d \ll \lambda$), then the periodic cladding behaves approximately as a homogeneous medium whose **effective refractive index** n_2 is equal to the average refractive index of the holey material [see Fig. 10.4-1(a)]. Waveguiding is then achieved by making use of a solid core with index $n_1 > n_2$, so that the light is guided by total internal reflection as with conventional optical fibers. In this configuration, the holes serve merely as distributed "negative dopants" that reduce the refractive index of the cladding below that of the solid core. The holes can therefore be randomly, rather than periodically, arrayed and they need not be axially continuous.

If the size of the holes is not much smaller than the wavelength, then the holey cladding must be treated as a two-dimensional periodic medium [Fig. 10.4-1(b)]. The effective refractive index n_2 is then given by the average refractive index, weighted by the optical intensity distribution in the medium, and is therefore strongly dependent on the wavelength, as well as on the size and the geometry of the holes. Since waves of shorter wavelength are more confined in the medium of higher refractive index, the effective refractive index of the cladding $n_2(\lambda)$ is a decreasing function of the wavelength. A similar effect occurs in a 1D photonic crystal, for which the effective refractive index is an increasing function of frequency at frequencies in the lowest photonic band (see Fig. 7.2-6). The holey fiber is therefore endowed with strong waveguide dispersion, which can be a useful feature.

One consequence of the waveguide dispersion is that the holey fiber may operate as a single-mode structure over a broad range of wavelengths, possibly stretching from the infrared to the ultraviolet.[†] This property, called **endless single-mode guidance**, results when the V parameter of the fiber, $V = (2\pi a/\lambda)\sqrt{n_1^2 - n_2^2(\lambda)}$, is approximately independent of λ. This condition arises when the effective index $n_2(\lambda)$ decreases with increasing λ in such a way that $\sqrt{n_1^2 - n_2^2(\lambda)} \propto \lambda$. For a conventional optical fiber, in contrast, V is inversely proportional to λ so that single-mode behavior at a particular wavelength ($V < 2.405$) morphs into multimode behavior for wavelengths that are sufficiently short such that V exceeds 2.405.

Another interesting feature is the feasibility of achieving **large mode-area** (LMA) single-mode operation. Optical fibers with large mode areas are useful for applications

[†] See T. A. Birks, J. C. Knight, and P. St J. Russell, Endlessly Single-Mode Photonic Crystal Fibre, *Optics Letters*, vol. 22, pp. 961–963, 1997.

requiring the delivery of high optical powers. In a conventional optical fiber the condition for single-mode operation [$V = 2\pi(a/\lambda_o)\mathrm{NA} < 2.405$] can be met for a large core diameter $2a$ by making use of a small numerical aperture. Similarly, in holey fibers the guided-mode size can be increased by increasing the hole-to-hole spacing Λ (thereby increasing the core diameter) and concomitantly using holes of smaller diameter d (thereby reducing the numerical aperture and allowing the field to penetrate farther into the cladding). Dramatic increases in mode area for relatively small changes in hole size have been observed and mode areas several order of magnitudes greater than those in conventional optical fibers have been reported.

Photonic-Bandgap Guidance

The cladding of a holey fiber may be regarded as a two-dimensional **photonic crystal**. The triangular-hole microstructure shown in Fig. 10.4-1(b), for example, has a dispersion diagram endowed with photonic bandgaps, as shown in Fig. 7.3-3 and discussed in Sec. 7.3A. If the optical frequency lies within the photonic bandgap, propagation through the cladding is prohibited and the fiber serves as a photonic-crystal waveguide (see Sec. 9.5).

A **photonic-crystal fiber** (PCF) may have a solid or hollow core, as illustrated in Figs. 10.4-1(b) and (c), respectively. Fibers with hollow cores cannot operate by means of effective-index guidance, i.e., guidance cannot be based on total internal reflection. In any case, there are a number of merits to using hollow-core PCFs: (1) A guided wave traveling in an air-core PCF suffers lower losses as well as reduced nonlinear effects and can thus carry greater optical power; (2) light in the mid-ultraviolet region that would cause damage and degradation in solid-core fibers can be guided and transmitted; (3) light can be guided at wavelengths where transparent materials are not available; and (4) atoms, molecules, gases, and other subwavelength structures can be placed within the hollow core.

Applications

Photonic-crystal fibers offer many unique design possibilities and applications. As an example, dispersion flattening over broad wavelength ranges can be achieved, and the dispersion can be shifted to wavelengths below that at which the material dispersion is zero. PCFs support the propagation of high-intensity femtosecond pulses and their compression to sub-carrier-cycle durations. Long interaction lengths and tight confinement permit nonlinear optical effects such as low-threshold stimulated Raman scattering, harmonic generation, and electromagnetic-induced transparency to be investigated. The birefringence properties of such fibers can be tailored in interesting and useful ways. Acousto-optic interactions can be fostered and examined. PCFs can be used to construct broadband (mid infrared to mid ultraviolet) supercontinuum sources, mode-locked fiber soliton lasers, and powerful fiber lasers operating over a broad range of wavelengths.

Photonic-crystal fibers can be readily used as dynamic sensors for strain, temperature, and electric field. Pressure and heat serve to modify the sizes of the air holes, which in turn modifies their optical properties. Hollow-core fibers can be used to inspect, characterize, manipulate, trap, and accelerate objects such as living cells, colloids, clusters, and nm-size particles. Gases, liquids, or metabolites can be diffused into a hollow-core PCF, where light is trapped, thereby allowing ultrasensitive optical measurements to be carried out. Moreover, the hollow channels within the PCFs can be filled with materials such as metals, semiconductors, or soft glasses. This allows nanometer-scale features to be incorporated in the fibers, thereby leading to devices with both fiber-optic and plasmonic features.

10.5 FIBER MATERIALS

As discussed throughout this chapter, a principal use of optical fibers is for near-infrared optical fiber communications and data-transmission systems. The preeminent material for fabricating these fibers is silica glass since this medium exhibits low loss in the 1.3–1.6-μm telecommunications band, as described in Sec. 10.3A. **Specialty fibers**, which are optical fibers endowed with at least one special property that distinguishes them from standard silica-glass optical fibers, are also used in a many applications. Such fibers may, for example, be double-clad, polarization-maintaining, radiation-resistant, or doped with laser-active rare-earth ions.

Mid-Infrared Fibers

Interest in fiber optics extends beyond the near-infrared region. The mid infrared, for example, offers a substantial number of material choices, including fluoride, germanate, tellurite, and chalcogenide glasses. Of these, the chalcogenides (which include sulfides, selenides, and tellurides) exhibit the broadest transparency windows. The melting/processing temperatures of these so called *soft glasses* are substantially lower than that of silica glass so they can be conveniently used to fabricate infrared fibers via thermal fiber drawing; indeed, fibers fabricated from these three glasses are commercially available.

Interest in these materials is significant since the Rayleigh inverse fourth-power law (see Secs. 5.6B and 10.3A) predicts a reduction of Rayleigh scattering, and absorption further into the infrared, than that in silica glass. The attenuation arising from Rayleigh scattering in fluoride-glass infrared fibers, for example, is expected to be approximately ten times smaller than that for silica-glass optical fibers, reaching a minimum of ≈ 0.01 dB/km at $\lambda_o \approx 2.5$ μm. This issue is, of course, significant only when extrinsic loss mechanisms do not dominate transmission loss.

In spite of the extensive choice of optical fiber materials in the mid infrared, in the current state of technology the optical and mechanical properties of these fibers are substantially inferior to those of conventional silica optical fibers. With the notable exception of fluoride-glass fibers, most mid-infrared fibers exhibit transmission losses in the dB/m range, whereas silica-glass fibers are far more transparent with losses in the dB/km range. The use of most mid-infrared fibers is thus currently limited to applications that involve fiber lengths of the order of meters rather than kilometers, i.e., to short-haul applications such as sensing, metrology, biomedicine, and the delivery of infrared-laser power. These limitations are ultimately expected to be overcome, however.

Multimaterial and Multifunctional Fibers

Hybrid mid-infrared fibers. The ability to co-draw fibers comprising heterogeneous materials has led to the development of hybrid fibers that offer promise for ameliorating some of the optical and mechanical shortcomings of conventional mid-infrared fibers discussed above. Several examples of hybrid mid-infrared fibers are:[†] (1) Step-index fibers comprising a chalcogenide glass core with claddings of polymer (as a protective jacket), silica-glass, or tellurite-glass; (2) Step-index fibers comprising a semiconductor crystalline core such as InSb, ZnSe, Si, or Ge, with borosilicate-glass or silica-glass cladding. In connection with photonic-crystal fibers (PCFs), as discussed in Sec. 10.4, examples are: (3) Chalcogenide or fluoride glass solid-core PCFs; and (4) Silica or chalcogenide hollow-core PCFs.

[†] See G. Tao, H. Ebendorff-Heidepriem, A. M. Stolyarov, S. Danto, J. V. Badding, Y. Fink, J. Ballato, and A. F. Abouraddy, Infrared Fibers, *Advances in Optics and Photonics*, vol. 7, pp. 379–458, 2015.

Multimaterial fibers. Multimaterial fibers comprise combinations of materials. As alluded to in Sec. 10.4, optical fibers have been developed that incorporate not only glasses, but also conductors, semiconductors, insulators, and gases, along with more specialized materials such as liquid crystals and piezoelectric media. Nowadays, thermally drawn fibers that are kilometers long can be internally structured with elaborate macroscopic device architectures that incorporate these materials. Moreover, the ultimate composition of a fiber can differ from that of the initial materials by virtue of chemical reactions initiated by the heating and drawing processes (e.g., initial ingredients of aluminum metal and silica glass can result in a fiber whose core is crystalline silicon). Multimaterial fibers can be configured to sense light, sound, and/or heat impinging on their lateral surfaces. They can also be operated in reverse, i.e., to emit light and sound.

Multifunctional fibers and fiber assemblies. Multifunctional fibers, on the other hand, accommodate multitudinal functionalities. Beyond integrating multiple functional components into individual fibers, large-scale fiber arrays and fiber textiles can be fabricated. Multimaterial and multifunctional fibers and fiber assemblies offer a panoply of functionalities: photonic, electronic, mechanical, thermal, acoustic, chemical, and biomedical. Moreover, subwavelength internal structures of nanometer length scales can be organized within fibers of kilometer length scales, resulting in devices that merge fiber-optic and plasmonic features.

Examples. Some examples of multimaterial and multifunctional fibers are:[†]

- An axially pumped photonic-bandgap fiber laser, comprising a core of organic dye dissolved in a solid host as the gain medium, that emits light *radially*.
- Fibers engineered to detect sound by incorporating a piezoelectric plastic component along the length of the fiber (a pressure wave induces a charge in the piezoelectric material).
- Fibers designed to detect heat by incorporating a semiconductor component along the length of the fiber (temperature modifies the conductivity).
- A flexible nonimaging textile camera woven from fibers that incorporate a photoconductive-semiconductor or photoconductive-glass component along the length of the fiber (light modifies the conductivity).
- A textile display woven from fibers that incorporate liquid-crystal channels (an applied field serves to block or transmit light).
- A multifunctional polymer-fiber probe that allows concurrent optical, electrical, and chemical interactions with a cell in a neural circuit.[‡]

READING LIST

Books

See also the reading list on photonic crystals in Chapter 7 and the reading list in Chapter 9.
F. Mitschke, *Fiber Optics: Physics and Technology*, Springer-Verlag, 2nd ed. 2016.

[†] See A. F. Abouraddy, M. Bayindir, G. Benoit, S. D. Hart, K. Kuriki, N. Orf, O. Shapira, F. Sorin, B. Temelkuran, and Y. Fink, Towards Multimaterial Multifunctional Fibres that See, Hear, Sense and Communicate, *Nature Materials*, vol. 6, pp. 336–347, 2007; G. Tao, A. F. Abouraddy, and A. M. Stolyarov, Multimaterial Fibers, *International Journal of Applied Glass Science*, vol. 3, pp. 349–368, 2012.

[‡] See A. Canales, X. Jia, U. P. Froriep, R. A. Koppes, C. M. Tringides, J. Selvidge, C. Lu, C. Hou, L. Wei, Y. Fink, and P. Anikeeva, Multifunctional Fibers for Simultaneous Optical, Electrical and Chemical Interrogation of Neural Circuits *In Vivo*, *Nature Biotechnology*, vol. 33, pp. 277–284, 2015.

Y. Koike, *Fundamentals of Plastic Optical Fibers*, Wiley–VCH, 2015.

J. Hecht, *Understanding Fiber Optics*, Laser Light Press, 5th ed. 2015.

F. Zolla, G. Renversez, A. Nicolet, B. Kuhlmey, S. Guenneau, D. Felbacq, A. Argyros, and S. Leon-Saval, *Foundations of Photonic Crystal Fibres*, Imperial College Press (London), 2nd ed. 2012.

A. Kumar and A. Ghatak, *Polarization of Light with Applications in Optical Fibers*, SPIE Optical Engineering Press, 2011.

R. Paschotta, *Field Guide to Optical Fiber Technology*, SPIE Optical Engineering Press, 2010.

F. Poli, A. Cucinotta, and S. Selleri, *Photonic Crystal Fibers: Properties and Applications*, Springer-Verlag, 2007, paperback ed. 2010.

M. G. Kuzyk, *Polymer Fiber Optics: Materials, Physics, and Applications*, CRC Press/Taylor & Francis, 2007.

A. Méndez and T. F. Morse, eds., *Specialty Optical Fibers Handbook*, Academic Press, 2007.

C. DeCusatis and C. J. Sher DeCusatis, *Fiber Optic Essentials*, Academic Press/Elsevier, 2005.

A. Galtarossa and C. R. Menyuk, eds., *Polarization Mode Dispersion*, Springer-Verlag, 2005.

J. A. Harrington, *Infrared Fibers and Their Applications*, SPIE Optical Engineering Press, 2004.

A. Bjarklev, J. Broeng, and A. S. Bjarklev, *Photonic Crystal Fibers*, Springer-Verlag, 2003, paperback ed. 2012.

J. Hecht, *City of Light: The Story of Fiber Optics*, Oxford University Press, 1999, paperback ed. 2004.

C. K. Kao, *Optical Fiber Systems: Technology, Design, and Applications*, McGraw–Hill, 1982.

D. Marcuse, *Light Transmission Optics*, Van Nostrand Reinhold, 1972, 2nd ed. 1982; Krieger, reissued 1989.

D. Marcuse, *Principles of Optical Fiber Measurements*, Academic Press, 1981.

Review Articles on Photonic-Crystal, Multicore, Multimaterial, and Infrared Fibers

G. Tao, H. Ebendorff-Heidepriem, A. M. Stolyarov, S. Danto, J. V. Badding, Y. Fink, J. Ballato, and A. F. Abouraddy, Infrared Fibers, *Advances in Optics and Photonics*, vol. 7, pp. 379–458, 2015.

T. Hayashi, Multi-Core Optical Fibers, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

G. Tao, A. F. Abouraddy, and A. M. Stolyarov, Multimaterial Fibers, *International Journal of Applied Glass Science*, vol. 3, pp. 349–368, 2012.

A. F. Abouraddy, M. Bayindir, G. Benoit, S. D. Hart, K. Kuriki, N. Orf, O. Shapira, F. Sorin, B. Temelkuran, and Y. Fink, Towards Multimaterial Multifunctional Fibres that See, Hear, Sense and Communicate, *Nature Materials*, vol. 6, pp. 336–347, 2007.

P. Russell, Photonic Crystal Fibers, *Science*, vol. 299, pp. 358–362, 2003.

J. C. Knight, J. Broeng, T. A. Birks, and P. St. J. Russell, Photonic Band Gap Guidance in Optical Fibers, *Science*, vol. 282, pp. 1476–1478, 1998.

Seminal Articles

P. J. Winzer, C. J. Chang-Hasnain, A. E. Willner, R. C. Alferness, R. W. Tkach, and T. G. Giallorenzi, eds., *A Third of a Century of Lightwave Technology: January 1983–April 2016*, IEEE–OSA, 2016.

C. K. Kao, Sand from Centuries Past: Send Future Voices Fast (Nobel Lecture in Physics, 2009), in L. Brink, ed., *Nobel Lectures in Physics 2006–2010*, World Scientific, 2014, pp. 253–264.

D. B. Keck, ed., *Selected Papers on Optical Fiber Technology*, SPIE Optical Engineering Press (Milestone Series Volume 38), 1992.

J. P. Pocholle, L. Jeunhomme, and L. D'Auria, Caractéristiques de la propagation guidée dans les fibres multimodes à gradient d'indice (FMGI), *Revue technique Thomson-CSF*, vol. 13, pp. 727–807, 1981.

P. J. B. Clarricoats, Optical Fibre Waveguides – A Review, in E. Wolf, ed., *Progress in Optics*, North-Holland, 1976, vol. 14, pp. 327–402.

D. B. Keck, R. D. Maurer, and P. C. Schultz, On the Ultimate Lower Limit of Attenuation in Glass Optical Waveguides, *Applied Physics Letters*, vol. 22, pp. 307–309, 1973.

D. Gloge and E. A. J. Marcatili, Multimode Theory of Graded-Core Fibers, *Bell System Technical Journal*, vol. 52, pp. 1563–1578, 1973.

D. Gloge, Weakly Guiding Fibers, *Applied Optics*, vol. 10, pp. 2252–2258, 1971.

D. Gloge, Dispersion in Weakly Guiding Fibers, *Applied Optics*, vol. 10, pp. 2442–2445, 1971.

PROBLEMS

10.1-1 **Coupling Efficiency.**

(a) A source emits light with optical power P_0 and a distribution $I(\theta) = (1/\pi)P_0 \cos\theta$, where $I(\theta)$ is the power per unit solid angle in the direction making an angle θ with the axis of a fiber. Show that the power collected by the fiber is $P = (\mathrm{NA})^2 P_0$, so that the coupling efficiency is $(\mathrm{NA})^2$, where NA is the numerical aperture of the fiber.

(b) If the source is a planar light-emitting diode of refractive index n_s bonded to the fiber, and the fiber cross-sectional area is larger than the LED emitting area, calculate the numerical aperture of the fiber and the coupling efficiency when $n_1 = 1.46$, $n_2 = 1.455$, and $n_s = 3.5$.

10.1-2 **Numerical Aperture of a Graded-Index Fiber.** Compare the numerical apertures of a step-index fiber with $n_1 = 1.45$ and $\Delta = 0.01$ and a graded-index fiber with $n_1 = 1.45$, $\Delta = 0.01$, and a parabolic refractive-index profile ($p = 2$). (See Exercise 1.3-2.)

10.2-1 **Modes.** A step-index fiber has radius $a = 5$ μm, core refractive index $n_1 = 1.45$, and fractional refractive-index change $\Delta = 0.002$. Determine the shortest wavelength λ_c for which the fiber is a single-mode waveguide. If the wavelength is changed to $\lambda_c/2$, identify the indices (l, m) of all the guided modes.

10.2-2 **Modal Dispersion.** A step-index fiber of numerical aperture NA $= 0.16$, core radius $a = 45$ μm, and core refractive index $n_1 = 1.45$ is used at $\lambda_o = 1.3$ μm, where material dispersion is negligible. If a light pulse of very short duration enters the fiber at $t = 0$ and travels a distance 1 km, sketch the shape of the received pulse:

(a) using ray optics and assuming that only meridional rays are allowed;

(b) using wave optics and assuming that only meridional ($l = 0$) modes are allowed.

10.2-3 **Propagation Constants and Group Velocities.** A step-index fiber with refractive indices $n_1 = 1.444$ and $n_2 = 1.443$ operates at $\lambda_o = 1.55$ μm. Determine the core radius at which the fiber V parameter is 10. Use Fig. 10.2-3 to estimate the propagation constants of all the guided modes with $l = 0$. If the core radius is now changed so that $V = 4$, use Fig. 10.2-6(a) to determine the phase velocity, the propagation constant, and the group velocity of the LP_{01} mode. Ignore the effect of material dispersion.

*10.2-4 **Propagation Constants and Wavevector (Step-Index Fiber).** A step-index fiber of radius $a = 20$ μm and refractive indices $n_1 = 1.47$ and $n_2 = 1.46$ operates at $\lambda_o = 1.55$ μm. Using the quasi-plane wave theory and considering only guided modes with $l = 1$:

(a) determine the smallest and largest propagation constants;

(b) for the mode with the smallest propagation constant, determine the radii of the cylindrical shell within which the wave is confined, and determine the components of the wavevector $\mathbf{k}$ at $r = 5$ μm.

*10.2-5 **Propagation Constants and Wavevector (Graded-Index Fiber).** Carry out the same calculations as in Prob. 10.2-4, but for a graded-index fiber with parabolic profile ($p = 2$).

10.3-3 **Scattering Loss.** At a wavelength of $\lambda_o = 820$ nm, the absorption loss of a fiber is 0.25 dB/km and the scattering loss is 2.25 dB/km. If the fiber is instead used at $\lambda_o = 600$ nm, and calorimetric measurements of the heat generated by light absorption reveal a loss of 2 dB/km, estimate the total attenuation at $\lambda_o = 600$ nm.

10.3-4 **Modal Dispersion in Step-Index Fibers.** Determine the core radius of a multimode step-index fiber with a numerical aperture NA $= 0.1$ if the number of modes $M = 5000$ when the wavelength is 0.87 μm. If the core refractive index $n_1 = 1.445$, the group index $N_1 = 1.456$, and Δ are approximately independent of wavelength, determine the modal-dispersion response time σ_τ for a 2-km-long fiber.

10.3-5 **Modal Dispersion in Graded-Index Fibers.** Consider a graded-index fiber with $a/\lambda_o = 10$, $n_1 = 1.45$, $\Delta = 0.01$, and power-law profile with index p. Determine the number of modes M, and the modal-dispersion pulse-broadening rate σ_τ/L, for $p = 1.9$, 2, 2.1, ∞.

10.3-6 **Pulse Propagation.** Consider a pulse of initial temporal width τ_0 transmitted through a graded-index fiber of length L km and power-law refractive-index profile with index p. The peak refractive index n_1 is wavelength-dependent with $D_\lambda = -(\lambda_o/c_o)\, d^2 n_1/d\lambda_o^2$, Δ is approximately independent of wavelength, σ_λ is the spectral width of the source, and λ_o is the operating wavelength. Discuss the effect of increasing each of the following parameters on the temporal width of the received pulse: L, τ_0, p, $|D_\lambda|$, σ_λ, and λ_o.

CHAPTER
11

RESONATOR OPTICS

11.1 PLANAR-MIRROR RESONATORS 436
- A. Resonator Modes
- B. Off-Axis Resonator Modes

11.2 SPHERICAL-MIRROR RESONATORS 447
- A. Ray Confinement
- B. Gaussian Modes
- C. Resonance Frequencies
- D. Hermite–Gaussian Modes
- *E. Finite Apertures and Diffraction Loss

11.3 TWO- AND THREE-DIMENSIONAL RESONATORS 459
- A. Two-Dimensional Rectangular Resonators
- B. Circular Resonators and Whispering-Gallery Modes
- C. Three-Dimensional Rectangular Resonators

11.4 MICRORESONATORS AND NANORESONATORS 463
- A. Rectangular Microresonators
- B. Micropillars, Microdisks, and Microtoroids
- C. Microspheres
- D. Photonic-Crystal Microcavities
- E. Plasmonic Resonators: Metallic Nanodisks and Nanospheres

Charles Fabry
(1867–1945)

Alfred Perot
(1863–1925)

Working together, the French physicists Charles Fabry and Alfred Perot constructed an optical resonator for use as an interferometer. Now known as the Fabry–Perot etalon, it is used extensively in lasers.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

An optical resonator is the optical counterpart of an electronic resonant circuit. It confines and stores light at resonance frequencies determined by its configuration. It is conveniently viewed as an optical transmission system that incorporates feedback: light is repeatedly reflected, or circulates, within its boundaries. Various optical-resonator configurations are depicted in Fig. 11.0-1. The simplest of these, the **Fabry–Perot resonator**, comprises two parallel planar mirrors; light is repeatedly reflected between them while experiencing little loss. Other mirror configurations include spherical mirrors, ring arrangements, and rectangular two- and three-dimensional resonators.

Fiber-ring resonators and **integrated-optic-ring resonators** are widely used. Light can also be trapped in defects within dielectric photonic-bandgap periodic structures, forming **photonic-crystal resonators**. **Guided-wave Fabry–Perot resonators** make use of Fresnel reflection at the boundaries between semiconductors and air. Periodic dielectric structures such as distributed Bragg reflectors (DBRs) can serve in lieu of conventional mirrors in Fabry–Perot resonators, providing feedback in structures such as **micropillar resonators**. **Dielectric resonators** make use of total internal reflection, in place of conventional reflection, at the boundary between low-loss dielectric materials. **Microdisks**, **microtoroids**, and **microspheres** support light that circulates via reflection at near-grazing incidence, in what are known as whispering-gallery modes. The confined rays skim around the inside rim of the resonator with an angle of incidence that is always greater than the critical angle so they do not refract out of the resonator. **Plasmonic resonators** are metallic structures of subwavelength dimensions, such as **nanodisks** and **nanospheres**, that support surface plasmon polariton waves and localized surface plasmon oscillations.

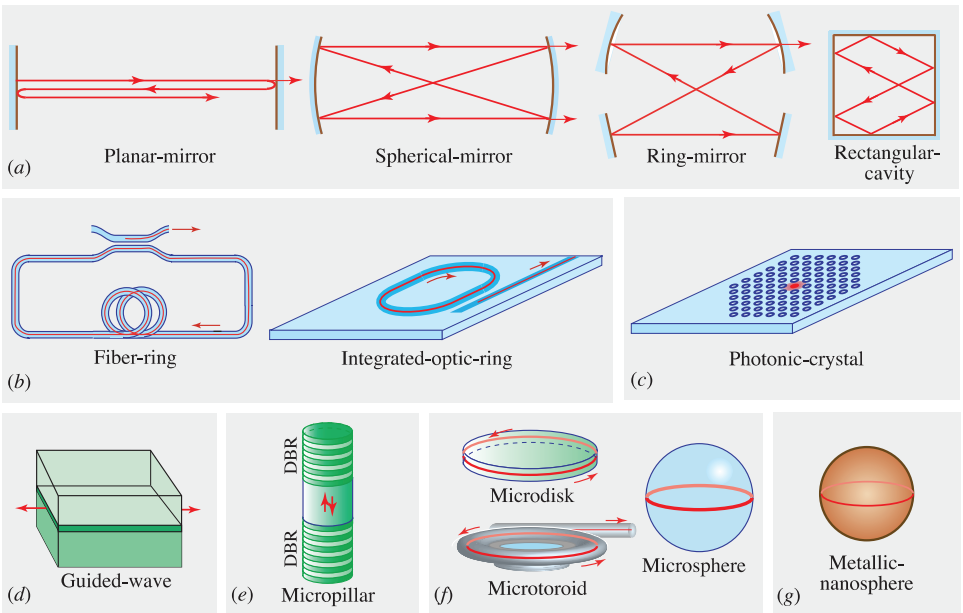


Figure 11.0-1 Storage of light in optical resonators via: (*a*) multiple reflections from mirrors; (*b*) propagation through closed-loop optical fibers and integrated-photonic waveguides; (*c*) trapping of light within defects in photonic crystals; (*d*) multiple Fresnel reflections at semiconductor–air boundaries; (*e*) reflections from periodic structures such as distributed Bragg reflectors (DBRs); (*f*) whispering-gallery mode reflections near the surfaces of dielectric microresonators such as disks, toroids, and spheres; and (*g*) localized surface plasmon oscillations in metallic nanospheres.

The size of an optical resonator can be of the same order of magnitude as its resonance wavelength, as with microresonators. But optical resonators can also be orders of magnitude larger than the resonance wavelength, as with bulk mirror resonators; or orders of magnitude smaller, as with metallic nanospheres. Figure 11.0-2 displays the relationship between resonator size and resonance wavelength for a number of electromagnetic resonators. Optical resonators are frequently characterized by two

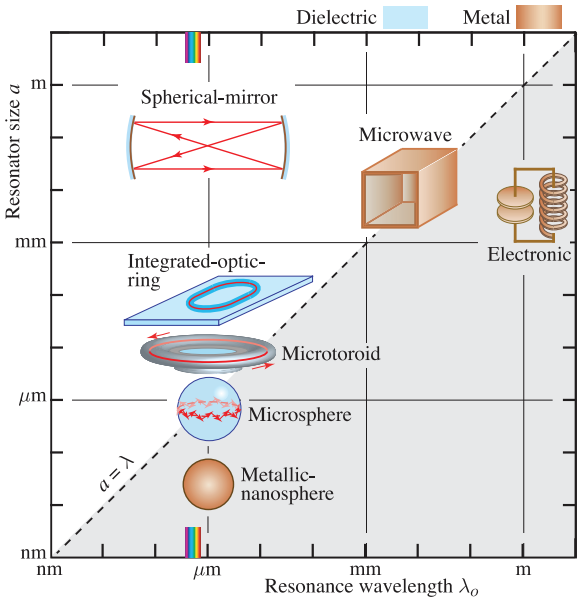


Figure 11.0-2 Resonator size a vs. resonance wavelength λ_o for various dielectric and metallic electromagnetic resonators. The size-to-wavelength ratio belongs to one of three regimes: $a/\lambda_o > 1$ (unshaded region), $a/\lambda_o \approx 1$ (dotted diagonal), and $a/\lambda_o < 1$ (shaded region). Metallic-nanosphere and electronic resonators lie well within the shaded region.

key parameters, representing the degrees of temporal and spatial light confinement, respectively:

1. The quality factor Q, which is proportional to the storage time of the resonator in units of optical period; a large value of Q indicates strong temporal confinement;
2. The modal volume V, which is the volume occupied by the confined optical mode; this measure is of particular interest for microresonators in which a small value of V indicates tight spatial confinement.

Because of their frequency selectivity, optical resonators can also serve as optical filters or spectrum analyzers (Sec. 7.1B). Their most important use, however, is as a "container" within which laser light can be generated and built up. The laser medium (Sec. 16.1A) amplifies light inside the optical resonator while the resonator determines, in part, the frequency and spatial distribution of the laser beam that is generated. Because resonators have the capacity to store energy, they can also be used to produce pulses of laser energy (Sec. 16.4A). Lasers are discussed in Chapters 16 and 18, and the material presented in this chapter is essential to their understanding.

This Chapter

The theoretical approaches considered in previous chapters are useful for describing the operation of optical resonators:

1. The simplest approach is based on *Ray Optics* (Chapter 1); optical rays are traced as they repeatedly reflect within the resonator and geometrical conditions are established that assure that the rays are confined.

2. *Wave Optics* (Chapter 2) is used to determine the modes of a resonator, i.e., the resonance frequencies and wavefunctions of the optical waves that are permitted to exist self-consistently within the resonator.
3. The study of *Beam Optics* (Chapter 3) is useful for understanding the behavior of spherical-mirror resonators; the modes of a resonator with spherical mirrors give rise to Gaussian and Laguerre–Gaussian optical beams.
4. *Fourier Optics* and the theory of light propagation and diffraction (Chapter 4) determine how the finite sizes of resonator mirrors affect resonator loss and the spatial characteristics of the ensuing modes.
5. *Electromagnetic Optics* (Chapter 5) provides a treatment of scattering that is fundamental to understanding the resonant feedback in coherent random lasers.
6. *Polarization Optics* (Chapter 6) sets forth the Fresnel relations, which establish the reflectances at the boundaries between semiconductors and air in guided-wave Fabry–Perot resonators.
7. *Photonic-Crystal Optics*, including the optics of multilayer media (Chapter 7), is important for describing optical resonators that make use of multiple dielectric layers and periodic media (e.g., distributed Bragg reflectors and photonic crystals) in lieu of conventional mirrors.
8. The analysis of oscillating electric charges (localized surface plasmons) in metals, considered in the *Optics of Metals and Metamaterials* (Chapter 8), is crucial for understanding the behavior of nanoresonators such as the metallic nanosphere.
9. The methods used in *Guided-Wave Optics* (Chapter 9) for determining the modes of planar-mirror and planar dielectric waveguides are similar to those required for the analysis of resonator modes.
10. *Fiber Optics* (Chapter 10) provides an analysis of the modes supported by optical fibers; the fiber laser resonator is an optical fiber bounded by end reflectors that result in the repeated reflection and confinement of the propagating light.

The optical resonator evidently provides a comprehensive venue for applying the theories of light discussed in earlier chapters. We begin with a study of planar-mirror resonators in Sec. 11.1 and spherical-mirror resonators in Sec. 11.2. We then introduce two- and three-dimensional resonators in Sec. 11.3 and finally consider microresonators and nanoresonators in Sec. 11.4.

11.1 PLANAR-MIRROR RESONATORS

A. Resonator Modes

In this section we examine the modes of an optical resonator constructed from two parallel, highly reflective, flat mirrors separated by a distance d (Fig. 11.1-1). This simple one-dimensional resonator is known as a **Fabry–Perot resonator**. We first consider an idealized version of this resonator in which the mirrors are lossless; the effect of losses is included subsequently.

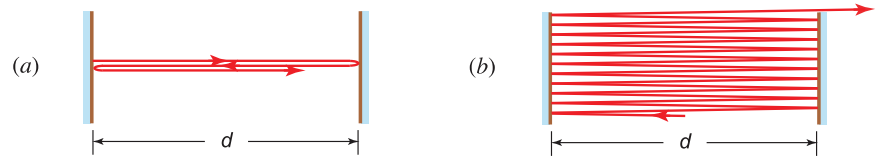


Figure 11.1-1 Two-mirror planar resonator (Fabry–Perot resonator). (*a*) Light rays perpendicular to the mirrors reflect back and forth without escaping. (*b*) Rays that are only slightly inclined eventually escape. Rays also escape if the mirrors are not perfectly parallel.

Resonator Modes as Standing Waves

As discussed in Secs. 2.2, 5.3, and 5.4, a monochromatic wave of frequency ν has a wave function

$$u(\mathbf{r}, t) = \text{Re}\left\{U(\mathbf{r}) \exp(j2\pi\nu t)\right\}, \tag{11.1-1}$$

representing a transverse component of the electric field. The complex amplitude $U(\mathbf{r})$ satisfies the Helmholtz equation, $\nabla^2 U(\mathbf{r}) + k^2 U(\mathbf{r}) = 0$, where $k = 2\pi\nu/c$ is the wavenumber and c is the speed of light in the medium. The resonator modes are the solutions to the Helmholtz equation under the appropriate boundary conditions. For the lossless planar-mirror resonator, the transverse components of the electric field vanish at the mirror surfaces (see Sec. 5.1), so that $U(\mathbf{r}) = 0$ at the planes $z = 0$ and $z = d$ in Fig. 11.1-2.

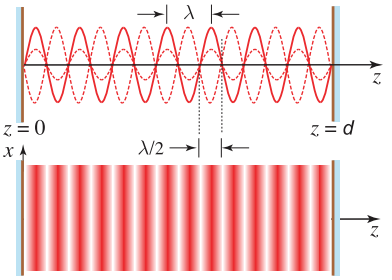


Figure 11.1-2 (a) Wave function $u(\mathbf{r}, t)$ for an ideal planar-mirror mode as a function of z (for $x = y = 0$), portrayed at several different times. In this illustration, 14 half-wavelengths match the length of the resonator so that the mode number $q = d/(\lambda/2) = 14$. (b) Spatial distribution of the magnitude $|u(\mathbf{r}, t)|$ as a function of x and z (for $y = 0$) at a particular time, represented on a color scale where red represents a large magnitude and white represents zero.

The standing wave $U(\mathbf{r}) = A \sin kz$, where A is a constant, satisfies the Helmholtz equation and vanishes at $z = 0$ and $z = d$ if k satisfies the condition $kd = q\pi$, where q is an integer. This restricts k to the values

$$k_q = q\frac{\pi}{d}, \qquad q = 1, 2, \ldots, \tag{11.1-2}$$

so that the modes have complex amplitudes

$$U(\mathbf{r}) = A_q \sin k_q z, \tag{11.1-3}$$

where the A_q are constants. Negative values of q do not constitute independent modes since $\sin k_{-q} z = -\sin k_q z$. Furthermore, the value $q = 0$ is associated with a mode that carries no energy since $k_0 = 0$ and $\sin k_0 z = 0$. The modes of the resonator are therefore the standing waves $A_q \sin k_q z$, where the positive integer $q = 1, 2, \ldots$ is called the **mode number**. An arbitrary wave inside the resonator can be written in terms of a superposition of the resonator modes:

$$U(\mathbf{r}) = \sum_q A_q \sin k_q z. \tag{11.1-4}$$

It follows from (11.1-2) that the associated frequencies $\nu = ck/2\pi$ are restricted to the discrete values

$$\nu_q = q\frac{c}{2d}, \qquad q = 1, 2, \ldots, \tag{11.1-5}$$

which are the resonance frequencies. As illustrated in Fig. 11.1-3, adjacent resonance frequencies are separated by a constant frequency difference, known as the **free spectral range**:

$$\nu_F = \frac{c}{2d} = \frac{c_o}{2nd}. \tag{11.1-6}$$

Frequency Spacing of Resonator Modes

The associated resonance wavelengths are $\lambda_q = c/\nu_q = 2d/q$. The round-trip distance traversed at resonance must therefore precisely equal an integer number of wavelengths:

$$2d = q\lambda_q, \qquad q = 1, 2, \ldots. \tag{11.1-7}$$

It is important to keep in mind that $c = c_o/n$ is the speed of light in the medium between the two mirrors, and that the λ_q represent wavelengths within that medium.

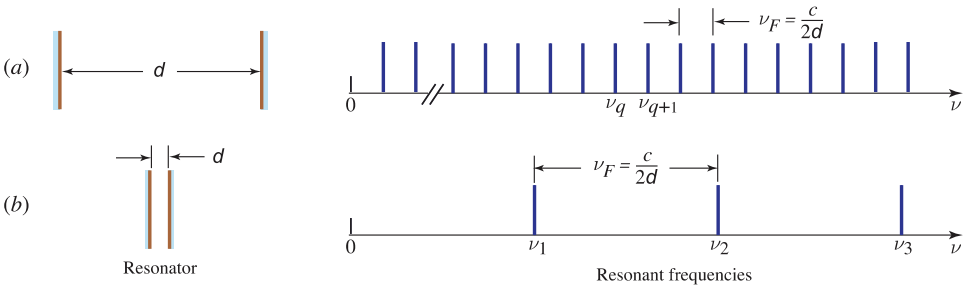


Figure 11.1-3 The adjacent resonance frequencies of a planar-mirror resonator are separated by $\nu_F = c/2d = c_o/2nd$, as illustrated by two examples: (a) A 30-cm long resonator ($d = 30$ cm) with air between the mirrors ($n = 1$) has a frequency spacing between modes given by $\nu_F = 500$ MHz. (b) A much shorter resonator with $d = 3$ μm has $\nu_F = 50$ THz, so that the first mode has a frequency corresponding to a wavelength of 6 μm and there are only two modes within the 700–900-nm optical band, which occupies a frequency range of 95 THz.

Resonator Modes as Traveling Waves

Alternatively, the resonator modes can be determined by following a wave as it travels back and forth between the two mirrors [Fig. 11.1-4(a)]. A mode is a wave that reproduces itself after a single round trip (see Appendix C). The phase shift imparted by a single round trip of propagation (a distance $2d$), $\varphi = k2d = 4\pi\nu d/c$, must therefore be a multiple of 2π:

$$\varphi = k2d = q2\pi, \qquad q = 1, 2, \ldots. \tag{11.1-8}$$

This result is not altered by an additional phase shift of 2π, which can be imparted by reflections at the two mirrors (see Sec. 6.2). As expected, we therefore obtain $kd = q\pi$, as in (11.1-2), and the same resonance frequencies as set forth in (11.1-5). Equation (11.1-8) may be viewed as a condition of positive feedback in the system displayed in Fig. 11.1-4(b); this requires that the output of the system be fed back *in phase* with the input.

We now demonstrate that only self-reproducing waves, or combinations thereof, can exist within the resonator under steady-state conditions. Consider a monochromatic

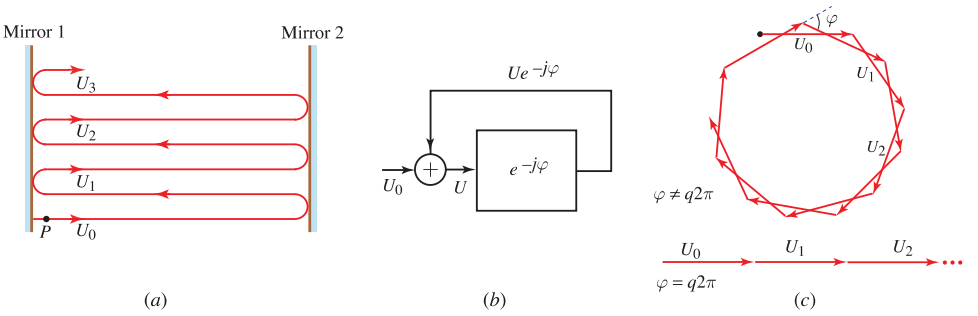


Figure 11.1-4 (a) A wave reflects back and forth between the resonator mirrors, suffering a phase shift φ on each round trip. (b) Block diagram of an optical feedback system with a phase delay φ. (c) Phasor diagram representing the sum $U = U_0 + U_1 + \cdots$ for $\varphi \neq q2\pi$ and for $\varphi = q2\pi$.

plane wave of complex amplitude U_0 at point P traveling to the right along the axis of the resonator [see Fig. 11.1-4(a)]. The wave is reflected from mirror 2 and propagates back to mirror 1 where it is again reflected. Its amplitude then becomes U_1. Yet another round trip results in a wave of complex amplitude U_2, and so on *ad infinitum*. Because the original wave U_0 is monochromatic, it is "eternal." Indeed, all of the partial waves, $U_0, U_1, U_2, \ldots$ are monochromatic and perpetually coexist. Moreover, their magnitudes are identical because it has been assumed that there is no loss associated with reflection and propagation. The total wave U is therefore represented by the sum of an infinite number of phasors of equal magnitude,

$$U = U_0 + U_1 + U_2 + \cdots, \tag{11.1-9}$$

as shown in Figs. 11.1-4(b) and (c).

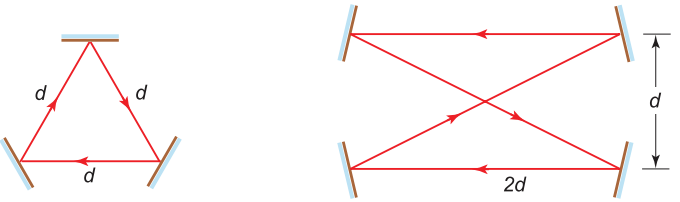
The phase difference of two consecutive phasors imparted by a single round trip of propagation is $\varphi = k2d$. If the magnitude of the initial phasor is infinitesimal, the magnitude of each of these phasors must also be infinitesimal. The magnitude of the sum of this infinite number of infinitesimal phasors is itself infinitesimal unless they are aligned, i.e., unless $\varphi = q2\pi$, as illustrated at the bottom of Fig. 11.1-4(c). Thus, an infinitesimal initial wave can result in the buildup of finite power in the resonator, but only if $\varphi = q2\pi$.

Traveling-Wave Resonators

In a traveling-wave resonator, an optical mode travels in one direction along a closed path representing a round trip and retraces itself without reversing direction. Examples are the **ring resonator** and the **bow-tie resonator** illustrated in Fig. 11.1-5. The resonance frequencies of the modes may be obtained by equating the round-trip phase shift to 2π. Each of the set of modes traveling in the clockwise direction has a corresponding mode of the same resonance frequency traveling in the counterclockwise direction, and the matching modes are said to be degenerate.

EXERCISE 11.1-1

Resonance Frequencies of a Traveling-Wave Resonator. Derive expressions for the resonance frequencies ν_q and their frequency spacing ν_F for the three-mirror ring and the four-mirror bow-tie resonator shown in Fig. 11.1-5. Assume that each mirror reflection introduces a phase shift of π.



(*a*) Ring resonator (*b*) Bow-tie resonator

Figure 11.1-5 Traveling-wave resonators. (*a*) Three-mirror ring resonator. (*b*) Four-mirror bow-tie resonator.

Density of Modes

The number of modes per unit frequency is the inverse of the frequency spacing between modes, i.e., $1/\nu_F = 2d/c$ in each of the two orthogonal polarizations. The density of modes $M(\nu)$, which is the number of modes per unit frequency per unit length of the resonator, is therefore

$$M(\nu) = \frac{4}{c}. \tag{11.1-10}$$

Density of Modes (1D Resonator)

The number of modes in a resonator of length d, in the frequency interval $\Delta\nu$, is thus $(4/c)d\Delta\nu$. This represents the number of degrees of freedom for the optical waves existing in the resonator, i.e., the number of independent ways in which these waves may be arranged.

Losses and Resonance Spectral Width

The strict condition on the frequencies of optical waves that are permitted to exist inside a resonator is relaxed when the resonator has losses. Consider again Fig. 11.1-4(*a*) and follow the initial wave inside the resonator, U_0, in its excursions between the two mirrors. As discussed above, the result is the infinite sum of phasors shown in Fig. 11.1-4(*c*) and the phase difference imparted by propagation through a single round trip is

$$\varphi = 2kd = 4\pi\nu d/c. \tag{11.1-11}$$

Reflection at the two mirrors can impart an additional phase shift, usually 2π.

However, in the presence of loss the phasors are *not* all of equal magnitude. Two successive phasors are related by a complex round-trip amplitude attenuation factor $h = |\mathrm{r}|e^{-j\varphi}$ resulting from losses associated with the *two* mirror reflections and the absorption in the medium (the corresponding intensity attenuation factor for a round trip is $|\mathrm{r}|^2$ with $|\mathrm{r}| < 1$). Thus, $U_1 = hU_0$ and, in fact, U_2 is related to U_1 by this same complex factor h, as are all consecutive phasor pairs. The net result is the superposition of an infinite number of waves, each distinguished from the previous one by a constant phase shift and an amplitude that is geometrically reduced. It is readily seen that $U = U_0 + U_1 + U_2 + \cdots = U_0 + hU_0 + h^2U_0 + \cdots = U_0(1 + h + h^2 + \cdots) = U_0/(1 - h)$. The net result, $U = U_0/(1 - h)$, is easily understood in terms of the simple feedback configuration pictured in Fig. 11.1-4(*b*).

The intensity of the light in the resonator is therefore given by

$$I = |U|^2 = \frac{|U_0|^2}{|1 - |\mathrm{r}|e^{-j\varphi}|^2} = \frac{I_0}{1 + |\mathrm{r}|^2 - 2|\mathrm{r}|\cos\varphi}, \tag{11.1-12}$$

which can be written as

$$I = \frac{I_{\max}}{1 + (2\mathcal{F}/\pi)^2 \sin^2(\varphi/2)}, \qquad I_{\max} = \frac{I_0}{(1 - |\mathrm{r}|)^2}. \tag{11.1-13}$$

Here, $I_0 = |U_0|^2$ is the intensity of the initial wave, and the **finesse** of the resonator is

$$\boxed{\mathcal{F} = \frac{\pi\sqrt{|\mathrm{r}|}}{1 - |\mathrm{r}|}.} \tag{11.1-14}$$

Finesse

Again, $|\mathrm{r}|$ is the magnitude of the *round-trip* attenuation factor.

The treatment offered above is nearly identical to that provided earlier in Sec. 2.5B, where the complex round-trip amplitude attenuation factor was chosen to be $h = |h|e^{+j\varphi}$. In the current context we instead select this factor to be $h = |\mathrm{r}|e^{-jk2d} = |\mathrm{r}|e^{-j\varphi}$ by dint of the fact that successive phasors arise from the *delay* of the wave as it bounces between the mirrors. This distinction is superficial, however, and has no bearing on the results.

Indeed, (11.1-13) is identical to (2.5-18), which is plotted in Fig. 2.5-10(*b*). The intensity $I(\varphi)$ is a periodic function of φ with period 2π. For large $\mathcal{F}$, $I(\varphi)$ has sharp peaks centered about the values $\varphi = q2\pi$, which correspond to the alignment of all phasors. The peaks have a full-width at half-maximum (FWHM) described by $\Delta\varphi \approx 2\pi/\mathcal{F}$, in accordance with (2.5-21).

The internal resonator intensity $I(\varphi)$ in (11.1-13) can alternately be expressed as a function of the optical frequency of an internal monochromatic wave, $I(\nu)$, by virtue of (11.1-11), which shows that $\varphi = 4\pi\nu d/c$. This function then takes the form

$$\boxed{I = \frac{I_{\max}}{1 + (2\mathcal{F}/\pi)^2 \sin^2(\pi\nu/\nu_F)},} \qquad I_{\max} = \frac{I_0}{(1 - |\mathrm{r}|)^2}, \tag{11.1-15}$$

with $\nu_F = c/2d$. This result is displayed in Fig. 11.1-6 and indeed it mirrors that depicted in Figs. 2.5-10 and 7.1-5. The maximum internal intensity $I = I_{\max}$ is attained when the second term in the denominator is zero, i.e., at the resonance frequencies

$$\nu = \nu_q = q\nu_F, \qquad q = 1, 2, \ldots. \tag{11.1-16}$$

The minimum intensity, attained at the midpoints between the resonances, is

$$I_{\min} = \frac{I_{\max}}{1 + (2\mathcal{F}/\pi)^2}. \tag{11.1-17}$$

When the finesse is large ($\mathcal{F} \gg 1$), it is clear that the spectral response of the resonator is sharply peaked about the resonance frequencies and $I_{\min}/I_{\max}$ is small. In

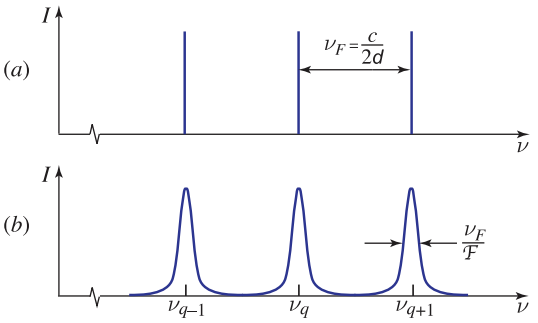


Figure 11.1-6 (a) In the steady state, a lossless resonator ($\mathcal{F} = \infty$) sustains light waves only at the precise resonance frequencies ν_q. (b) A lossy resonator best sustains waves in the immediate vicinity of the resonance frequencies, but it can sustain waves at other frequencies as well.

that case, the FWHM of the resonance peaks is $\delta\nu \approx \nu_F/\mathcal{F}$ since $\delta\nu = (c/4\pi d)\Delta\varphi$ and $\Delta\varphi \approx 2\pi/\mathcal{F}$ in accordance with (2.5-21). This simple result provides the rationale for the definition of the finesse given in (11.1-14).

In short, the spectral response of the Fabry–Perot optical resonator is characterized by two parameters:

- The frequency spacing ν_F between adjacent resonator modes:

$$\boxed{\nu_F = \frac{c}{2d}.} \qquad (11.1\text{-}18)$$

Frequency Spacing

- The spectral width $\delta\nu$ of the individual resonator modes:

$$\boxed{\delta\nu \approx \frac{\nu_F}{\mathcal{F}}.} \qquad (11.1\text{-}19)$$

Spectral Width

Equation (11.1-19) is valid in the usual case when $\mathcal{F} \gg 1$. The spectral width $\delta\nu$ is inversely proportional to the finesse $\mathcal{F}$. As the loss increases, $\mathcal{F}$ decreases and $\delta\nu$ therefore increases.

Sources of Resonator Loss

The two principal sources of loss in optical resonators are:

- Losses arising from imperfect reflection at the mirrors. There are two underlying sources of reduced reflection: (1) a partially transmitting mirror is often deliberately used in a resonator to permit laser light generated in the resonator to escape through it; and (2) the finite size of the mirrors causes a fraction of the light to leak around them and thereby to be lost. This latter effect also modifies the spatial distribution of the reflected wave by truncating it to the size of the mirror. The reflected light produces a diffraction pattern at the opposite mirror that is again truncated. Such diffraction loss may be regarded as an effective reduction of the mirror reflectance. Further details regarding diffraction loss are provided in Sec. 11.2E.
- Losses attributable to absorption and scattering that occurs in the medium between the mirrors. The round-trip power attenuation factor associated with these effects is $\exp(-2\alpha_s d)$, where α_s is the loss coefficient of the medium associated with absorption and scattering.

For mirrors of reflectances $\mathcal{R}_1 = |\mathrm{r}_1|^2$ and $\mathcal{R}_2 = |\mathrm{r}_2|^2$, the wave intensity decreases by the factor $\mathcal{R}_1\mathcal{R}_2$ as a result of the two reflections associated with a single round trip. These are referred to as "lumped losses" since they occur only at the discrete locations where the mirrors are located. Accounting also for the "distributed losses" that take place within the intervening medium yields a round-trip intensity attenuation factor

$$|\mathrm{r}|^2 = \mathcal{R}_1\mathcal{R}_2 \exp(-2\alpha_s d), \tag{11.1-20}$$

which is usually written in the form

$$|\mathrm{r}|^2 = \exp(-2\alpha_r d), \tag{11.1-21}$$

where α_r is an effective overall distributed-loss coefficient. Equating (11.1-20) and (11.1-21), and taking the natural logarithm of both sides, allows α_r to be written in terms of the distributed and lumped loss parameters, α_s and $\mathcal{R}_1\mathcal{R}_2$, respectively:

$$\boxed{\alpha_r = \alpha_s + \frac{1}{2d} \ln \frac{1}{\mathcal{R}_1\mathcal{R}_2}.} \tag{11.1-22}$$

Loss Coefficient

This can also be written as

$$\alpha_r = \alpha_s + \alpha_{m1} + \alpha_{m2}, \tag{11.1-23}$$

where the quantities

$$\alpha_{m1} = \frac{1}{2d} \ln \frac{1}{\mathcal{R}_1}, \qquad \alpha_{m2} = \frac{1}{2d} \ln \frac{1}{\mathcal{R}_2} \tag{11.1-24}$$

represent the effective distributed-loss coefficients associated with mirrors 1 and 2, respectively.

These loss coefficients can be cast in a simpler form for mirrors of high reflectance. If $\mathcal{R}_1 \approx 1$, then $\ln(1/\mathcal{R}_1) = -\ln(\mathcal{R}_1) = -\ln[1-(1-\mathcal{R}_1)] \approx 1-\mathcal{R}_1$, where we have used the Taylor-series approximation $\ln(1-\Delta) \approx -\Delta$, which is valid for $|\Delta| \ll 1$. This allows us to write

$$\alpha_{m1} \approx \frac{1-\mathcal{R}_1}{2d}. \tag{11.1-25}$$

Similarly, if $\mathcal{R}_2 \approx 1$, we have $\alpha_{m2} \approx (1-\mathcal{R}_2)/2d$. If, furthermore, $\mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R} \approx 1$, then

$$\alpha_r \approx \alpha_s + \frac{1-\mathcal{R}}{d}. \tag{11.1-26}$$

The finesse $\mathcal{F}$ can be expressed as a function of the effective loss coefficient α_r by substituting (11.1-21) in (11.1-14). The result is

$$\mathcal{F} = \frac{\pi \exp(-\alpha_r d/2)}{1-\exp(-\alpha_r d)}, \tag{11.1-27}$$

which is plotted in Fig. 11.1-7. It is clear that the finesse decreases as the loss increases. If the loss factor $\alpha_r d \ll 1$, then $\exp(-\alpha_r d) \approx 1 - \alpha_r d$, whereupon

$$\mathcal{F} \approx \frac{\pi}{\alpha_r d}. \tag{11.1-28}$$

Finesse and Loss Factor

The finesse is thus inversely proportional to the loss factor $\alpha_r d$ in this limit.

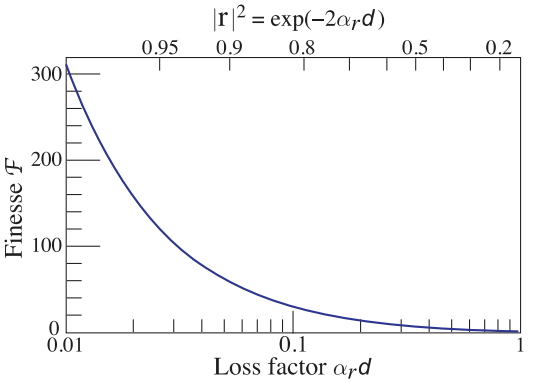


Figure 11.1-7 Finesse of an optical resonator versus the loss factor $\alpha_r d$, where α_r is the effective overall distributed-loss coefficient. The round-trip intensity attenuation factor $|r|^2 = \exp(-2\alpha_r d)$ is shown on the upper abscissa.

EXERCISE 11.1-2

Resonator Modes and Spectral Width. Determine the frequency spacing, and spectral width, of the modes of a Fabry–Perot resonator whose mirrors have power reflectances of 0.98 and 0.99 and are separated by a distance $d = 100$ cm. Assume that the medium has refractive index $n = 1$ and negligible losses. Is the approximation used to derive (11.1-28) appropriate in this case?

Photon Lifetime

The relationship between the resonance linewidth and resonator loss may be viewed as a manifestation of the time–frequency uncertainty relation, as we now demonstrate. Substituting (11.1-18) and (11.1-28) in (11.1-19), we obtain

$$\delta\nu \approx \frac{c/2d}{\pi/\alpha_r d} = \frac{c\alpha_r}{2\pi}. \tag{11.1-29}$$

Because α_r is the loss per unit length, $c\alpha_r$ represents the loss per unit time. Defining the characteristic decay time

$$\tau_p = \frac{1}{c\alpha_r} \tag{11.1-30}$$

as the **resonator lifetime** or **photon lifetime**, we obtain

$$\delta\nu \approx \frac{1}{2\pi\tau_p}. \tag{11.1-31}$$

The time–frequency uncertainty product is thus $\delta\nu \cdot \tau_p = 1/2\pi$. Resonance-line broadening may therefore be considered to be a consequence of optical-energy decay arising from resonator losses. An electric field that decays as $\exp(-t/2\tau_p)$, corresponding to an energy that decays as $\exp(-t/\tau_p)$, has a Fourier transform that is proportional to $1/(1 + j4\pi\nu\tau_p)$, which has a (FWHM) spectral width $\delta\nu = 1/2\pi\tau_p$ (see Sec. 14.3D).

Quality Factor Q

The **quality factor** Q is often used to characterize electrical resonance circuits and microwave resonators. This parameter is defined as

$$Q = 2\pi \frac{\text{stored energy}}{\text{energy loss per cycle}}. \tag{11.1-32}$$

Large values of Q are associated with low-loss resonators. A series RLC circuit has resonance frequency $\nu_0 \approx 1/2\pi\sqrt{LC}$ and quality factor $Q = 2\pi\nu_0 L/R$, where R, L, and C are the resistance, inductance, and capacitance of the resonance circuit, respectively.

The quality factor of an optical resonator is determined by observing that the stored energy E is lost at the rate $c\alpha_r E$ (per unit time), which is equivalent to the rate $c\alpha_r E/\nu_0$ (per cycle of the optical field), so that

$$Q = \frac{2\pi\nu_0}{c\alpha_r}. \tag{11.1-33}$$

Since $\delta\nu \approx c\alpha_r/2\pi$ in accordance with (11.1-29), we have

$$Q \approx \frac{\nu_0}{\delta\nu}. \tag{11.1-34}$$

By virtue of (11.1-33), the quality factor is related to the resonator lifetime (photon lifetime) $\tau_p = 1/c\alpha_r$ via

$$Q = 2\pi\nu_0\tau_p. \tag{11.1-35}$$

The quality factor Q is thus understood to be the storage time of the resonator in units of the optical period $T = 1/\nu_0$.

Finally, combining (11.1-19) and (11.1-34) leads to a relationship between Q and the finesse $\mathcal{F}$ of the resonator:

$$Q \approx \frac{\nu_0}{\nu_F}\mathcal{F}. \tag{11.1-36}$$

Since optical resonator frequencies ν_0 are typically much greater than the mode spacing ν_F, we have $Q \gg \mathcal{F}$. Moreover, the quality factor of an optical resonator is typically far greater than that of a resonator at microwave frequencies.

Summary

- Two parameters are convenient for characterizing the losses in an optical resonator: the loss coefficient α_r (cm^{-1}) and the photon lifetime $\tau_p = 1/c\alpha_r$ (s).
- Two dimensionless parameters characterize the quality of an optical resonator of length d operated at frequency ν_0: the finesse $\mathcal{F} \approx \pi/\alpha_r d$ and the quality factor $Q = 2\pi\nu_0\tau_p$.
- Two frequencies describe the spectral characteristics of an optical resonator: the frequency spacing between the modes $\nu_F = c/2d$, known as the free spectral range, and the spectral width $\delta\nu \approx \nu_F/\mathcal{F}$.

B. Off-Axis Resonator Modes

An optical resonator with perfectly parallel planar mirrors of infinite dimensions can also support oblique, or off-axis, modes. A plane wave traveling at an angle θ with respect to the axis of the resonator (the z direction) bounces back and forth between lossless mirrors [see Fig. 11.1-8(a)] as a guided wave traveling in the transverse direction (the x direction). Such guided waves were described in Sec. 9.1.

The boundary conditions at the mirrors dictate that the axial component of the propagation constant, $k_z = k\cos\theta$, is an integer multiple of π/d. However, no such condition is imposed on the transverse component k_x since the resonator is open in the x direction. Since $k = 2\pi\nu/c$, the condition $k\cos\theta = q\pi/d$, where q is an integer, can be written in the form

$$\nu = q\,\nu_F \sec\theta\,, \qquad q = 1, 2, \ldots, \tag{11.1-37}$$

where $\nu_F = c/2d$. This relation, which is plotted in Fig 11.1-8(b), is equivalent to the self-consistency condition for guided modes in planar-mirror waveguides (see Sec. 9.1). It is also identical to the condition (7.1-37) for the peak transmittance of an oblique wave through a Fabry–Perot etalon. As illustrated in Fig 11.1-8(c), at a given frequency ν, there are modes at a discrete set of angles θ_q that satisfy the condition $\cos\theta_q = q\nu_F/\nu$. These are the complements of the bounce angles of the guided modes of a waveguide. Also, at any fixed angle θ, the modal frequencies are $\nu_q = q\nu_F \sec\theta$, as illustrated in Fig 11.1-8(d). The larger the inclination angle, the greater the spacing between the modal frequencies.

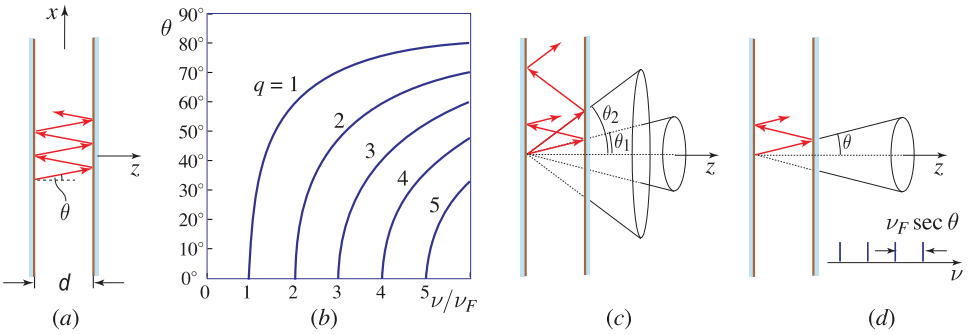


Figure 11.1-8 (a) Off-axis mode in a planar-mirror resonator. (b) Relation between mode angles and resonance frequencies. (c) Off-axis modes at a fixed frequency $\nu > \nu_F$. (d) Resonance frequencies of an off-axis mode at a prescribed angle θ.

11.2 SPHERICAL-MIRROR RESONATORS

The planar-mirror resonator configuration discussed in the preceding section is highly sensitive to misalignment. If the mirrors are not perfectly parallel, or the rays are not perfectly normal to the mirror surfaces, they undergo a sequence of lateral displacements that eventually causes them to wander out of the resonator [see Fig. 11.1-1(b)]. Spherical-mirror resonators, in contrast, provide a more stable configuration for the confinement of light that renders them less sensitive to misalignment under appropriate geometrical conditions.

A spherical-mirror resonator is constructed from two spherical mirrors of radii R_1 and R_2, separated by a distance d (Fig. 11.2-1). A line connecting the centers of the mirrors defines the optical axis (z axis), about which the system exhibits circular symmetry. Each of the mirrors can be concave ($R < 0$) or convex ($R > 0$). The planar-mirror resonator is a special case for which $R_1 = R_2 = \infty$. We first make use of the results set forth in Sec. 1.4D and examine the conditions required for ray confinement. Then, using the results derived in Chapter 3, we determine the resonator modes and resonance frequencies. Finally, we briefly discuss the implications of finite mirror size.

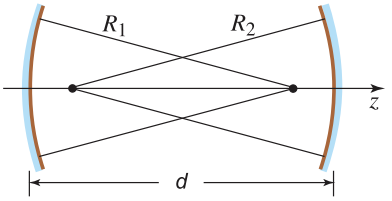


Figure 11.2-1 Geometry of a spherical-mirror resonator. In this illustration both mirrors are concave (their radii of curvature are negative).

A. Ray Confinement

We begin with ray optics to determine the conditions of confinement for light rays in a spherical-mirror resonator. We consider only meridional rays (rays lying in a plane that passes through the optical axis) and limit our consideration to paraxial rays (rays that make small angles with the optical axis). The matrix-optics methods introduced in Sec. 1.4, which are valid only for meridional and paraxial rays in a circularly symmetric system, are thus suitable for studying the trajectories of these rays as they travel inside the resonator.

A resonator is a periodic optical system, since a ray travels through the same system after a round trip of two reflections. We may therefore make use of the analysis of periodic optical systems presented in Sec. 1.4D. Let y_m and θ_m be the position and inclination of an optical ray after m round trips, as illustrated in Fig. 11.2-2. Given y_m and θ_m, we determine y_{m+1} and θ_{m+1} by tracing the ray through the system.

For paraxial rays, where all angles are small, the relation between (y_{m+1}, θ_{m+1}) and (y_m, θ_m) is linear and can be written in matrix form as [see (1.4-3)]

$$\begin{bmatrix} y_{m+1} \\ \theta_{m+1} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_m \\ \theta_m \end{bmatrix}. \tag{11.2-1}$$

Beginning at the left of Fig. 11.2-2 with y_0 and θ_0, the round-trip ray-transfer matrix for the ray pattern shown is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{2}{R_1} & 1 \end{bmatrix} \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{2}{R_2} & 1 \end{bmatrix} \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}. \tag{11.2-2}$$

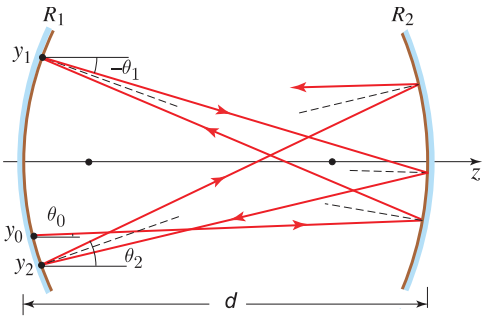


Figure 11.2-2 The position and inclination of a ray after m round trips are represented by y_m and θ_m, respectively, where $m = 0, 1, 2, \ldots$. In this diagram, $\theta_1 < 0$ since the ray is directed downward. Angles are exaggerated for the purposes of illustration; all rays are paraxial so that $\sin\theta \approx \tan\theta \approx \theta$ and the propagation distance of all rays between the mirrors is $\approx d$.

This cascade of ray-transfer matrices represents, from right to left [see (1.4-4) and (1.4-9)]:

- Propagation a distance d through free space
- Reflection from a mirror of radius R_2
- Propagation a distance d through free space
- Reflection from a mirror of radius R_1

As shown in Sec. 1.4D, the solution of the difference equation (11.2-1) is $y_m = y_{\max} F^m \sin(m\varphi + \varphi_0)$, where $F^2 = AD - BC$, $\varphi = \cos^{-1}(b/F)$, $b = (A + D)/2$, and $y_{\max}$ and φ_0 are constants determined from the initial position and inclination of the ray. For the case at hand $F = 1$, so that

$$y_m = y_{\max} \sin(m\varphi + \varphi_0), \tag{11.2-3}$$

$$\varphi = \cos^{-1} b, \qquad b = 2\left(1 + \frac{d}{R_1}\right)\left(1 + \frac{d}{R_2}\right) - 1. \tag{11.2-4}$$

The solution (11.2-3) is harmonic, and therefore bounded, provided $\varphi = \cos^{-1} b$ is real. This is ensured if $|b| \le 1$, i.e., if $-1 \le b \le 1$, so that

$$0 \le \left(1 + \frac{d}{R_1}\right)\left(1 + \frac{d}{R_2}\right) \le 1. \tag{11.2-5}$$

It is convenient to write this **confinement condition** in terms of the quantities $g_1 = 1 + d/R_1$ and $g_2 = 1 + d/R_2$, which are known as the g**-parameters**:

$$\boxed{0 \le g_1\, g_2 \le 1.} \tag{11.2-6}$$

Confinement Condition

The resonator is said to be **stable** when this condition is satisfied. This result also emerges from wave optics, as will be demonstrated subsequently [see (11.2-17)].

When the confinement condition (11.2-6) is not satisfied, φ is imaginary so that y_m in (11.2-3) becomes a hyperbolic sine function of m and increases without bound. The resonator is then said to be **unstable**. At the boundary of the confinement condition (when the inequalities are equalities), the resonator is said to be **conditionally stable**.

A useful graphical representation of the confinement condition (Fig. 11.2-3) identifies each combination (g_1, g_2) of the two g-parameters of a resonator as a point in a g_2 versus g_1 diagram. The left inequality in (11.2-6) is equivalent to $\{g_1 \ge 0 \text{ and } g_2 \ge 0;\ \text{or } g_1 \le 0 \text{ and } g_2 \le 0\}$ so that all stable points (g_1, g_2) must lie in the first or third

quadrants. The right inequality in (11.2-6) signifies that stable points (g_1, g_2) must lie in a region bounded by the hyperbola $g_1\, g_2 = 1$. The unshaded area in Fig. 11.2-3 represents the region for which both inequalities are satisfied, indicating that the resonator is stable.

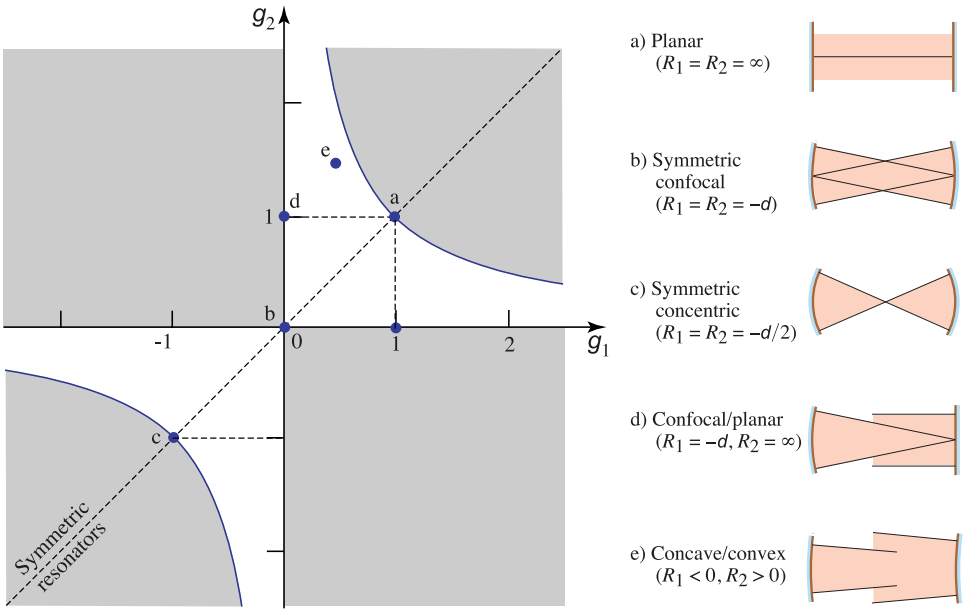


Figure 11.2-3 Resonator stability diagram. A spherical-mirror resonator is stable if the parameters $g_1 = 1 + d/R_1$ and $g_2 = 1 + d/R_2$ lie in the unshaded regions, which are bounded by the lines $g_1 = 0$ and $g_2 = 0$, and the hyperbola $g_2 = 1/g_1$. R is negative for a concave mirror and positive for a convex mirror. Commonly used resonator configurations are indicated by letters and are sketched at the right; shaded areas represent collections of rays perpendicular to the mirrors. All symmetric resonators lie along the line $g_2 = g_1$.

Symmetric resonators, by definition, have identical mirrors $(R_1 = R_2 = R)$ so that $g_1 = g_2 = g$. Resonators in this class are thus represented in Fig. 11.2-3 by points lying along the line $g_2 = g_1$. The condition of stability then becomes $g^2 \leq 1$, or $-1 \leq 1 + d/R \leq 1$, which implies

$$\boxed{0 \leq \frac{d}{(-R)} \leq 2.} \qquad \begin{matrix}(11.2\text{-}7)\\ \text{Confinement Condition}\\ \text{(Symmetric Resonator)}\end{matrix}$$

To satisfy (11.2-7) a stable symmetric resonator must use concave mirrors $(R < 0)$ whose radii are greater than half the resonator length. Three examples within this class are of special interest: $d/(-R) = 0$, 1, and 2, corresponding to **planar**, **confocal**, and **concentric resonators**, respectively.

In the **symmetric confocal resonator**, $(-R) = d$ so that the center of curvature of each mirror lies on the other. Thus, $b = -1$ and $\varphi = \pi$ so that the ray position in (11.2-3) is prescribed to be $y_m = y_{\max} \sin(m\pi + \varphi_0)$, i.e., $y_m = (-1)^m y_0$. Rays initiated at position y_0, at any inclination, are thus imaged to position $y_1 = -y_0$, and then reimaged again to position $y_2 = y_0$, and so on, repeatedly. Each ray thus retraces

itself after two round trips (Fig. 11.2-4). All paraxial rays are therefore confined, whatever their original position and inclination. This is a substantial improvement in comparison with the planar-mirror resonator, for which only rays of zero inclination retrace themselves as schematized in Fig. 11.1-1.

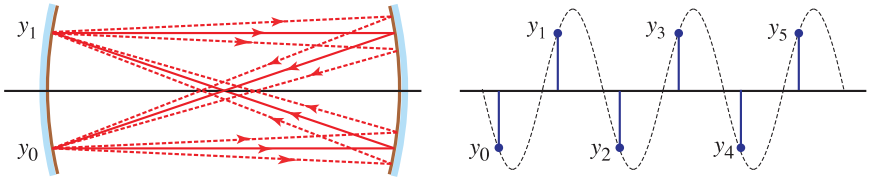


Figure 11.2-4 All paraxial rays in a symmetric confocal resonator retrace themselves after two round trips, whatever their original position and inclination. Angles are exaggerated in this drawing for purposes of illustration.

Summary

The confinement condition for paraxial rays in a spherical-mirror resonator, comprising mirrors of radii R_1 and R_2 separated by a distance d, is $0 \leq g_1 g_2 \leq 1$, where $g_1 = 1 + d/R_1$ and $g_2 = 1 + d/R_2$. The confinement condition for symmetric resonators is $0 \leq d/(-R) \leq 2$; this condition governs planar, symmetric confocal, and symmetric concentric mirror configurations.

EXERCISE 11.2-1

Maximum Resonator Length for Confined Rays. A resonator is constructed using concave mirrors of radii 50 cm and 100 cm. Determine the maximum resonator length for which rays satisfy the confinement condition.

B. Gaussian Modes

Although the ray-optics approach considered in the preceding section is useful for determining the geometrical conditions under which rays are confined, it cannot provide information about the resonance frequencies and spatial intensity distributions of the resonator modes. For those quantities we must appeal to wave optics. We now proceed to show that Gaussian beams are solutions of the paraxial Helmholtz equation for the boundary conditions imposed by a pair of spherical mirrors in a resonator configuration. More generally, we demonstrate that Hermite–Gaussian beams are modes of the spherical-mirror resonator. In the course of our analysis, we obtain expressions for the resonance frequencies and spatial intensity distributions of the resonator modes.

Gaussian Beams

As discussed in Chapter 3, the Gaussian beam is a circularly symmetric wave whose energy is confined about its axis (the z axis) and whose wavefront normals are paraxial rays (Fig. 11.2-5). In accordance with (3.1-12), at an axial distance z from the beam waist, the beam intensity I varies in the transverse x–y plane as the Gaussian distribution $I = I_0[W_0/W(z)]^2 \exp[-2(x^2 + y^2)/W^2(z)]$. Its width is given by (3.1-8):

$$W(z) = W_0 \sqrt{1 + \left(\frac{z}{z_0}\right)^2}, \tag{11.2-8}$$

where z_0 is the distance, known as the Rayleigh range, at which the beam wavefronts are most curved. The beam width (radius) $W(z)$ increases in both directions from its minimum value W_0 at the beam waist ($z = 0$). The radius of curvature of the wavefronts, given by (3.1-9),

$$R(z) = z \left[1 + \left(\frac{z_0}{z}\right)^2\right] \tag{11.2-9}$$

decreases from ∞ at $z = 0$, to a minimum value at $z = z_0$, and thereafter grows linearly with z for large z. For $z > 0$, the wave diverges and $R(z) > 0$; for $z < 0$, the wave converges and $R(z) < 0$. The Rayleigh range z_0 is related to the beam waist radius W_0 by (3.1-11):

$$z_0 = \frac{\pi W_0^2}{\lambda}. \tag{11.2-10}$$

The depth of focus is $2z_0$, i.e., twice the Rayleigh range.

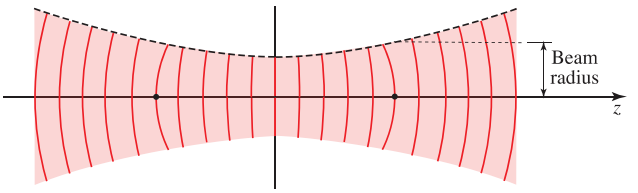


Figure 11.2-5 Gaussian beam wavefronts (solid curves) and beam width (dashed curve).

The Gaussian Beam is a Mode of the Spherical-Mirror Resonator

A Gaussian beam reflected from a spherical mirror will retrace the incident beam if the radius of curvature of its wavefront is the same as that of the mirror radius (see Sec. 3.2C). Hence, if the radii of curvature of the wavefronts of a Gaussian beam, at planes separated by a distance d, match the radii of two mirrors separated by the same distance d, a beam incident on the first mirror will reflect and retrace itself to the second mirror, where it once again will reflect and retrace itself back to the first mirror, and so on. The beam can then exist self-consistently within that spherical-mirror resonator, satisfying the Helmholtz equation and the boundary conditions imposed by the mirrors. Provided that the phase also retraces itself, as discussed in Sec. 11.2C, the Gaussian beam is then said to be a mode of the spherical-mirror resonator.

We now proceed to determine the Gaussian beam that matches a spherical-mirror resonator, whose mirrors have radii of curvature R_1 and R_2 and are separated by the

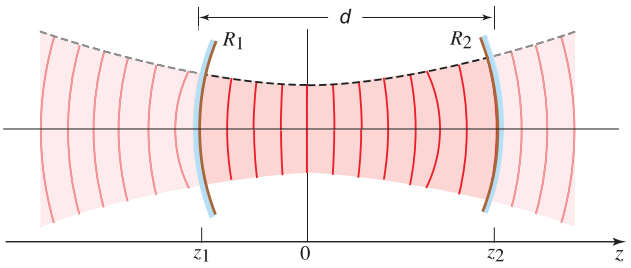


Figure 11.2-6 Fitting a Gaussian beam to two mirrors separated by a distance d. Their radii of curvature are R_1 and R_2. Both mirrors are taken to be concave so that R_1 and R_2 are negative, as is z_1.

distance d. The task is illustrated in Fig. 11.2-6 for the special case when both mirrors are concave ($R_1 < 0$ and $R_2 < 0$).

The z axis is defined by the centers of the mirrors. The center of the beam, which is yet to be determined, is assumed to be located at the origin $z = 0$; mirrors R_1 and R_2 are located at positions z_1 and

$$z_2 = z_1 + d, \tag{11.2-11}$$

respectively. A negative value for z_1 indicates that the center of the beam lies to the right of mirror 1; a positive value indicates that it lies to the left. The values of z_1 and z_2 are determined by matching the radius of curvature of the beam, $R(z) = z + z_0^2/z$, to the radii R_1 at z_1 and R_2 at z_2. Careful attention must be paid to the signs. If both mirrors are concave, they have negative radii. But the beam radius of curvature was defined to be positive for $z > 0$ (at mirror 2) and negative for $z < 0$ (at mirror 1). We therefore equate $R_1 = R(z_1)$, but $-R_2 = R(z_2)$, to obtain

$$R_1 = z_1 + z_0^2/z_1 \tag{11.2-12}$$

$$-R_2 = z_2 + z_0^2/z_2. \tag{11.2-13}$$

Solving (11.2-11), (11.2-12), and (11.2-13) for z_1, z_2, and z_0 leads to

$$z_1 = \frac{-d(R_2 + d)}{R_2 + R_1 + 2d}, \qquad z_2 = z_1 + d, \tag{11.2-14}$$

$$z_0^2 = \frac{-d(R_1 + d)(R_2 + d)(R_2 + R_1 + d)}{(R_2 + R_1 + 2d)^2}, \tag{11.2-15}$$

which accord with (3.1-27) and (3.1-28) (if R_2 is replaced with $-R_2$).

Having determined the location of the beam center and the depth of focus $2z_0$, everything about the beam is known (see Sec. 3.1B). The waist radius is $W_0 = \sqrt{\lambda z_0/\pi}$, and the beam radii at the mirrors are

$$W_i = W_0\sqrt{1 + \left(\frac{z_i}{z_0}\right)^2}, \qquad i = 1, 2. \tag{11.2-16}$$

In order that the solution (11.2-14)–(11.2-15) indeed represents a Gaussian beam, z_0 must be real. An imaginary value of z_0 would signify that the Gaussian beam is a paraboloidal wave, which is an unconfined solution of the paraxial Helmholtz equation

(see Sec. 3.1A). Using (11.2-15), it is not difficult to show that the condition $z_0^2 > 0$ is equivalent to

$$0 \le \left(1 + \frac{d}{R_1}\right)\left(1 + \frac{d}{R_2}\right) \le 1. \tag{11.2-17}$$

This is precisely the confinement condition derived from ray optics as set forth in (11.2-5).

EXERCISE 11.2-2

A Plano-Concave Resonator. If mirror 1 is planar ($R_1 = \infty$), determine the confinement condition and the depth of focus, as well as the beam width at the waist and at each of the mirrors, as a function of $d/|R_2|$.

Gaussian Mode of a Symmetric Spherical-Mirror Resonator

The results provided in (11.2-11)–(11.2-15) simplify considerably for symmetric resonators with concave mirrors. Substituting $R_1 = R_2 = -|R|$ into (11.2-14) provides $z_1 = -d/2$ and $z_2 = d/2$. The beam center thus lies at the center of the resonator, and

$$z_0 = \frac{d}{2}\sqrt{2\frac{|R|}{d} - 1}\,, \tag{11.2-18}$$

$$W_0^2 = \frac{\lambda d}{2\pi}\sqrt{2\frac{|R|}{d} - 1}\,, \tag{11.2-19}$$

$$W_1^2 = W_2^2 = \frac{\lambda d/\pi}{\sqrt{(d/|R|)\,[2 - (d/|R|)]}}\,. \tag{11.2-20}$$

The confinement condition (11.2-17) becomes

$$0 \le \frac{d}{|R|} \le 2. \tag{11.2-21}$$

Given a resonator of fixed mirror separation d, we now examine the effect of increasing mirror curvature on the beam radius at the waist W_0, and at the mirrors $W_1 = W_2$. (Increasing curvature corresponds to increasing $d/|R|$ since the *radius of curvature* diminishes as the *curvature* increases.) The results are illustrated in Fig. 11.2-7. For a planar-mirror resonator, $d/|R| = 0$, so that W_0 and W_1 are infinite, corresponding to a plane wave rather than a Gaussian beam. As $d/|R|$ increases, W_0 decreases until it vanishes for the concentric resonator ($d/|R| = 2$); at this point $W_1 = W_2 = \infty$ and $W_0 = 0$. In this limit, the resonator supports a spherical wave instead of a Gaussian beam.

The width of the beam at the mirrors attains its minimum value, $W_1 = W_2 = \sqrt{\lambda d/\pi}$, when $d/|R| = 1$, i.e., for the symmetric confocal resonator. In this case

$$z_0 = d/2, \tag{11.2-22}$$

$$W_0 = \sqrt{\lambda d/2\pi}\,, \tag{11.2-23}$$

$$W_1 = W_2 = \sqrt{2}\,W_0. \tag{11.2-24}$$

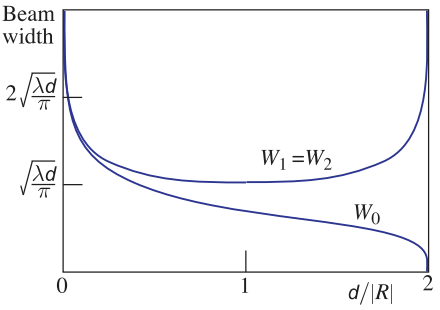


Figure 11.2-7 The beam width at the waist, W_0, and at the mirrors, $W_1 = W_2$, for a symmetric spherical-mirror resonator with concave mirrors, as a function of the ratio $d/|R|$. The planar-mirror resonator corresponds to $d/|R| = 0$. Symmetric confocal and concentric resonators correspond to $d/|R| = 1$ and $d/|R| = 2$, respectively.

The depth of focus $2z_0$ is then equal to the length of the resonator d, as shown in Fig. 11.2-8. This explains why the parameter $2z_0$ is sometimes called the confocal parameter. A long resonator has a long depth of focus. The waist radius is proportional to the square root of the mirror spacing. A Gaussian beam at $\lambda_0 = 633$ nm (a He–Ne laser wavelength) in a resonator with $d = 100$ cm, for example, has a waist radius $W_0 = \sqrt{\lambda d/2\pi} = 0.32$ mm, whereas a 25-cm-long resonator supports a Gaussian beam with a waist radius that is half as big at the same wavelength: 0.16 mm. The width of the beam at each of the mirrors is greater than it is at the waist by a factor of $\sqrt{2}$.

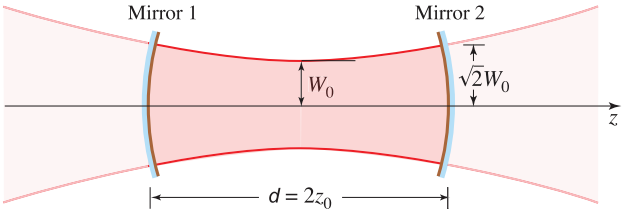


Figure 11.2-8 Gaussian beam in a symmetric confocal resonator with concave mirrors. The depth of focus $2z_0$ equals the length of the resonator d. The beam width at the mirrors is a factor of $\sqrt{2}$ greater than that at the waist.

C. Resonance Frequencies

As indicated in Sec. 11.2B, a Gaussian beam is a mode of the spherical-mirror resonator provided that the wavefront normals reflect back onto themselves, always retracing the same path, and that the phase retraces itself as well.

The phase of a Gaussian beam, in accordance with (3.1-23), is

$$\varphi(\rho, z) = kz - \zeta(z) + \frac{k\rho^2}{2R(z)}, \tag{11.2-25}$$

where $\zeta(z) = \tan^{-1}(z/z_0)$ and $\rho^2 = x^2 + y^2$. At points on the optical axis ($\rho = 0$), $\varphi(0, z) = kz - \zeta(z)$, so that the phase retardation relative to a plane wave is $\zeta(z)$. At the locations of the mirrors, z_1 and z_2, we therefore have

$$\varphi(0, z_1) = kz_1 - \zeta(z_1), \tag{11.2-26}$$

$$\varphi(0, z_2) = kz_2 - \zeta(z_2). \tag{11.2-27}$$

Because the mirror surface coincides with the wavefronts, all points on each mirror share the same phase. As the beam propagates from mirror 1 to mirror 2, its phase changes by

$$\begin{aligned}\varphi(0, z_2) - \varphi(0, z_1) &= k(z_2 - z_1) - [\zeta(z_2) - \zeta(z_1)] \\ &= kd - \Delta\zeta,\end{aligned} \tag{11.2-28}$$

where

$$\Delta\zeta = \zeta(z_2) - \zeta(z_1). \tag{11.2-29}$$

As the traveling wave completes a round trip between the two mirrors, therefore, its phase changes by $2kd - 2\Delta\zeta$.

In order that the beam truly retrace itself, the round-trip phase change must be zero or a multiple of $\pm 2\pi$, i.e., $2kd - 2\Delta\zeta = 2\pi q$, $q = 0, \pm 1, \pm 2, \ldots$. Using the substitutions $k = 2\pi\nu/c$ and $\nu_F = c/2d$, the frequencies ν_q that satisfy this condition are

$$\boxed{\nu_q = q\nu_F + \frac{\Delta\zeta}{\pi}\nu_F.} \tag{11.2-30}$$

Resonance Frequencies
Gaussian Modes

The frequency spacing of adjacent modes is therefore $\nu_F = c/2d$, which is identical to the result obtained in Sec. 11.1A for the planar-mirror resonator. For spherical-mirror resonators, this frequency spacing is evidently independent of the curvatures of the mirrors. The second term in (11.2-30), which does depend on the mirror curvatures, simply represents a displacement of all resonance frequencies.

EXERCISE 11.2-3

Resonance Frequencies of a Confocal Resonator. A symmetric confocal resonator has a length $d = 30$ cm, and the medium has refractive index $n = 1$. Determine the frequency spacing ν_F and the displacement frequency $(\Delta\zeta/\pi)\,\nu_F$. Determine all resonance frequencies that lie within the band $5 \times 10^{14} \pm 2 \times 10^9$ Hz.

D. Hermite–Gaussian Modes

In Sec. 3.3 it was shown that the Gaussian beam is not the only beam-like solution of the paraxial Helmholtz equation. The family of Hermite–Gaussian beams also provides solutions. Although a Hermite–Gaussian beam of order (l, m) has an amplitude distribution that differs from that of the Gaussian beam, their wavefronts are identical. As a result, the design of a resonator that "matches" a given beam (or the design of a beam that "fits" a given resonator) is the same as for the Gaussian beam, whatever the values of (l, m). It follows that all members of the family of Hermite–Gaussian beams represent modes of the spherical-mirror resonator.

The resonance frequencies of the (l, m) mode do, however, depend on the indices (l, m). This is because of the dependence of the Gouy phase shift on l and m. As is evident from (3.3-10), the phase of the (l, m) mode on the beam axis is

$$\varphi(0, z) = kz - (l + m + 1)\zeta(z). \tag{11.2-31}$$

Again, the phase shift encountered by a traveling wave undergoing a single round trip through a resonator of length d must be set equal to zero or an integer multiple of $\pm 2\pi$ in order that the beam retrace itself. Thus,

$$2kd - 2(l + m + 1)\Delta\zeta = 2\pi q, \qquad q = 0, \pm 1, \pm 2, \ldots, \tag{11.2-32}$$

where, as previously, $\Delta\zeta = \zeta(z_2) - \zeta(z_1)$ and z_1, z_2 represent the positions of the two mirrors. With $k = 2\pi\nu/c$ and $\nu_F = c/2d$, this yields the resonance frequencies

$$\boxed{\nu_{l,m,q} = q\nu_F + (l + m + 1)\,\frac{\Delta\zeta}{\pi}\,\nu_F.} \tag{11.2-33}$$

Resonance Frequencies
Hermite–Gaussian Modes

Modes of different q, but the same (l, m), have identical intensity distributions [see (3.3-12)]. They are known as **longitudinal** or **axial modes**. The indices (l, m) label different spatial dependencies on the transverse coordinates x, y; these therefore represent different **transverse modes**, as illustrated in Fig. 3.3-2.

Equation (11.2-33) dictates that the resonance frequencies of the Hermite–Gaussian modes satisfy the following properties:

- Longitudinal modes corresponding to a given transverse mode have resonance frequencies spaced by $\nu_F = c/2d$ since $\nu_{l,m,q+1} - \nu_{l,m,q} = \nu_F$. This result is the same as that obtained for the (0,0) Gaussian mode and for the planar-mirror resonator.
- All transverse modes, for which the sum of the indices $l + m$ is the same, have the same resonance frequencies.
- Two transverse modes (l, m), (l', m') corresponding to the same longitudinal mode q have resonance frequencies spaced by

$$\nu_{l,m,q} - \nu_{l',m',q} = [(l + m) - (l' + m')]\,\frac{\Delta\zeta}{\pi}\,\nu_F. \tag{11.2-34}$$

  This expression determines the frequency shift between the sets of longitudinal modes of indices (l, m) and (l', m').

EXERCISE 11.2-4

Resonance Frequencies of the Symmetric Confocal Resonator. Show that for a symmetric confocal resonator, the longitudinal modes associated with different transverse modes are either the same, or are displaced by $\nu_F/2$, as illustrated in Fig. 11.2-9.

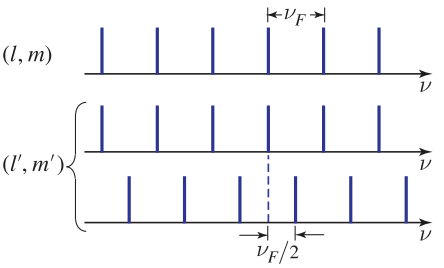


Figure 11.2-9 In a symmetric confocal resonator, the longitudinal modes associated with two transverse modes of indices (l, m) and (l', m') are either aligned or displaced by half a longitudinal mode spacing.

*E. Finite Apertures and Diffraction Loss

Since Gaussian and Hermite–Gaussian beams have infinite transverse extent whereas the resonator mirrors are of finite extent, a portion of the optical power leaks around the mirrors and escapes from the resonator on each pass. An estimate of the power loss may be obtained by calculating the fractional power of the beam that is not intercepted by the mirror. If the beam is Gaussian with width W and the mirror is circular with radius $a = 2W$, for example, a small fraction, $\exp(-2a^2/W^2) \approx 3.35 \times 10^{-4}$, of the beam power escapes on each pass [see (3.1-17)], the remainder being reflected (or transmitted through the mirror). Higher-order transverse modes suffer greater losses since they have greater spatial extent in the transverse plane.

When the mirror radius a is smaller than $2W$, the losses are greater. The resonator modes are then less well described by Gaussian and Hermite–Gaussian beams. The problem of determining the modes of a spherical-mirror resonator with finite-size mirrors is difficult. A wave is a mode if it retraces its amplitude (to within a multiplicative constant) and reproduces its phase (to within an integer multiple of 2π) after completing a round trip through the resonator. One oft-used method for determining the modes involves following a wave repeatedly as it bounces through the resonator, thereby determining its amplitude and phase, much as we determined the position and inclination of a ray bouncing within a resonator. After many round trips this process converges to one of the modes.

The configuration for implementing this approach in a spherical-mirror resonator is schematized in Fig. 11.2-10.

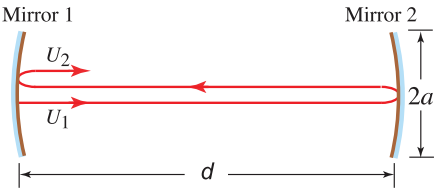


Figure 11.2-10 Propagation of a wave through a spherical-mirror resonator. The complex amplitude $U_1(x, y)$ corresponds to a mode if it reproduces itself after a round trip, i.e., if $U_2(x, y) = \mu U_1(x, y)$ and $\arg\{\mu\} = q2\pi$.

If $U_1(x, y)$ is the complex amplitude of a wave immediately to the right of mirror 1 in Fig. 11.2-10, and if $U_2(x, y)$ is the complex amplitude after one round trip of travel through the resonator, then $U_1(x, y)$ is a mode provided that $U_2(x, y) = \mu U_1(x, y)$ and provided that $\arg\{\mu\}$ is an integer multiple of 2π (i.e., μ is real and positive). After a single round trip, the mode intensity is attenuated by the factor μ^2, and the phase is reproduced. The methods of Fourier optics (Chapter 4) may be used to determine $U_2(x, y)$ from $U_1(x, y)$. These quantities may be regarded as the output and input, respectively, of a linear system (see Appendix B) characterized by an impulse response function $h(x, y; x', y')$, so that

$$U_2(x, y) = \int_{-\infty}^{\infty}\int h(x, y; x', y')\, U_1(x', y')\, dx'\, dy'. \tag{11.2-35}$$

If the impulse response function h is known, the modes can be determined by solving the eigenvalue problem described by the integral equation (see Appendix C)

$$\int_{-\infty}^{\infty}\int h(x, y; x', y')\, U(x', y')\, dx'\, dy' = \mu\, U(x, y). \tag{11.2-36}$$

The solutions determine the eigenfunctions $U_{l,m}(x, y)$, and the eigenvalues $\mu_{l,m}$, labeled by the indices (l, m). The eigenfunctions are the modes and the eigenvalues are the round-trip multiplicative factors. The squared magnitude $|\mu_{l,m}|^2$ is the round-trip intensity reduction factor for the (l, m) mode. Clearly, when the mirrors are infinite in size and the paraxial approximation is satisfied, the modes reduce to the family of Hermite–Gaussian beams discussed earlier.

It remains to determine $h(x, y; x', y')$ and to solve the integral equation (11.2-36). A single pass inside the resonator involves traveling a distance d, truncation by the mirror aperture, and reflection by the mirror. The remaining pass, needed to comprise a single round trip, is similar. The impulse response function $h(x, y; x', y')$ can then be determined by applying the theory of Fresnel diffraction (Sec. 4.3B). In general, however, the modes and their associated losses can be determined only by numerically solving the integral equation (11.2-36). An iterative numerical solution begins with an initial guess U_1, from which U_2 is computed and passed through the system one more round trip, and so on until the process converges.

This technique has been used to determine the losses associated with the various modes of a spherical-mirror resonator with circular mirror apertures of radius a. The results are illustrated in Fig. 11.2-11 for a symmetric confocal resonator. The loss is governed by a single parameter, the Fresnel number $N_F = a^2/\lambda d$. This is because the Fresnel number governs Fresnel diffraction between the two mirrors, as discussed in Sec. 4.3B. For the symmetric confocal resonator described by (11.2-23) and (11.2-24), the beam width at the mirrors is $W = \sqrt{\lambda d/\pi}$, so that $\lambda d = \pi W^2$, from which the Fresnel number is readily determined to be $N_F = a^2/\pi W^2$. N_F is therefore proportional to the ratio a^2/W^2; a higher Fresnel number corresponds to a smaller loss. From Fig. 11.2-11 we find that the loss per pass of the lowest-order symmetric-confocal-resonator mode $(l, m) = (0, 0)$ is about 0.1% when $N_F \approx 0.94$. This Fresnel number corresponds to $a/W = 1.72$. If the beam were Gaussian with width W, the percentage of power contained outside a circle of radius $a = 1.72\,W$ would be $\exp(-2a^2/W^2) \approx 0.27\%$. This is larger than the 0.1% loss per pass for the actual resonator mode. Higher-order modes suffer from greater losses because of their greater spatial extent.

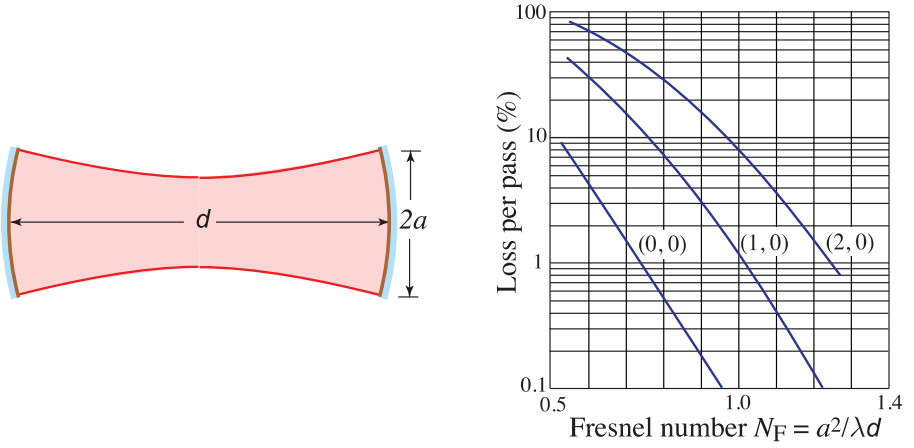


Figure 11.2-11 Percent diffraction loss per pass (half a round trip) as a function of the Fresnel number $N_F = a^2/\lambda d$ for the $(0, 0)$, $(1, 0)$, and $(2, 0)$ modes in a symmetric confocal resonator. (Adapted from A. E. Siegman, *Lasers*, University Science, 1986, Fig. 19.19 left.)

11.3 TWO- AND THREE-DIMENSIONAL RESONATORS

A. Two-Dimensional Rectangular Resonators

A two-dimensional (2D) planar-mirror resonator is constructed from two orthogonal pairs of parallel mirrors, e.g., a pair normal to the z axis and another pair normal to the y axis. Light is confined in the z–y plane by a sequence of ray reflections, as illustrated in Fig. 11.3-1(a).

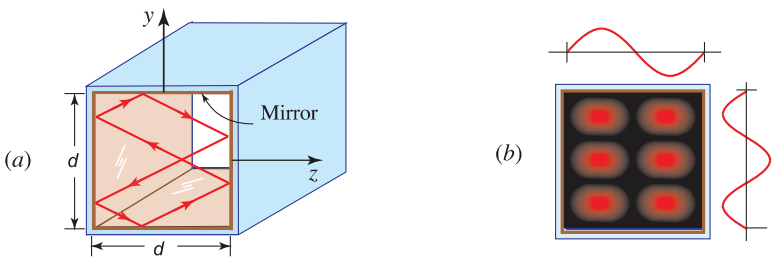


Figure 11.3-1 A two-dimensional planar-mirror resonator: (a) ray pattern; (b) standing-wave pattern with mode numbers $q_y = 3$ and $q_z = 2$. The curves represent the modal amplitudes while the brightness pattern depicts intensity.

The boundary conditions establish the resonator modes, much as for the one-dimensional Fabry–Perot resonator. If the mirror spacing is d, then for standing waves the components of the wavevector $\mathbf{k} = (k_y, k_z)$ are restricted to the values

$$k_y = q_y \frac{\pi}{d}, \quad k_z = q_z \frac{\pi}{d}, \quad q_y = 1, 2, \ldots, \quad q_z = 1, 2, \ldots, \tag{11.3-1}$$

where q_y and q_z are mode numbers for the y and z directions, respectively. These conditions are a generalization of (11.1-2). Each pair of integers (q_y, q_z) represents a resonator mode $U(\mathbf{r}) \propto \sin(q_y \pi y/d) \sin(q_z \pi z/d)$, as illustrated in Fig. 11.3-1(b). The lowest-order mode is $(1, 1)$ since the modes $(q_y, 0)$ and $(0, q_z)$ have zero amplitude, i.e., $U(\mathbf{r}) = 0$. Modes are conveniently represented by dots that indicate their values of k_y and k_z on a periodic lattice of spacing π/d (Fig. 11.3-2).

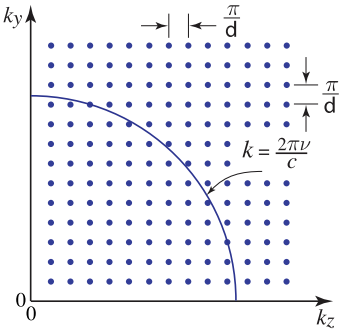


Figure 11.3-2 Dots denote the endpoints of the wavevectors $\mathbf{k} = (k_y, k_z)$ for modes in a two-dimensional resonator.

The wavenumber k of a mode is the distance of the dot from the origin. The associated frequency of the mode is $\nu = ck/2\pi$. The frequencies of the resonator modes are

thus determined from

$$k^2 = k_y^2 + k_z^2 = \left(\frac{2\pi\nu}{c}\right)^2, \tag{11.3-2}$$

so that

$$\boxed{\nu_{\mathbf{q}} = \nu_F\sqrt{q_y^2 + q_z^2}, \quad q_y, q_z = 1, 2, \ldots,} \qquad \nu_F = \frac{c}{2d}, \tag{11.3-3}$$

(11.3-3) Resonance Frequencies

where $\mathbf{q} = (q_y, q_z)$.

The number of modes in a given frequency band, $\nu_1 < \nu < \nu_2$, is established by drawing two circles, of radii $k_1 = 2\pi\nu_1/c$ and $k_2 = 2\pi\nu_2/c$ in the k diagram of Fig. 11.3-2, and counting the number of dots that lie within the annulus. This procedure converts the allowed values of the vector $\mathbf{k}$ into allowed values of the frequency ν.

EXERCISE 11.3-1

Density of Modes in a Two-Dimensional Resonator.

(a) Determine an approximate expression for the number of modes in a two-dimensional resonator with frequencies lying between 0 and ν, assuming that $2\pi\nu/c \gg \pi/d$, i.e., $d \gg \lambda/2$, and allowing for two orthogonal polarizations per mode.

(b) Show that the number of modes per unit area lying within the frequency interval between ν and $\nu + d\nu$ is $M(\nu)d\nu$, where the density of modes $M(\nu)$ (modes per unit area per unit frequency) at frequency ν is given by

$$\boxed{M(\nu) = \frac{4\pi\nu}{c^2}.} \tag{11.3-4}$$

(11.3-4) Density of Modes (2D Resonator)

The resonator modes described thus far in this section are in-plane modes, traveling in the plane of the 2D resonator (the y–z plane). Off-plane modes have a propagation constant with a component in the orthogonal direction (the x direction). These are guided modes traveling along the axis of a 2D waveguide such as that described in Sec. 9.3. Whereas the k_y and k_z components of the wavevector take discrete values dictated by the boundary conditions, the k_x component takes continuous values since the 2D resonator is open in the x direction.

B. Circular Resonators and Whispering-Gallery Modes

Light may be confined in a two-dimensional circular resonator by repeated reflections from the circular boundary. As illustrated in Fig. 11.3-3, a ray that self-reproduces after N reflections traces a path with round-trip pathlength Nd, where $d = 2a\sin(\pi/N)$ and a is the radius. For a traveling-wave mode, the resonance frequencies are determined by equating the round-trip pathlength to an integer number of wavelengths, as in (11.1-7). Ignoring the phase shift associated with each reflection, this leads to $Nd = q\lambda = qc/\nu$, i.e., to resonant frequencies $\nu_q = qc/Nd$, where $q = 1, 2, \ldots$. The spacing between these frequencies is therefore $\nu_F = c/Nd$.

For $N = 2$, we have $\nu_F = c/2d = c/4a$, which is identical to (11.1-5). Similarly, $N = 3$ yields $\nu_F = c/3d = c/3\sqrt{3}\,a$, which coincides with the result for the three-mirror resonator (Exercise 11.1-1). In the limit $N \to \infty$, the pathlength Nd approaches the cylindrical circumference $2\pi a$ and the corresponding spacing of the resonance frequencies becomes

$$\boxed{\nu_F = \frac{c}{2\pi a}.} \qquad (11.3\text{-}5)$$

Spacing of Resonance Frequencies

The rays then hug the interior boundary of the resonator, reflecting at near-grazing incidence, as illustrated in Fig. 11.3-3. Such optical modes are known as **whispering-gallery modes** (WGM). The optical modes then behave similarly to acoustic modes in the familiar acoustical whispering gallery, so-named because of the ease with which an acoustic whisper can bounce along the convex surface of a church dome or gallery.

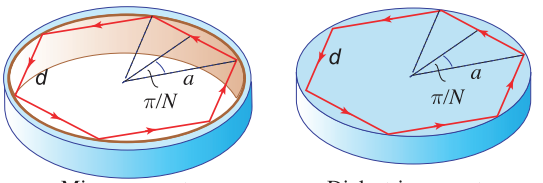


Figure 11.3-3 Reflections in a circular resonator.

Two-dimensional resonators with other cross sections are also used. For example, the circular cross section can be squeezed into a stadium-shaped structure. This oblong configuration supports bow-tie modes [see Fig. 11.1-5(*b*)] in which the ray executes a round-trip path comprising localized reflections from the four locations on the perimeter of the resonator that match the curvature of a conventional spherical-mirror confocal resonator (see Sec. 11.2A).

C. Three-Dimensional Rectangular Resonators

A three-dimensional (3D) planar-mirror resonator is constructed from three pairs of parallel mirrors forming the walls of a closed rectangular box of dimensions d_x, d_y, and d_z. The structure is a three-dimensional resonator, as depicted in Fig. 11.3-4(*a*). Standing-wave solutions within the resonator require that the components of the wavevector $\mathbf{k} = (k_x, k_y, k_z)$ are discretized to obey

$$k_x = q_x \frac{\pi}{d_x}, \quad k_y = q_y \frac{\pi}{d_y}, \quad k_z = q_z \frac{\pi}{d_z}, \qquad q_x, q_y, q_z = 1, 2, \ldots, \qquad (11.3\text{-}6)$$

where q_x, q_y, and q_z are positive integers representing the respective mode numbers. Each mode $\mathbf{q}$, which is characterized by the three integers (q_x, q_y, q_z), is represented by a dot in (k_x, k_y, k_z)-space. The spacing between these dots in a given direction is inversely proportional to the width of the resonator along that direction. Figure 11.3-4(*b*) illustrates the concept of the k-space for a cubic resonator with $d_x = d_y = d_z = d$.

The values of the wavenumbers k, and the corresponding resonance frequencies ν, satisfy

$$k^2 = k_x^2 + k_y^2 + k_z^2 = \left(\frac{2\pi\nu}{c}\right)^2. \qquad (11.3\text{-}7)$$

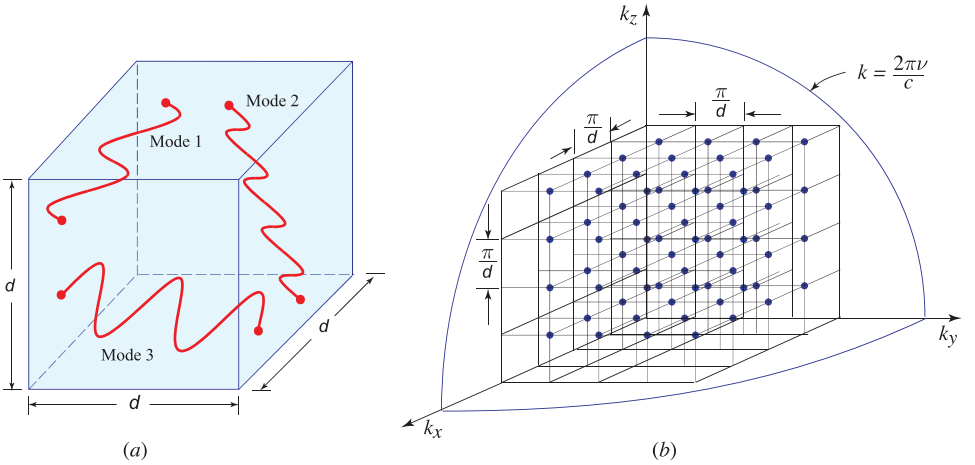


Figure 11.3-4 (*a*) Waves in a three-dimensional cubic resonator ($d_x = d_y = d_z = d$). (*b*) The endpoints of the wavevectors (k_x, k_y, k_z) of the modes in a three-dimensional resonator are marked by dots. The wavenumber k of a mode is the distance from the origin to the dot. Each point in k-space occupies a volume $(\pi/d)^3$. All modes of frequency smaller than ν lie inside the positive octant of a sphere of radius $k = 2\pi\nu/c$.

The surface of constant frequency ν is a sphere of radius $k = 2\pi\nu/c$. The resonance frequencies are determined from (11.3-6) and (11.3-7):

$$\nu_{\mathbf{q}} = \sqrt{q_x^2\,\nu_{Fx}^2 + q_y^2\,\nu_{Fy}^2 + q_z^2\,\nu_{Fz}^2}\,, \quad q_x, q_y, q_z = 1, 2, \ldots, \tag{11.3-8}$$

Resonance Frequencies

where

$$\nu_{Fx} = \frac{c}{2d_x}\,, \quad \nu_{Fy} = \frac{c}{2d_y}\,, \quad \nu_{Fz} = \frac{c}{2d_z} \tag{11.3-9}$$

are frequency spacings that are inversely proportional to the resonator widths in the x, y, and z direction, respectively. For resonators whose dimensions are much greater than a wavelength, the frequency spacing is much smaller than the optical frequency. For example, for $d = 1$ cm and $n = 1$, $\nu_F = 15$ GHz. This is not so for microresonators, however, as will be discussed in Sec. 11.4.

Density of Modes

When all dimensions of the resonator are much greater than a wavelength, the frequency spacing $\nu_F = c/2d$ is small, and it is analytically difficult to enumerate the modes. In this case, it is useful to resort to a continuous approximation and introduce the concept of density of modes, the validity of which depends on the relative values of the bandwidth of interest and the frequency interval between successive modes.

The number of modes lying in the frequency interval between 0 and ν corresponds to the number of points lying in the volume of the positive octant of a sphere of radius k in the k diagram [Fig. 11.3-4(*b*)]. The number of modes in the positive octant of a sphere of radius k is $2(\frac{1}{8})(\frac{4}{3}\pi k^3)/(\pi/d)^3 = (k^3/3\pi^2)d^3$. The initial factor of 2 accounts for the two possible polarizations of each mode, whereas the denominator $(\pi/d)^3$ represents the volume in k-space per point. It follows that the number of

modes with wavenumbers between k and $k + \Delta k$, per unit volume, is $\varrho(k)\Delta k = [(d/dk)(k^3/3\pi^2)]\Delta k = (k^2/\pi^2)\Delta k$, so that the density of modes in k-space is $\varrho(k) = k^2/\pi^2$. It is worthy of mention that this derivation is identical to that used for determining the density of allowed quantum states for electron waves confined within perfectly reflecting walls in a bulk semiconductor [see Sec. 17.1C and (17.1-6)].

Since $k = 2\pi\nu/c$, the number of modes lying between 0 and ν is $[(2\pi\nu/c)^3/3\pi^2]d^3 = (8\pi\nu^3/3c^3)d^3$. The number of modes in the incremental frequency interval lying between ν and $\nu + \Delta\nu$ is therefore given by $(d/d\nu)[(8\pi\nu^3/3c^3)d^3]\Delta\nu = (8\pi\nu^2/c^3)d^3\Delta\nu$. The density of modes $M(\nu)$, i.e., the number of modes per unit volume of the resonator, per unit bandwidth surrounding the frequency ν, is thus

$$M(\nu) = \frac{8\pi\nu^2}{c^3}. \tag{11.3-10}$$

Density of Modes (3D Resonator)

This formula was used by Rayleigh and Jeans in connection with the spectrum of blackbody radiation (see Sec. 14.4B). The quantity $M(\nu)$ is a quadratically increasing function of frequency so that the number of modes within a fixed bandwidth $\Delta\nu$ increases with the frequency ν in the manner indicated in Fig. 11.3-5. As an example, at $\nu = 3 \times 10^{14}$ ($\lambda_o = 1\ \mu$m), $M(\nu) = 0.08\ \text{modes/cm}^3\text{-Hz}$. Within a frequency band of width 1 GHz, there are therefore $\approx 8 \times 10^7\ \text{modes/cm}^3$. The number of modes per unit volume within an arbitrary frequency interval $\nu_1 < \nu < \nu_2$ is simply the integral $\int_{\nu_1}^{\nu_2} M(\nu)\, d\nu$.

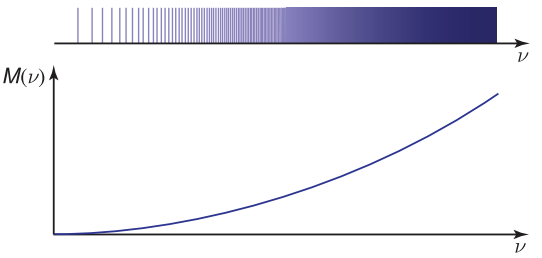


Figure 11.3-5 (a) The frequency spacing between adjacent modes decreases as the frequency increases. (b) The density of modes $M(\nu)$ for a three-dimensional optical resonator is a quadratically increasing function of frequency.

The density of modes in two and three dimensions were derived on the basis of square and cubic geometry, respectively. Nevertheless, the results are applicable for arbitrary geometries, provided that the resonator dimensions are large in comparison with the wavelength.

11.4 MICRORESONATORS AND NANORESONATORS

Microresonators are resonators in which one or more of the spatial dimensions assumes the size of a few wavelengths of light or smaller. The term **microcavity resonator**, or **microcavity** for short, is usually reserved for a microresonator that has small dimensions in all spatial directions, so that the modes exhibit large spacings in all directions of k-space and the resonance frequencies are sparse. However, these terms are often used interchangeably.

The absence of resonance modes in extended spectral bands can inhibit the emission of light from sources placed within a microcavity. At the same time, the emission of light into particular modes of a high-Q, small-volume microcavity can be enhanced relative to emission into ordinary optical modes, as described in Sec. 14.3E. These effects can be important in the operation of resonant-cavity light-emitting diodes (RCLEDs) and microcavity lasers (see Secs. 18.1B and 18.5, respectively).

Microresonators can be fabricated using dielectric materials configured in various geometries, such as (1) micropillars with Bragg-grating reflectors; (2) microdisks and microspheres in which light reflects near the surface in whispering-gallery modes; (3) microtoroids, which resemble small fiber rings; and (4) 2D photonic crystals containing light-trapping defects that function as microcavities. These technologies embrace two principal design objectives:

- The reduction of the modal volume V, the spatial integral of the optical energy density $\frac{1}{2}\epsilon\,\mathcal{E}^2$ of the mode, normalized to the maximum energy density.
- The enhancement of the quality factor Q.

Spatial confinement is improved by fabricating microresonators with special geometries, while enhancement of temporal confinement is realized by making use of low-loss materials and low-leakage configurations. Typical modal volumes and quality factors for these structures are summarized in Table 11.4-1.

Table 11.4-1 Normalized modal volume V/λ^3 and quality factor Q for various microresonators.

	Micropillar	Microdisk	Microtoroid	Microsphere	Photonic-Crystal
V/λ^3	5	5	10^3	10^3	1
Q	10^3	10^4	10^8	10^{10}	10^4

An exact analysis of the resonator modes of dielectric microresonators requires the full electromagnetic theory. The Helmholtz equation is solved in a coordinate system suitable for the geometry of the structure, and appropriate boundary conditions are applied to the electric and magnetic fields at the planar, cylindrical, or spherical boundaries. The solution yields the resonance frequencies of the modes and their spatial distributions, which may be used to determine the modal volume for each mode. Since the analysis is complex for all practical geometries, numerical solutions are often necessary.

In the next section, we describe some of the properties of a simple rectangular (box) microresonator whose walls are made of perfect mirrors. A simple analysis of the modes of such a structure provides the resonance frequencies and the spatial distributions of the modes. High-Q microresonators do not make use of mirrors because of their relatively high losses, and the box structure is also not among the geometries typically used in practical microresonators. Nevertheless, the analysis is useful for elucidating the relation between the resonance frequencies and the dimensions of the resonator, and for illustrating the frequency dependence of the density of modes for boxes with different aspect ratios.

A. Rectangular Microresonators

The simplest microresonator structure is a rectangular (box) resonator made of planar parallel mirrors. The modes are then sinusoidal standing waves in all three directions and the resonance frequencies are given by (11.3-8). When the dimensions of the box are small, only the lowest order modes lie within the optical band. For a cubic resonator, the resonance frequencies are provided in Table 11.4-2 in units of $\nu_F = c/2d$. As an

example, if $d = 1\ \mu$m and the medium has refractive index $n = 1.5$, we obtain $\nu_F = 100$ THz. The frequencies of the lowest-order modes then correspond to the free-space wavelengths $\lambda_o = 2.13, 1.73, 1.34, 1.22, 1.06, 1.00,$ and 0.87μm, which are widely spaced.

Table 11.4-2 Resonance frequencies for the lowest-order modes of a cubic microcavity resonator.

Mode $(q_x\ q_y\ q_z)^{(a)}$	$(011)^{(3)}$	$(111)^{(1)}$	$(012)^{(6)}$	$(112)^{(3)}$	$(022)^{(3)}$	$(122)^{(3)}$	$(222)^{(1)}$
Frequency (units of ν_F)	1.41	1.73	2.24	2.45	2.83	3	3.46

[a] Superscripts in parentheses indicate the modal degeneracy, i.e., the number of modes of the same resonance frequency. As an example, three modes have the same resonance frequency $1.41\ \nu_F$: (011), (101), and (110).

If the resonator has a mixture of dimensions both small and large, as with a box of large aspect ratio, the modes are placed at the points of an anisotropic grid in k-space [see Fig. 11.3-4(b)]. The grid is finely divided along the directions of the large dimensions and coarsely divided along the directions of the small dimensions. Mode counting may then be implemented by use of a continuous approximation only in those directions for which the grid is fine. The resultant modal density is displayed in Fig. 11.4-1 for various cases.

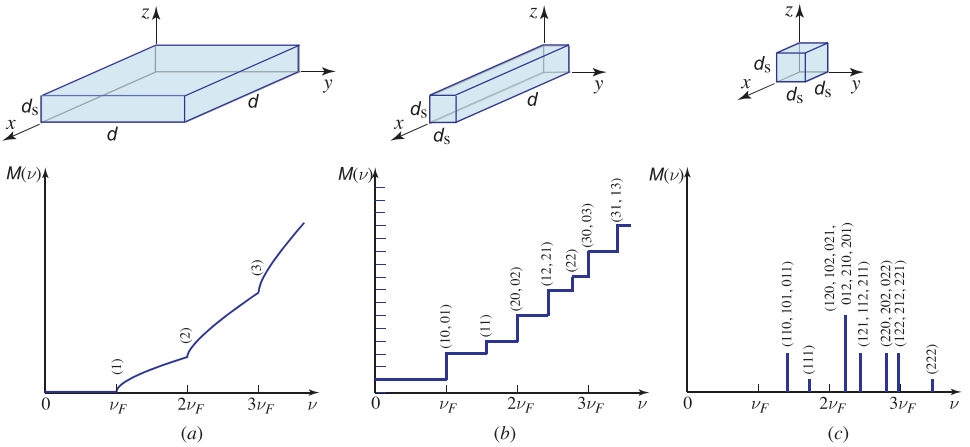


Figure 11.4-1 Modal density $M(\nu)$ for rectangular microresonators with (a) one; (b) two; and (c) three sides of small dimension $d_S \ll d$. The frequency spacing associated with the small dimension is $\nu_F = c/2d_S$. When all dimensions are small, as in (c), the resonance frequencies are discrete and their values are those provided in Table 11.4-2 for the cubic microcavity resonator. The result shown in (b) represents a combination of discrete modes associated with a 2D microresonator and continuous modes associated with a 1D large resonator, which has a uniform modal density [see (11.1-10)]. The result provided in (a) illustrates a combination of discrete modes associated with a 1D microresonator and a continuum of modes associated with a 2D large resonator, which has a modal density that is linearly proportional to frequency [see (11.3-4)].

B. Micropillars, Microdisks, and Microtoroids

Dielectric microresonators have been fabricated in a number of configurations, including micropillars, microdisks, and microtoroids, as illustrated in Fig. 11.4-2. Light is confined in these structures by total internal reflection (see Fig. 11.3-3).

The **micropillar**, or **micropost**, resonator is a cylinder of high-refractive-index material sandwiched between dielectric layers comprising distributed Bragg-grating reflectors, as illustrated in Fig. 11.4-2(a). Light is confined in the axial direction by reflection from the DBRs, as in a Fabry–Perot resonator; light is confined in the lateral direction by total internal reflection from the walls of the cylinder. Micropillars are typically fabricated from compound semiconductors via conventional lithographic and etching processes; DBR layers are often made of AlAs/GaAs or AlGaAs/GaAs. The pillar itself can contain an active region such as a multiquantum-well structure that provides optical gain when pumped (Sec. 18.5A).

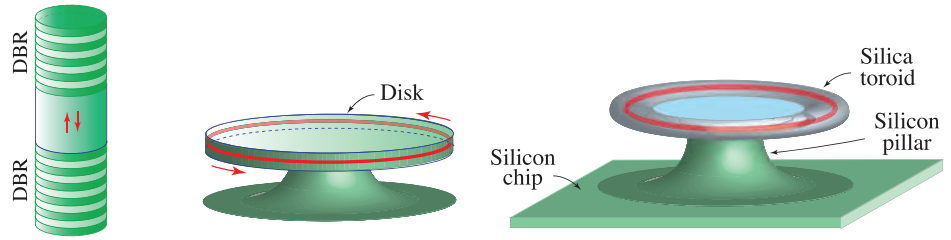


Figure 11.4-2 Micropillar, microdisk, and microtoroid resonators.

The **microdisk** resonator displayed in Fig. 11.4-2(b) is a circular resonator in which light travels at near-grazing incidence in whispering-gallery modes and is confined by total internal reflection from the circular boundary (see Sec. 11.3B). Micropillar and microdisk diameters usually range from 1 μm to tens of μm and their quality factors Q are substantially larger than those of mirror resonators since their losses are significantly lower (Table 11.4-1). Microdisk resonators fabricated from semiconductor materials are widely used as microdisk lasers because of their many salutary features (Sec. 18.5B).

The **microtoroid** dielectric microresonator illustrated in Fig. 11.4-2(c) is much like a fiber-ring resonator, in which the resonator modes are circulating guided waves. These microresonators are often fabricated from silica and are supported on a silicon chip by a silicon pillar. The toroid is formed by surface tension while the material is in a molten state; the outer boundary thus assumes a near atomic-scale surface finish and has significantly lower scattering losses than the microdisk resonator. Silica toroidal microresonators-on-a-chip exhibit exceptionally high values of the quality factor, $Q > 10^8$ (Table 11.4-1).

C. Microspheres

Dielectric **microspheres** are used as three-dimensional optical microresonators. Certain modes are guided along trajectories (orbits) that are tightly confined near a great circle of the sphere, resulting in whispering-gallery modes.

The modes of a dielectric sphere may be determined by solving the Helmholtz equation (5.3-16) for the electric and magnetic-field vectors, together with the appropriate boundary conditions. These modes are similar to the wavefunctions of an electron in a hydrogen atom (see Sec. 14.1A) because of the spherical symmetry of both problems, but there are also differences in view of the vector nature of the electromagnetic field.

The electric and magnetic vector fields are directly related to a scalar potential function U that satisfies the Helmholtz equation.[†] For a sphere of radius a and refractive

[†] For a detailed mathematical description, see, for example, A. N. Oraevsky, Whispering-Gallery Waves, *Kvantovaya Elektronika* (*Quantum Electronics*), vol. 32, pp. 377–400, 2002.

index n, located in air, the separation-of-variables method in a spherical coordinate system (r, θ, ϕ) results in a solution of the form

$$U(r,\theta,\phi) \propto \sqrt{r}\, J_{\ell+1/2}(nk_o r)\, P_{\rm m}^{\ell}(\cos\theta)\, \exp(\pm j{\rm m}\phi), \quad r \le a, \quad (11.4\text{-}1)$$

$$\propto \sqrt{r}\, {\rm H}_{\ell+1/2}^{(1)}(nk_o r)\, P_{\rm m}^{\ell}(\cos\theta)\, \exp(\pm j{\rm m}\phi), \quad r > a, \quad (11.4\text{-}2)$$

where $J_\ell(\cdot)$ is the Bessel function of the first kind of order ℓ, ${\rm H}_\ell^{(1)}(\cdot)$ is the Hankel function of the first kind of order ℓ, $P_{\rm m}^{\ell}(\cdot)$ is the adjoint Legendre function, and m and ℓ are nonnegative integers. The boundary conditions at $r = a$ yield a characteristic equation that provides a discrete set of values for k_o, corresponding to the resonance frequencies. These are indexed by a third integer n. In addition, there are two polarization modes — an E mode for which $H_r = 0$ and an H mode for which $E_r = 0$.

The modes are generally oscillatory functions of r, θ, and ϕ characterized by the radial, polar, and azimuthal mode numbers n, ℓ, and m, respectively. There are n maxima in the radial direction within the sphere. The number of field maxima in the azimuthal direction is 2ℓ, while the number of field maxima in the polar direction (between the two poles) is $\ell - {\rm m} + 1$.

The fundamental mode $({\rm n} = 1, {\rm m} = \ell)$ has a single peak in the radial direction within the sphere, and a single peak in the polar direction at $\theta = \pi/2$. For large ${\rm m} = \ell$, the modes are highly confined near the equator. This is because $P_\ell^\ell(\cos\theta) \approx \sin^\ell\theta$ vanishes rapidly at angles slightly different from $\theta = \pi/2$, and $J_\ell(nk_o r)$ is small everywhere within the sphere except for a sharp peak near $r = a$. The mode therefore represents an optical beam traveling along the equator, as shown in Fig. 11.4-3(a), much like the whispering-gallery modes of the disk resonator displayed in Fig. 11.3-3. For sufficiently large $\ell = {\rm m}$, the resonance frequencies of these modes are approximately equal to $\nu_\ell \approx \ell\, c/2\pi a$. This is to be expected since the angular mode number ℓ is close to the number of wavelengths that comprise the optical length of the equator.

The whispering-gallery mode may be viewed from a ray-optics perspective in terms of quasi-plane waves with wavevectors parallel to the local rays (see Sec. 2.3 and Fig. 10.2-8) that zigzag near the equator, as shown in Fig. 11.4-3(b). The wavevector $\mathbf{k}$ has magnitude $k = \sqrt{\ell(\ell+1)}/a$ and azimuthal component $k_\phi = {\rm m}/a$. The inclination angle of the zigzagging rays is smallest ($\approx 1/\sqrt{\ell}$) for the fundamental mode ${\rm m} = \ell$, while the ${\rm m} = 0$ mode has a 90° inclination.

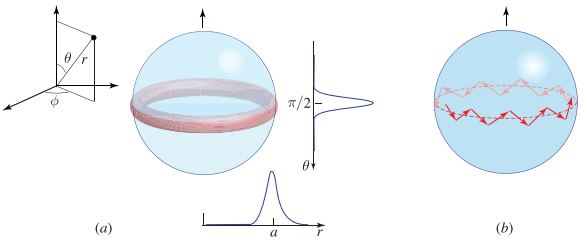


Figure 11.4-3 (a) Whispering-gallery mode in a microsphere resonator. (b) Ray model of the whispering-gallery mode.

Microspheres fabricated from low-loss fused silica have been used as optical resonators with ultrahigh values of Q. Like the toroidal resonator depicted in Fig. 11.4-2(c), the shape and surface finish of the sphere are determined by the surface tension in the molten state during fabrication; the result is near atomic perfection in the surface finish. The reduced surface scattering losses lead to remarkably high quality factors,

$Q > 10^{10}$ (see Table 11.4-1). Optical power may be coupled into the sphere via an optical fiber that is locally stripped of its cladding, as illustrated in Fig. 11.4-4.

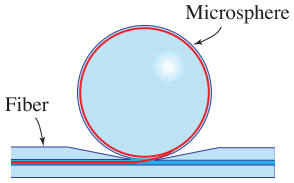


Figure 11.4-4 Coupling optical power from an optical fiber into a microsphere resonator.

D. Photonic-Crystal Microcavities

As described in Chapter 7, photonic crystals are periodic dielectric structures exhibiting photonic bandgaps, i.e., spectral bands within which light cannot propagate. The Bragg grating reflector (BGR) is an example of a 1D photonic crystal that serves as a reflector for frequencies within a photonic bandgap. The micropillar resonator shown in Fig. 11.4-2(a), for example, uses BGRs in lieu of mirrors. If the height of the microresonator equals one or just a few periods of the BGR, as illustrated in Fig. 11.4-5(a), the structure may also be regarded as an extended photonic crystal with the cavity acting as a defect in the crystal structure. The resonator is then called a **photonic-crystal resonator**.

This concept is also applicable to 2D photonic crystals. As schematized in Fig. 11.4-5(b), a *defect* in the 2D periodic crystal structure is a local alteration such as a missing hole in a periodic array of air holes drilled in a slab. For wavelengths that fall within the photonic-crystal bandgap, the periodic structure surrounding the defect does not support light propagation, so that light is trapped within the defect, much like electrons or holes are trapped by a defect in a semiconductor crystal. The defect then serves as a microcavity resonator. Stated differently, the defect produces new resonance frequencies that lie within the bandgap and correspond to optical modes that have spatial distributions centered within the microcavity and that decay rapidly in the surrounding photonic crystal.

Two-dimensional photonic crystals can be fabricated by using electron-beam lithography and reactive ion etching in semiconductor materials. Microcavities of dimensions close to a period of the photonic crystal, which can be of the order of a wavelength of light, can support modal volumes as small as λ^3. Hence, photonic-crystal microcavities have the smallest normalized modal volumes of the family of microresonators (see Table 11.4-1). The quality factors Q can be as high as 10^4. Because of these features, photonic-crystal resonators are often co-opted for use in photonic-crystal microcavity lasers (Sec. 18.5C).

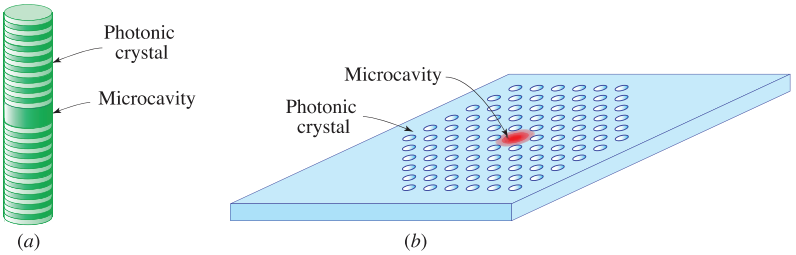


Figure 11.4-5 Photonic-crystal microresonators. (a) The micropillar resonator as a 1D photonic crystal in which the microcavity acts as a defect. (b) A 2D photonic-crystal resonator may be fabricated by drilling holes in a dielectric slab at the points of a planar hexagonal lattice; a missing hole serves as the microcavity.

E. Plasmonic Resonators: Metallic Nanodisks and Nanospheres

Metallic electromagnetic resonators are routinely used at microwave and radio frequencies. As illustrated in Fig. 11.0-2, microwave cavity resonators have dimensions that are close to the resonance wavelength (cm), whereas radio-frequency resonant circuits, comprising metallic capacitors and inductors, have dimensions (cm) much smaller than the resonance wavelength (m).

Plasmonic resonators contain metallic structures of subwavelength dimensions that operate at optical frequencies by supporting surface plasmon polariton (SPP) waves or localized surface plasmon (LSP) oscillations at their boundaries with dielectric media (Secs. 8.2B and 8.2C, respectively). They can take the form of nanodisks, nanospheres, or other nanoparticles. Their dimensions ($\sim$10 nm) can be much smaller than the resonance wavelength ($\sim\mu$m), while maintaining a size-to-wavelength ratio ($\sim 10^{-2}$) not unlike that of the radio-frequency electronic resonator portrayed in Fig. 11.0-2.

The mathematical modeling of structures whose dimensions are much larger than the resonance wavelength, such as resonators with large mirrors, is readily achieved by making use of electromagnetic optics. The electric and magnetic fields are then of paramount importance, and these are the quantities that we customarily encounter in the optics literature. At the opposite extreme, metallic structures whose dimensions are much smaller than the resonance wavelength can be well-modeled in terms of electrical voltages and currents. The key elements in this domain are lumped electrical components such as inductors and capacitors, as encountered in the electrical-engineering literature — as a simple example, the resonance frequency of the electronic resonator portrayed in Fig. 11.0-2 is $\omega_0 = 1/\sqrt{LC}$. Metallic structures whose dimensions are comparable to the resonance wavelength pose the greatest challenge in terms of modeling because the analysis must then generally be carried out in terms of voltages and currents as well as fields.

Metallic Nanodisk

A metallic cylinder with an interior dielectric material supports whispering-gallery *photonic* modes of light that reflect at the metallic boundary, as displayed in Fig. 11.4-6(a). This disk resonator also supports *plasmonic* modes in the form of SPP waves (Sec. 8.2B) at the interior boundary, as illustrated in Fig. 11.4-6(b). The internal optical field of the plasmonic mode is more tightly confined near the boundary than is the field of the photonic mode. Moreover, since the plasmonic wave penetrates more deeply into the metal than does the photonic mode, it suffers greater losses and the resonator exhibits a smaller quality factor Q. By comparison, a dielectric disk resonator (Sec. 11.3B) supports whispering-gallery modes with evanescent optical fields that extend into the exterior dielectric medium, as shown in Fig. 11.4-6(c).

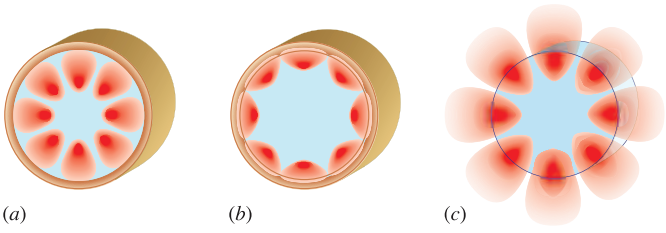


Figure 11.4-6 Schematic of optical-field distributions in disk resonators. (a) Photonic mode in a metal–dielectric disk: light is confined to the interior by multiple reflections at the metallic boundary. (b) Plasmonic mode in a metal–dielectric disk: a SPP wave travels along the interior boundary. (c) Photonic mode in a dielectric–dielectric disk: light is confined by total internal reflection at the boundary.

Nominal values for the normalized modal volume and quality factor of a plasmonic mode in a metal–dielectric disk resonator with a diameter of 100 nm are $V/\lambda^3 \sim 10^{-4}$ and $Q \sim 10$, respectively.[†] These values are substantially below those for photonic-crystal microcavities (see Table 11.4-1). Nevertheless, structures of this kind find use in surface plasmon polariton disk and ring nanolasers (Sec. 18.6).

Metallic Nanosphere

As discussed in Sec. 8.2C, a metallic nanosphere embedded in a dielectric medium supports resonant localized surface plasmon (LSP) oscillations. When the nanosphere is illuminated by an optical wave, the scattered field, as well as the internal field, are substantially enhanced at frequencies near resonance. This field enhancement is accompanied by spatial localization of energy at the nanoscale, so that the nanosphere serves as a nanoresonator. Since metals are relatively lossy, however, the resonator quality factor Q is far lower than that of dielectric resonators. Nevertheless, a metal nanosphere embedded in a specialized dielectric medium can serve as a localized surface plasmon nanolaser (Sec. 18.6).

READING LIST

Books

See also the reading list on lasers in Chapter 16.

A. H. W. Choi, ed., *Handbook of Optical Microcavities*, CRC Press/Taylor & Francis, 2015.

J. Heebner, R. Grover, and T. A. Ibrahim, *Optical Microresonators: Theory, Fabrication, and Applications*, Springer-Verlag, 2008, paperback ed. 2010.

A. V. Kavokin, J. J. Baumberg, G. Malpuech, and F. P Laussy, *Microcavities*, Oxford University Press, 2007.

D. G. Rabus, *Integrated Ring Resonators: The Compendium*, Springer-Verlag, 2007.

N. Hodgson and H. Weber, *Laser Resonators and Beam Propagation: Fundamentals, Advanced Concepts and Applications*, Springer-Verlag, 2nd ed. 2005.

K. J. Vahala, ed., *Optical Microcavities*, World Scientific, 2004.

K. Staliūnas and V. J. Sánchez-Morcillo, *Transverse Patterns in Nonlinear Optical Resonators*, Springer-Verlag, 2003.

A. N. Oraevsky, *Gaussian Beams and Optical Resonators*, Nova, 1996.

Yu. A. Anan'ev, *Laser Resonators and the Beam Divergence Problem*, CRC/Taylor & Francis, 1992.

J. M. Vaughan, *The Fabry–Perot Interferometer*, CRC Press, 1989.

G. Hernandez, *Fabry–Perot Interferometers*, Cambridge University Press, 1986, paperback ed. 1988.

A. E. Siegman, *Lasers*, University Science, 1986.

L. A. Vaynshteyn, *Open Resonators and Open Waveguides*, Golem Press, 1969.

Seminal Articles

K. J. Vahala, Optical Microcavities, *Nature*, vol. 424, pp. 839–846, 2003.

J. U. Nöckel and A. D. Stone, Ray and Wave Chaos in Asymmetric Resonant Optical Cavities, *Nature*, vol. 385, pp. 45–47, 1997.

H. Kogelnik and T. Li, Laser Beams and Resonators, *Applied Optics*, vol. 5, pp. 1550–1567, 1966 (published simultaneously in *Proceedings of the IEEE*, vol. 54, pp. 1312–1329, 1966).

A. E. Siegman, Unstable Optical Resonators for Laser Applications, *Proceedings of the IEEE*, vol. 53, pp. 277–287, 1965.

A. G. Fox and T. Li, Resonant Modes in a Maser Interferometer, *Bell System Technical Journal*, vol. 40, pp. 453–488, 1961.

G. D. Boyd and J. P. Gordon, Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, *Bell System Technical Journal*, vol. 40, pp. 489–508, 1961.

[†] See, e.g., M. Kuttge, F. Javier García de Abajo, and A. Polman, Ultrasmall Mode Volume Plasmonic Nanodisk Resonators, *Nano Letters*, vol. 10, pp. 1537–1541, 2010.

PROBLEMS

11.1-3 **Resonance Frequencies of a Resonator with an Etalon.**

(a) Determine the spacing between adjacent resonance frequencies in a resonator constructed of two parallel planar mirrors separated by a distance $d = 15$ cm in air $(n = 1)$.

(b) A transparent plate of thickness $d_1 = 2.5$ cm and refractive index $n = 1.5$ is placed inside the resonator and is tilted slightly to prevent light reflected from the plate from reaching the mirrors. Determine the spacing between the resonance frequencies of the resonator.

11.1-4 **Mirrorless Resonators.** Laser diodes are often fabricated from crystals whose surfaces are cleaved along crystal planes. These surfaces act as reflectors by virtue of Fresnel reflection, and therefore serve as the mirrors of a Fabry–Perot resonator. An expression for the power reflectance is provided in (6.2-15). Consider a crystal placed in air $(n = 1)$ whose refractive index $n = 3.6$ and loss coefficient $\alpha_s = 1$ cm^{-1}. The light reflects between two parallel surfaces separated by a distance $d = 0.2$ mm. Determine the spacing between resonance frequencies ν_F, the overall distributed loss coefficient α_r, the finesse $\mathcal{F}$, the spectral width of a mode $\delta\nu$, and the quality factor Q. If the free-space wavelength of the generated light is $\lambda_o = 1.55$ μm, estimate the longitudinal mode number q.

11.1-5 **Fabry–Perot Etalon with Bragg Grating Reflectors.** A Fabry–Perot etalon is made by sandwiching a layer of GaAs between two of the GaAs/AlAs Bragg grating reflectors described in Prob. 7.1-8. Determine the finesse $\mathcal{F}$ of the resonator and quality factor Q. Determine the transmittance of a Bragg grating reflector comprised of $N = 10$ alternating layers of GaAs ($n_1 = 3.6$) and AlAs ($n_2 = 3.2$) of widths d_1 and d_2 equal to a quarter wavelength in each medium. Assume that the light is incident from an extended GaAs medium.

11.1-6 **Optical Energy Decay Time.** How much time does it take for the optical energy stored in a resonator of finesse $\mathcal{F} = 100$, length $d = 50$ cm, and refractive index $n = 1$, to decay to one-half of its initial value?

11.2-5 **Stability of Spherical-Mirror Resonators.**

(a) Can a resonator with two convex mirrors ever be stable?

(b) Can a resonator with one convex and one concave mirror ever be stable?

11.2-6 **A Planar-Mirror Resonator Containing a Lens.** A lens of focal length f is placed inside a planar-mirror resonator constructed of two flat mirrors separated by a distance d. The lens is located at a distance $d/2$ from each of the mirrors.

(a) Determine the ray-transfer matrix for a ray that begins at one of the mirrors and travels a round trip inside the resonator.

(b) Determine the condition of stability of the resonator.

(c) Under stable conditions sketch the Gaussian beam that fits this resonator.

11.2-7 **Self-Reproducing Rays in a Symmetric Resonator.** Consider a symmetric resonator using two concave mirrors of radii R separated by a distance $d = 3|R|/2$. After how many round trips through the resonator will a ray retrace its path?

11.2-8 **Ray Position in Unstable Resonators.** Show that for an unstable resonator the ray position after m round trips is given by $y_m = \alpha_1 h_1^m + \alpha_2 h_2^m$, where α_1 and α_2 are constants. Here $h_1 = b + \sqrt{b^2 - 1}$, $h_2 = b - \sqrt{b^2 - 1}$, and $b = 2(1 + d/R_1)(1 + d/R_2) - 1$. *Hint:* Use the results in Sec. 1.4D.

11.2-9 **Ray Position in Unstable Symmetric Resonators.** Verify that a symmetric resonator using two concave mirrors of radii $R = -30$ cm separated by a distance $d = 65$ cm is unstable. Find the position y_1 of a ray that begins at one of the mirrors, at position $y_0 = 0$ with an angle $\theta_0 = 0.1°$, and undergoes one round trip. If the mirrors have 5-cm-diameter apertures, after how many round trips does the ray leave the resonator? Plot y_m, $m = 2, 3, \ldots$, for $d = 50$ cm and $d = 65$ cm. You may use the results of Prob. 11.2-8.

11.2-10 **Gaussian-Beam Standing Waves.** Consider a wave formed by the sum of two identical Gaussian beams propagating in the $+z$ and $-z$ directions. Show that the result is a standing wave. Using the boundary conditions at two ideal mirrors placed such that they coincide with the wavefronts, derive the resonance frequencies (11.2-30).

11.2-11 **Gaussian Beam in a Symmetric Confocal Resonator.** A symmetric confocal resonator with mirror spacing $d = 16$ cm, mirror reflectances 0.995, and $n = 1$ is used in a laser operating at $\lambda_o = 1$ μm.

(a) Find the radii of curvature of the mirrors.

(b) Find the waist of the $(0, 0)$ (Gaussian) mode.

(c) Sketch the intensity distribution of the $(1, 0)$ modes at one of the mirrors and determine the distance between its two peaks.

(d) Determine the resonance frequencies of the $(0, 0)$ and $(1, 0)$ modes.

(e) Assuming that losses arise only from imperfect mirror reflectances, determine the distributed resonator loss coefficient α_r.

*11.2-12 **Diffraction Loss in a Symmetric Confocal Resonator.** The percent diffraction loss per pass for the different low-order modes of a symmetric confocal resonator is given in Fig. 11.2-11, as a function of the Fresnel number $N_{\mathrm{F}} = a^2/\lambda d$ (where d is the mirror spacing and a is the radius of the mirror aperture). Using the parameters provided in Prob. 11.2-11, determine the mirror radius for which the loss per pass of the $(1, 0)$ mode is 1%.

11.3-2 **Number of Modes in Resonators of Different Dimensions.** Consider light of wavelength $\lambda_o = 1.06$ μm and spectral width $\Delta\nu = 120$ GHz. How many modes have frequencies within this linewidth in the following resonators $(n = 1)$:

(a) A one-dimensional resonator of length $d = 10$ cm?

(b) A 10 cm $\times$ 10 cm two-dimensional resonator?

(c) A 10 cm $\times$ 10 cm $\times$ 10 cm three-dimensional resonator?

CHAPTER
12

STATISTICAL OPTICS

12.1 STATISTICAL PROPERTIES OF RANDOM LIGHT 475
- A. Optical Intensity
- B. Temporal Coherence and Spectrum
- C. Spatial Coherence
- D. Longitudinal Coherence

12.2 INTERFERENCE OF PARTIALLY COHERENT LIGHT 489
- A. Interference of Two Partially Coherent Waves
- B. Interferometry and Temporal Coherence
- C. Interferometry and Spatial Coherence

*12.3 TRANSMISSION OF PARTIALLY COHERENT LIGHT 497
- A. Propagation of Partially Coherent Light
- B. Image Formation with Incoherent Light
- C. Gain of Spatial Coherence by Propagation

12.4 PARTIAL POLARIZATION 506

Max Born (1882–1970)

Emil Wolf (1922–2018)

The book *Principles of Optics*, first published in 1959 by Max Born and Emil Wolf, drew attention to the importance of coherence in optics. Emil Wolf is responsible for many advances in the theory of optical coherence.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Statistical optics is the study of the properties of random light. Randomness in light arises because of unpredictable fluctuations in the source of light itself or in the medium through which it propagates. Natural light, such as that radiated by a hot object (e.g., the sun) is random in time because it comprises a superposition of emissions from a very large number of atoms that radiate independently, and with different frequencies and phases. Randomness in light can also arise as a result of scattering from a rough surface, or transmission through a ground-glass diffuser or a turbulent fluid, which impart random spatial variations to the optical wavefront. The study of the random fluctuations of light is also known as the **theory of optical coherence**.

In earlier chapters it was assumed that light is deterministic or "coherent." As discussed in Sec. 2.2, an example of coherent light is the monochromatic wavefunction $u(\mathbf{r},t) = \text{Re}\{U(\mathbf{r})\exp(j2\pi\nu t)\}$, for which the complex amplitude $U(\mathbf{r})$ is a deterministic complex function, e.g., $U(\mathbf{r}) = (A_0/r)\exp(-jkr)$ for a spherical wave. The dependence of the wavefunction on time and position is then perfectly periodic and predictable [Fig. 12.0-1(a)]. For random light, in contrast, the dependence of the wavefunction on time and position is not totally predictable and generally requires statistical methods for its characterization [Fig. 12.0-1(b)].

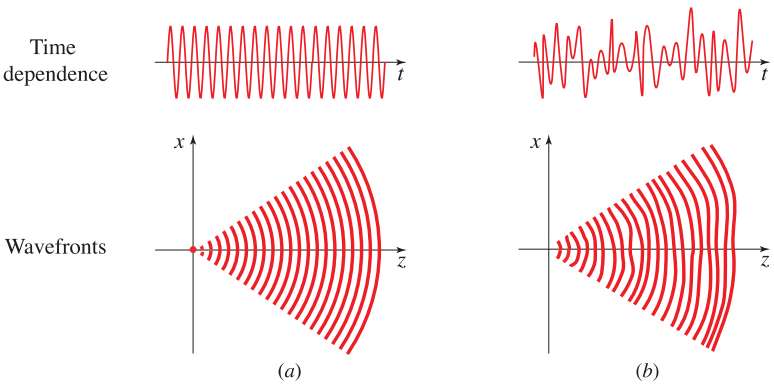


Figure 12.0-1 Time dependence and wavefronts of (a) a monochromatic spherical wave, which is an example of coherent light; (b) random light.

How can we go about extracting meaningful measures from the fluctuations of a random optical wave that will permit us to characterize it and distinguish it from other random waves? Examine, for instance, the three random optical waves displayed in Fig. 12.0-2, whose wavefunctions at some position vary with time as shown. A good beginning is to observe that wave (b) is more "intense" than wave (a), and that the envelope of wave (c) fluctuates "faster" than the envelopes of the other two waves.

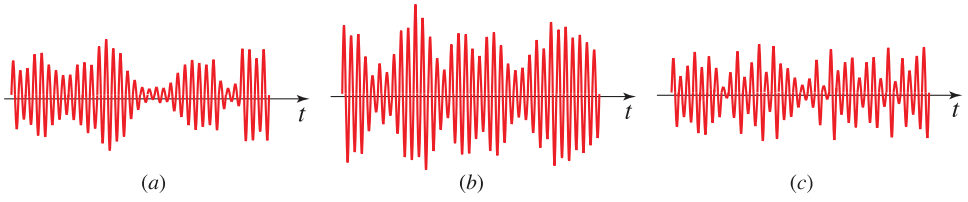


Figure 12.0-2 Time dependence of the wavefunctions of three random waves.

To translate these casual qualitative observations into quantitative measures, we use the concept of statistical averaging to define a number of (nonrandom) measures of a wave. Because the random function $u(\mathbf{r}, t)$ satisfies certain laws (the wave equation and boundary conditions) its statistical averages must also satisfy certain laws. The theory of optical coherence is concerned with the definitions of these statistical averages, the laws that govern them, and the measures by means of which light is classified as **coherent**, **incoherent**, or, in general, **partially coherent**.

This Chapter

This chapter serves as an introduction to the theory of optical coherence. Familiarity with the theory of random fields (random functions of many variables — space and time) is useful for fully understanding the theory of partial coherence. However, the notions presented in this chapter are limited in scope, so that knowledge of the concept of statistical averaging suffices.

In Sec. 12.1 we define two statistical averages used to describe random light: the optical intensity and the mutual coherence function. Temporal and spatial coherence are delineated, and the connection between temporal coherence and monochromaticity is established. The examples of partially coherent light provided in Sec. 12.1 reveal that spatially coherent light need not be temporally coherent, and that monochromatic light need not be spatially coherent. One of the basic manifestations of the coherence of light is its ability to produce visible interference fringes. Section 12.2 is devoted to the laws of interference obeyed by random light. The propagation of partially coherent light in free space, and its transmission through various optical systems (including image-formation systems) is the subject of Sec. 12.3. The theory of polarization of random light (partial polarization), which relies on statistical averaging of the components of the optical field vector, is introduced in Sec. 12.4. With the exception of this latter section, our exposition is framed in the context of scalar wave optics.

12.1 STATISTICAL PROPERTIES OF RANDOM LIGHT

An arbitrary optical wave is described by a wavefunction $u(\mathbf{r}, t) = \text{Re}\{U(\mathbf{r}, t)\}$, where $U(\mathbf{r}, t)$ is the complex wavefunction. For example, $U(\mathbf{r}, t)$ may take the form $U(\mathbf{r}) \exp(j2\pi\nu t)$ for monochromatic light, or it may comprise a sum of such functions with many different values of ν for polychromatic light (see Sec. 2.6A for a discussion of the complex wavefunction). For random light, both functions, $u(\mathbf{r}, t)$ and $U(\mathbf{r}, t)$, are random and are characterized by a number of statistical averages that are introduced in this section.

A. Optical Intensity

As discussed in Secs. 2.2A and 2.6A, the intensity $I(\mathbf{r}, t)$ of coherent (deterministic) light is related to the absolute square of the complex wavefunction $U(\mathbf{r}, t)$ via

$$I(\mathbf{r}, t) = |U(\mathbf{r}, t)|^2 . \tag{12.1-1}$$

For pulsed light, the intensity is time varying whereas for monochromatic deterministic light it is independent of time.

For random light, $U(\mathbf{r}, t)$ is a random function of time and position. The intensity $|U(\mathbf{r}, t)|^2$ is therefore also random. The **average intensity** is then defined as

$$\boxed{I(\mathbf{r}, t) = \langle |U(\mathbf{r}, t)|^2 \rangle ,} \tag{12.1-2}$$

Average Intensity

where the symbol $\langle \cdot \rangle$ denotes an ensemble average over many realizations of the random function. Thus, although a random wave that is repeatedly reproduced under the same conditions leads to a different wavefunction on each trial, the average intensity at each time and position is determined by making use of (12.1-2). We denote $I(\mathbf{r}, t)$ the *intensity* of the light (with the modifier *average* implied), when there is no ambiguity in meaning. The unaveraged quantity $|U(\mathbf{r}, t)|^2$, in contrast, is called the **random intensity** or **instantaneous intensity**. For deterministic light, the averaging operation is superfluous since all trials produce exactly the same wavefunction, whereupon (12.1-2) is equivalent to (12.1-1).

The average intensity may be time independent or it may be a function of time, as illustrated in Figs. 12.1-1(a) and (b), respectively. The former case applies when the optical wave is statistically **stationary**, i.e., when its statistical averages are invariant to time. The instantaneous intensity $|U(\mathbf{r}, t)|^2$ then fluctuates randomly with time, but its average is constant. We denote the average intensity in this case by $I(\mathbf{r})$. It is clear that stationarity does not necessarily mean constancy; rather it signifies constancy of the average properties. An example of stationary random light is that emitted by an ordinary incandescent lamp whose filament is heated by a constant electric current. The average intensity $I(\mathbf{r})$ is then a function of distance from the lamp, but it does not vary with time. On the other hand, the random intensity $|U(\mathbf{r}, t)|^2$ fluctuates with both position and time, as illustrated in Fig. 12.1-1(a).

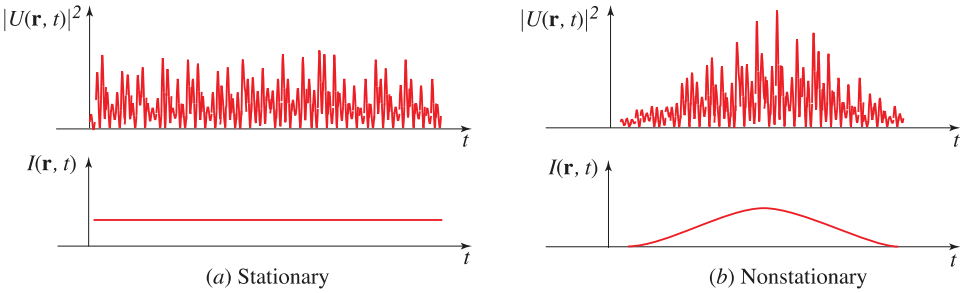


Figure 12.1-1 (a) A statistically stationary wave has an average intensity $I(\mathbf{r})$ that does not vary with time. (b) A statistically nonstationary wave has an average intensity $I(\mathbf{r}, t)$ that varies with time. These plots represent, for example, the intensity of light produced by an incandescent lamp driven by (a) a constant electric current, and (b) a pulse of electric current.

When the light is stationary, the statistical averaging operation over many realizations of the wave, as prescribed by (12.1-2), is usually equivalent to time averaging over a long time duration, which is written as

$$I(\mathbf{r}) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |U(\mathbf{r}, t)|^2 \, dt. \tag{12.1-3}$$

B. Temporal Coherence and Spectrum

Consider the fluctuations of *stationary* light at a fixed position $\mathbf{r}$, as a function of time. The stationary random function $U(\mathbf{r}, t)$ has a constant intensity $I(\mathbf{r}) = \langle |U(\mathbf{r}, t)|^2 \rangle$. For brevity, we need not explicitly indicate the $\mathbf{r}$ dependence since $\mathbf{r}$ is fixed, so that we may write $U(\mathbf{r}, t) = U(t)$ and $I(\mathbf{r}) = I$.

The random fluctuations of $U(t)$ are characterized by a time scale representing the "memory" of the random function. After this time, the process "forgets" itself, so that

fluctuations at points separated by a time interval longer than this memory time are independent. The function appears to be relatively smooth within its memory time, but "rough" or "erratic" when examined over longer time scales as portrayed in Fig. 12.0-2. A quantitative measure of this temporal behavior is established by defining a statistical *average* called the autocorrelation function. This function describes the extent to which the wavefunction fluctuates in unison at two instants of time separated by a given time delay, and thus serves to establish the time scale of the process that characterizes the wavefunction.

Temporal Coherence Function

The **autocorrelation function** of a stationary complex random function $U(t)$ is defined as the average of the product of $U^*(t)$ and $U(t+\tau)$, as a function of the time delay τ,

$$\boxed{G(\tau) = \langle U^*(t)\, U(t+\tau)\rangle} \qquad \text{(12.1-4)}$$

Temporal Coherence Function

or

$$G(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} U^*(t)\, U(t+\tau)\, dt \qquad \text{(12.1-5)}$$

(see Sec. A.1 in Appendix A).

To understand the significance of the definition in (12.1-4), consider the case in which the average value of the complex wavefunction $\langle U(t)\rangle = 0$. This is applicable when the phase of the phasor $U(t)$ is equally likely to have any value between 0 and 2π, as illustrated in Fig. 12.1-2. The phase of the product $U^*(t)U(t+\tau)$ is the angle between phasors $U(t)$ and $U(t+\tau)$. If these two quantities are uncorrelated, the angle between their phasors varies randomly between 0 and 2π. The phasor $U^*(t)U(t+\tau)$ then has a totally uncertain angle, so that it is equally likely to take any direction, making its average, the autocorrelation function $G(\tau)$, vanish. On the other hand if, for a given value of τ, $U(t)$ and $U(t+\tau)$ are correlated, their phasors will maintain some relationship. Their fluctuations are then linked together so that the product phasor $U^*(t)U(t+\tau)$ will have a preferred direction and its average $G(\tau)$ will not vanish.

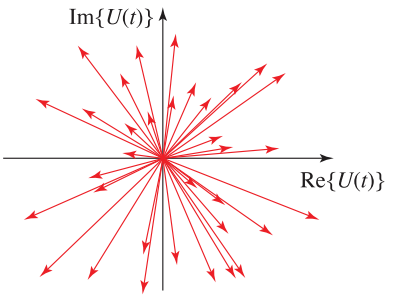


Figure 12.1-2 Variation of the phasor $U(t)$ with time when its argument is uniformly distributed between 0 and 2π. The average values of its real and imaginary parts are zero, so that $\langle U(t)\rangle = 0$.

In the language of optical coherence theory, the autocorrelation function $G(\tau)$ is known as the **temporal coherence function**. It is easy to show that $G(\tau)$ is a function with Hermitian symmetry, $G(-\tau) = G^*(\tau)$, and that the intensity I, defined by (12.1-2), is equal to $G(\tau)$ when $\tau = 0$:

$$I = G(0). \qquad \text{(12.1-6)}$$

Degree of Temporal Coherence

The temporal coherence function $G(\tau)$ carries information about both the intensity $I = G(0)$ and the degree of correlation (coherence) of stationary light. A measure of coherence that is insensitive to the intensity is provided by the normalized autocorrelation function,

$$g(\tau) = \frac{G(\tau)}{G(0)} = \frac{\langle U^*(t)U(t+\tau)\rangle}{\langle U^*(t)U(t)\rangle}, \tag{12.1-7}$$

Complex Degree of Temporal Coherence

which is called the **complex degree of temporal coherence**. Its absolute value cannot exceed unity,

$$0 \le |g(\tau)| \le 1. \tag{12.1-8}$$

The value of $|g(\tau)|$ is a measure of the degree of correlation between $U(t)$ and $U(t+\tau)$. When the light is deterministic and monochromatic, i.e., when $U(t) = A_0 \exp(j2\pi\nu_0 t)$ where A_0 is a constant, (12.1-7) yields

$$g(\tau) = \exp(j2\pi\nu_0\tau), \tag{12.1-9}$$

so that $|g(\tau)| = 1$ for all τ. The variables $U(t)$ and $U(t+\tau)$ are then completely correlated for all time delays τ. For most sources of light, $|g(\tau)|$ decreases from its maximum value $|g(0)| = 1$ as τ increases, and the fluctuations become uncorrelated for sufficiently large τ.

Coherence Time

If $|g(\tau)|$ decreases monotonically with time delay, the value τ_c at which it drops to a prescribed value ($1/2$ or $1/e$, for example) serves as a measure of the memory time of the fluctuations. The quantity τ_c is known as the **coherence time** (see Fig. 12.1-3).

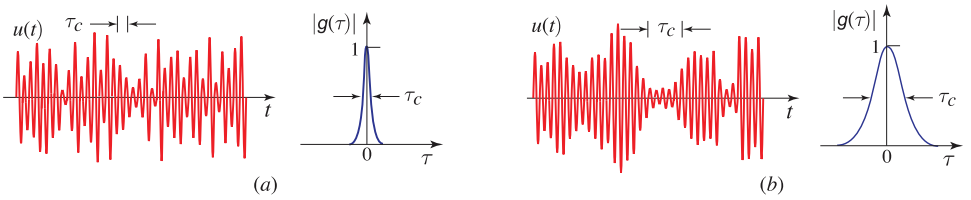


Figure 12.1-3 Illustrations of the wavefunction, magnitude of the complex degree of temporal coherence $|g(\tau)|$, and coherence time τ_c for an optical field with (a) short coherence time and (b) long coherence time. The amplitudes and phases of the wavefunctions vary randomly with time constants approximately equal to τ_c (which is greater than the duration of an optical cycle). Within the coherence time, the wave is rather predictable and can be approximated by a sinusoid. However, given the amplitude and phase of the wave at a particular time, the amplitude and phase at times beyond the coherence time cannot be predicted.

For $\tau < \tau_c$ the fluctuations are "strongly" correlated whereas for $\tau > \tau_c$ they are "weakly" correlated. In general, τ_c is the width of the function $|g(\tau)|$. Although the

definition of the width of a function can take many forms (see Sec. A.2 of Appendix A), the power-equivalent width is commonly used as the definition of coherence time:

$$\tau_c = \int_{-\infty}^{\infty} |g(\tau)|^2 \, d\tau \tag{12.1-10}$$

Coherence Time

[see (A.2-8) and note that $g(0) = 1$].

EXERCISE 12.1-1

Coherence Time. Verify that the following expressions for the complex degree of temporal coherence are consistent with the definition of τ_c given in (12.1-10):

$$g(\tau) = \begin{cases} \exp\left(-\dfrac{|\tau|}{\tau_c}\right) & \text{(exponential)} \\ \exp\left(-\dfrac{\pi\tau^2}{2\tau_c^2}\right). & \text{(Gaussian)} \end{cases} \tag{12.1-11}$$

By what factor does $|g(\tau)|$ drop as τ increases from 0 to τ_c in each case?

The coherence time of monochromatic light is infinite since $|g(\tau)| = 1$ everywhere. Light for which the coherence time τ_c is much longer than differences of the time delays encountered in an optical system of interest is effectively completely coherent. Thus, light is effectively coherent if the distance $c\tau_c$ is much greater than all optical pathlength differences encountered. This distance is known as the **coherence length**:

$$l_c = c\tau_c \tag{12.1-12}$$

Coherence Length

Power Spectral Density

To determine the *average* spectrum of random light, we carry out a Fourier decomposition of the random function $U(t)$. The amplitude of the component with frequency ν is the Fourier transform (see Appendix A)

$$V(\nu) = \int_{-\infty}^{\infty} U(t) \exp(-j2\pi\nu t) \, dt. \tag{12.1-13}$$

The average energy per unit area of those components with frequencies in the interval between ν and $\nu + d\nu$ is $\langle |V(\nu)|^2 \rangle \, d\nu$, so that $\langle |V(\nu)|^2 \rangle$ represents the energy spectral density of the light (energy per unit area per unit frequency). Note that the complex wave function $U(t)$ has been defined so that $V(\nu) = 0$ for negative ν (see Sec. 2.6A).

Since a truly stationary function $U(t)$ is eternal and carries infinite energy, we consider instead the *power* spectral density. We first determine the energy spectral density of the function $U(t)$ observed over a window of time width T by finding the truncated Fourier transform

$$V_T(\nu) = \int_{-T/2}^{T/2} U(t) \exp(-j2\pi\nu t) \, dt \tag{12.1-14}$$

and we then determine the energy spectral density $\langle |V_T(\nu)|^2 \rangle$. The power spectral density is the energy per unit time $(1/T)\langle |V_T(\nu)|^2 \rangle$. We can now extend the time window to infinity by taking the limit $T \to \infty$, which yields the **power spectral density**:

$$S(\nu) = \lim_{T \to \infty} \frac{1}{T} \langle |V_T(\nu)|^2 \rangle; \tag{12.1-15}$$

$S(\nu)$ is nonzero only for positive frequencies. Because $U(t)$ was defined such that $|U(t)|^2$ represents power per unit area, or intensity (W/cm^2), $S(\nu)\,d\nu$ represents the average power per unit area carried by frequencies between ν and $\nu + d\nu$, so that $S(\nu)$ actually represents the **intensity spectral density** (W/cm^2-Hz). It is often referred to simply as the **spectral density** or the **spectrum**. The total average intensity is the integral

$$I = \int_0^\infty S(\nu)\,d\nu. \tag{12.1-16}$$

The autocorrelation function $G(\tau)$, defined by (12.1-4), and the spectral density $S(\nu)$ defined by (12.1-15) can be shown to form a Fourier transform pair (see Prob. 12.1-5),

$$\boxed{S(\nu) = \int_{-\infty}^{\infty} G(\tau) \exp(-j2\pi\nu\tau)\,d\tau.} \tag{12.1-17}$$

Power Spectral Density

This relation is known as the **Wiener–Khinchin theorem**.

An optical wave representing a color image, such as that illustrated in Fig. 12.1-4, has a spectrum that varies with position $\mathbf{r}$; each spectral profile shown corresponds to a perceived color.

Figure 12.1-4 Spectral densities, plotted as a function of wavelength, at three locations in a color image (Georgia O'Keeffe, *Red Canna*, 1919, High Museum of Art, Atlanta).

Spectral Width

The spectrum of light is often confined to a narrow band centered about a central frequency ν_0. The **spectral width**, or **linewidth**, of light is the width $\Delta\nu$ of the spectral density $S(\nu)$. Because of the Fourier-transform relation between $S(\nu)$ and $G(\tau)$, their widths are inversely related. As illustrated in Fig. 12.1-5, a light source of broad spectral width has a short coherence time, whereas a light source of narrow spectral width has a long coherence time. In the limiting case of monochromatic light, $G(\tau) = I\exp(j2\pi\nu_0\tau)$, so that the corresponding intensity spectral density $S(\nu) = I\delta(\nu-\nu_0)$ contains only a single frequency component ν_0, in which case $\tau_c = \infty$ and $\Delta\nu = 0$. The coherence time of a source of light can be increased by passing it through an optical filter to reduce its spectral width. The resultant gain of coherence comes at the expense of a reduction of its intensity.

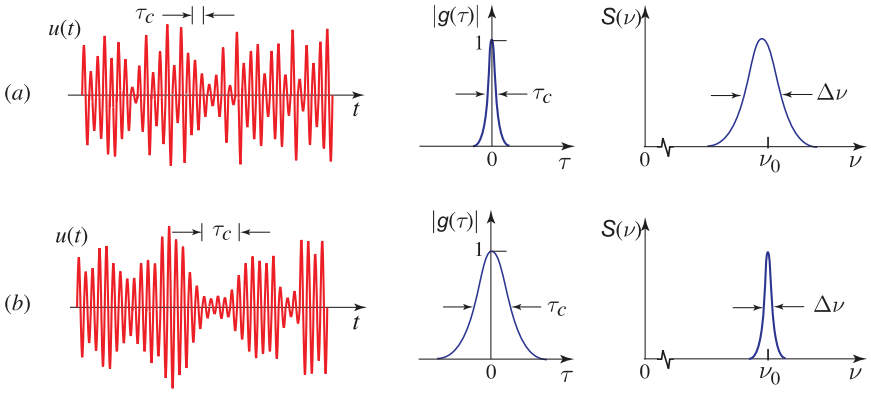


Figure 12.1-5 Two random waves together with the magnitudes of their complex degree of temporal coherence and their spectral densities. The widths of $S(\nu)$ and $|g(\tau)|$ are inversely related.

There are several definitions for spectral width (see Appendix A, Sec. A.2). The most common is the full-width at half-maximum (FWHM) of the function $S(\nu)$, which we denote $\Delta\nu_{\text{FWHM}} \equiv \Delta\nu$. The relation between $\Delta\nu_{\text{FWHM}}$ and the coherence time τ_c depends on the spectral profile of the source, as indicated in Table 12.1-1.

Table 12.1-1 Relation between spectral width $\Delta\nu_{\text{FWHM}}$ and coherence time τ_c for light with several different spectral profiles.

Spectral Profile	Rectangular	Lorentzian	Gaussian
Spectral Width $\Delta\nu_{\text{FWHM}}$	$\dfrac{1}{\tau_c}$	$\dfrac{1}{\pi\tau_c} \approx \dfrac{0.32}{\tau_c}$	$\dfrac{\sqrt{2\ln 2/\pi}}{\tau_c} \approx \dfrac{0.66}{\tau_c}$

An alternative convenient definition of the spectral width is

$$\Delta\nu_c = \frac{\left(\int_0^\infty S(\nu)\,d\nu\right)^2}{\int_0^\infty S^2(\nu)\,d\nu}. \tag{12.1-18}$$

Using the definition provided in (12.1-18) it can be shown that

$$\Delta\nu_c = \frac{1}{\tau_c}, \tag{12.1-19}$$

Spectral Width

regardless of the spectral profile (see Exercise 12.1-2). As an example, if $S(\nu)$ is a rectangular function extending over a frequency interval from $\nu_0 - B/2$ to $\nu_0 + B/2$, then (12.1-18) yields $\Delta\nu_c = B$. For this profile, the coherence time $\tau_c = 1/B$, so that (12.1-19) is obeyed. The two definitions of bandwidth, $\Delta\nu_c$ and $\Delta\nu_{\text{FWHM}}$, differ by a factor that ranges from 0.32 to 1 for the spectral profiles presented in Table 12.1-1.

EXERCISE 12.1-2

Relation Between Spectral Width and Coherence Time. Show that the coherence time τ_c defined in (12.1-10) is related to the spectral width $\Delta\nu_c$ defined in (12.1-18) by the simple inverse relation $\tau_c = 1/\Delta\nu_c$. *Hint:* Use the definitions of $\Delta\nu_c$ and τ_c, the Fourier-transform relation between $S(\nu)$ and $G(\tau)$, and Parseval's theorem provided in (A.1-7) [Appendix A].

Representative spectral widths for several different sources of light, along with their associated coherence times and coherence lengths, are provided in Table 12.1-2.

Table 12.1-2 Spectral widths $\Delta\nu_c$ for various sources of light together with their coherence times τ_c and coherence lengths in free space $l_c = c_o\tau_c$.

Source	$\Delta\nu_c$ (Hz)	$\tau_c = 1/\Delta\nu_c$	$l_c = c_o\tau_c$
Filtered sunlight ($\lambda_o = 0.4$–0.8 μm)	3.74×10^{14}	2.67 fs	800 nm
Light-emitting diode ($\lambda_o = 1$ μm, $\Delta\lambda_o = 50$ nm)	1.5×10^{13}	67 fs	20 μm
Low-pressure sodium lamp	5×10^{11}	2 ps	600 μm
Multimode He–Ne laser ($\lambda_o = 633$ nm)	1.5×10^{9}	0.67 ns	20 cm
Single-mode He–Ne laser ($\lambda_o = 633$ nm)	1×10^{6}	1 μs	300 m

EXAMPLE 12.1-1. ***A Wave Comprising a Random Sequence of Wavepackets.*** Light emitted from an incoherent source may be modeled as a sequence of wavepackets emitted at random times (Fig. 12.1-6).[†] Each wavepacket is taken to have a random phase since it is emitted by a different atom.

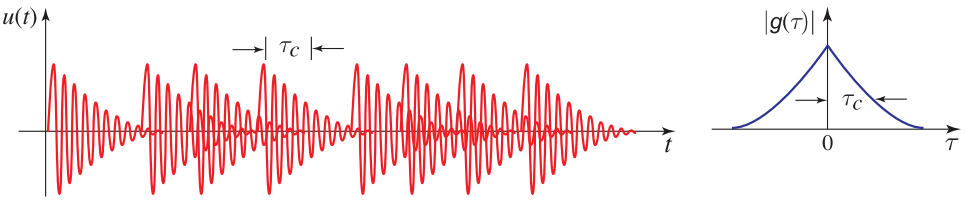


Figure 12.1-6 Light comprising wavepackets emitted at random times has a coherence time equal to the duration of a wavepacket.

[†] See B. E. A. Saleh, D. Stoler, and M. C. Teich, Coherence and Photon Statistics for Optical Fields Generated by Poisson Random Emissions, *Physical Review A*, vol. 27, pp. 360–374, 1983.

The individual wavepackets may be sinusoidal with an exponentially decaying envelope, for example, so that at a given position a wavepacket emitted at $t = 0$ has a complex wavefunction given by

$$U_p(t) = \begin{cases} A_p \exp\left(-\dfrac{t}{\tau_c}\right) \exp(j2\pi\nu_0 t), & t \geq 0 \\ 0, & t < 0. \end{cases} \tag{12.1-20}$$

The emission times are totally random, and the random independent phases of the different emissions are included in A_p. The statistical properties of the total field may be determined by performing the necessary averaging operations using the rules of mathematical statistics. The result is a complex degree of coherence given by $g(\tau) = \exp(-|\tau|/\tau_c)\exp(j2\pi\nu_0\tau)$, whose magnitude is a double-sided exponential function. The corresponding power spectral density is Lorentzian, $\mathcal{S}(\nu) = (\Delta\nu/2\pi)/[(\nu - \nu_0)^2 + (\Delta\nu/2)^2]$, where $\Delta\nu = 1/\pi\tau_c$ (see Table A.1-1 in Appendix A). The coherence time τ_c in this case turns out to be exactly the width of a wavepacket. The statement that this light is correlated within the coherence time therefore signifies that it is correlated within the duration of an individual wavepacket.

C. Spatial Coherence

Mutual Coherence Function

An important descriptor of the spatial and temporal fluctuations of the random function $U(\mathbf{r}, t)$ is the cross-correlation function of $U(\mathbf{r}_1, t)$ and $U(\mathbf{r}_2, t)$ at pairs of positions $\mathbf{r}_1$ and $\mathbf{r}_2$:

$$\boxed{G(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle U^*(\mathbf{r}_1, t)\, U(\mathbf{r}_2, t + \tau)\rangle.} \tag{12.1-21}$$

Mutual Coherence Function

This function of the two positions and the time delay τ is known as the **mutual coherence function**. Its normalized form is known as the **complex degree of coherence**:

$$\boxed{g(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{G(\mathbf{r}_1, \mathbf{r}_2, \tau)}{\sqrt{I(\mathbf{r}_1) I(\mathbf{r}_2)}}.} \tag{12.1-22}$$

Complex Degree of Coherence

When the two points coincide so that $\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}$, (12.1-21) and (12.1-22) reproduce the temporal coherence function and the complex degree of temporal coherence at the position $\mathbf{r}$, as provided in (12.1-4) and (12.1-7), respectively. When $\tau = 0$ as well, we recover the intensity $I(\mathbf{r}) = G(\mathbf{r}, \mathbf{r}, 0)$ at the position $\mathbf{r}$. The analogous cross-correlation functions in the *quantum* theory of optical coherence are defined in terms of operators rather than fields and the averages are carried out with respect to the quantum state of the light, in accordance with the precepts of quantum optics (Sec. 13.3).

The complex degree of coherence $g(\mathbf{r}_1, \mathbf{r}_2, \tau)$ is the cross-correlation coefficient of the random variables $U^*(\mathbf{r}_1, t)$ and $U(\mathbf{r}_2, t + \tau)$. As with the complex degree of temporal coherence [see (12.1-8)], its absolute value is bounded between zero and unity:

$$0 \leq |g(\mathbf{r}_1, \mathbf{r}_2, \tau)| \leq 1. \tag{12.1-23}$$

This quantity is therefore considered a measure of the degree of correlation between the fluctuations at $\mathbf{r}_1$ and those at $\mathbf{r}_2$ at a time τ later.

When the two phasors $U(\mathbf{r}_1, t)$ and $U(\mathbf{r}_2, t)$ fluctuate independently and their phases are totally random (each having a phase that is equally probable between 0 and 2π), $|g(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 0$ since the average of the product $U^*(\mathbf{r}_1, t)\, U(\mathbf{r}_2, t + \tau)$ vanishes. The light fluctuations at the two points are then uncorrelated. The other limit, $|g(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$, obtains when the light fluctuations at $\mathbf{r}_1$, and at $\mathbf{r}_2$ a time τ later, are fully correlated. Note that $|g(\mathbf{r}_1, \mathbf{r}_2, 0)|$ is not necessarily unity; however, by definition $|g(\mathbf{r}, \mathbf{r}, 0)| = 1$.

The dependence of $g(\mathbf{r}_1, \mathbf{r}_2, \tau)$ on the positions and on the time delay characterizes the spatial and temporal coherence of light. Two examples of the dependence of $|g(\mathbf{r}_1, \mathbf{r}_2, \tau)|$ on the distance $|\mathbf{r}_1 - \mathbf{r}_2|$ and on the time delay τ are illustrated in Fig. 12.1-7. The temporal and spatial fluctuations of light are interrelated since light propagates in waves and the complex wavefunction $U(\mathbf{r}, t)$ must satisfy the wave equation. This imposes certain conditions on the mutual coherence function (see Exercise 12.1-3). To illustrate this point, consider, for example, a plane wave of random light traveling in the z direction at velocity c in a homogeneous and nondispersive medium. Fluctuations at the points $\mathbf{r}_1 = (0, 0, z_1)$ and $\mathbf{r}_2 = (0, 0, z_2)$ are completely correlated when the time delay is $\tau = \tau_0 \equiv |z_2 - z_1|/c$, whereupon $|g(\mathbf{r}_1, \mathbf{r}_2, \tau_0)| = 1$. Considered as a function of τ, $|g(\mathbf{r}_1, \mathbf{r}_2, \tau)|$ then has its maximum value at $\tau = \tau_0$, as illustrated in Fig. 12.1-7(b). This example will be revisited in Sec. 12.1D.

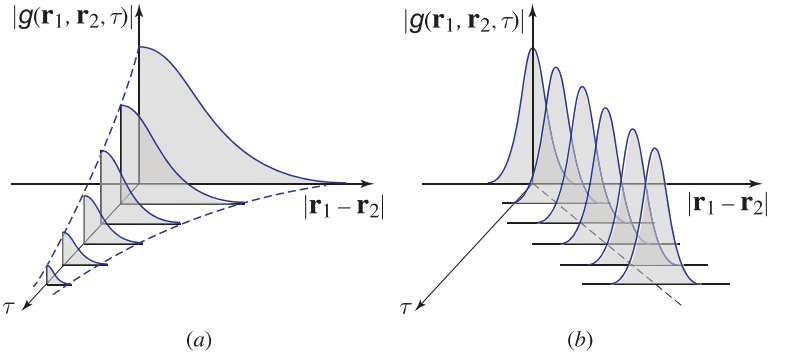


Figure 12.1-7 Two examples of $|g(\mathbf{r}_1, \mathbf{r}_2, \tau)|$ as a function of the separation $|\mathbf{r}_1 - \mathbf{r}_2|$ and the time delay τ. In (a) the maximum correlation for a given $|\mathbf{r}_1 - \mathbf{r}_2|$ occurs at $\tau = 0$, whereas in (b) the maximum correlation occurs at $|\mathbf{r}_1 - \mathbf{r}_2| = c\tau$.

EXERCISE 12.1-3

Differential Equations Governing the Mutual Coherence Function. In free space, $U(\mathbf{r}, t)$ must satisfy the wave equation, $\nabla^2 U - (1/c^2)\partial^2 U/\partial t^2 = 0$. Use the definition (12.1-21) to show that the mutual coherence function $G(\mathbf{r}_1, \mathbf{r}_2, \tau)$ satisfies a pair of partial differential equations known as the **Wolf equations**,

$$\nabla_1^2 G - \frac{1}{c^2}\frac{\partial^2 G}{\partial \tau^2} = 0 \tag{12.1-24a}$$

$$\nabla_2^2 G - \frac{1}{c^2}\frac{\partial^2 G}{\partial \tau^2} = 0, \tag{12.1-24b}$$

where ∇_1^2 and ∇_2^2 are the Laplacian operators with respect to $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively.

Mutual Intensity

The spatial correlation of light may be assessed by examining the dependence of the mutual coherence function on position at a specified fixed time delay τ. In many

situations the point $\tau = 0$ is the most appropriate to consider, as in the example illustrated in Fig. 12.1-7(a). The mutual coherence function at $\tau = 0$ is known as the **mutual intensity**,

$$G(\mathbf{r}_1, \mathbf{r}_2, 0) = \langle U^*(\mathbf{r}_1, t)\, U(\mathbf{r}_2, t)\rangle, \tag{12.1-25}$$

and is usually denoted by $G(\mathbf{r}_1, \mathbf{r}_2)$ for simplicity. The diagonal values of the mutual intensity ($\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}$) yield the intensity $I(\mathbf{r}) = G(\mathbf{r}, \mathbf{r})$. It is sometimes appropriate to use values of τ other than $\tau = 0$, however, as the example in Fig. 12.1-7(b) illustrates.

When the optical pathlength differences encountered in an optical system are much shorter than the coherence length $l_c = c\tau_c$, the light effectively possesses complete temporal coherence, in which case the mutual coherence function is a harmonic function of time [see (12.1-10)],

$$G(\mathbf{r}_1, \mathbf{r}_2, \tau) = G(\mathbf{r}_1, \mathbf{r}_2) \exp(j\omega_0 \tau), \tag{12.1-26}$$

where $\nu_0 = \omega_0/2\pi$ is the central frequency. The light is then referred to as **quasi-monochromatic** and the mutual intensity $G(\mathbf{r}_1, \mathbf{r}_2)$ completely describes the spatial coherence.

At $\tau = 0$, the complex degree of coherence $g(\mathbf{r}_1, \mathbf{r}_2, 0)$ is called the **normalized mutual intensity** and is denoted $g(\mathbf{r}_1, \mathbf{r}_2)$:

$$g(\mathbf{r}_1, \mathbf{r}_2) = \frac{G(\mathbf{r}_1, \mathbf{r}_2)}{\sqrt{I(\mathbf{r}_1) I(\mathbf{r}_2)}}. \tag{12.1-27}$$

Normalized Mutual Intensity

The magnitude $|g(\mathbf{r}_1, \mathbf{r}_2)|$ is bounded between zero and unity and is regarded as a measure of the degree of spatial coherence when the time delay $\tau = 0$. If the complex wavefunction $U(\mathbf{r}, t)$ is deterministic, $|g(\mathbf{r}_1, \mathbf{r}_2)| = 1$ for all $\mathbf{r}_1$ and $\mathbf{r}_2$, and the light is completely correlated everywhere.

Coherence Area

In a given plane, in the vicinity of a given position $\mathbf{r}_2$, the spatial coherence of quasi-monochromatic light is described by $|g(\mathbf{r}_1, \mathbf{r}_2)|$ as a function of the distance $|\mathbf{r}_1 - \mathbf{r}_2|$. This function is unity when $\mathbf{r}_1 = \mathbf{r}_2$ and decreases (but not necessarily monotonically) as $|\mathbf{r}_1 - \mathbf{r}_2|$ increases. The area scanned by the point $\mathbf{r}_1$ within which the function $|g(\mathbf{r}_1, \mathbf{r}_2)|$ is greater than some prescribed value ($1/2$ or $1/e$, for example) is called the **coherence area** A_c. It represents the spatial extent of $|g(\mathbf{r}_1, \mathbf{r}_2)|$ as a function of $\mathbf{r}_1$ for a fixed value of $\mathbf{r}_2$, as illustrated in Fig. 12.1-8. In the ideal limit of coherent light, the coherence area is infinite.

The coherence area is an important parameter for characterizing random light, but it must be viewed in relation to other pertinent dimensions of the optical system under consideration. For example, if the coherence area is greater than the size of an aperture through which the light is transmitted, we have $|g(\mathbf{r}_1, \mathbf{r}_2)| \approx 1$ at all points of interest, so that the light may be regarded as coherent (just as if A_c were infinite). Similarly, if the coherence area is smaller than the spatial resolution of the optical system, it can be regarded as infinitesimal, i.e., $g(\mathbf{r}_1, \mathbf{r}_2) \approx 0$ for practically all $\mathbf{r}_1 \neq \mathbf{r}_2$. In this limit, the light is said to be **incoherent**.

Light emitted from an extended radiating hot surface has a coherence area of the order of λ^2, where λ is the central wavelength, so that in most practical cases such light may be regarded as incoherent. Complete coherence and incoherence are therefore seen to be idealizations that represent the two limits of partial coherence.

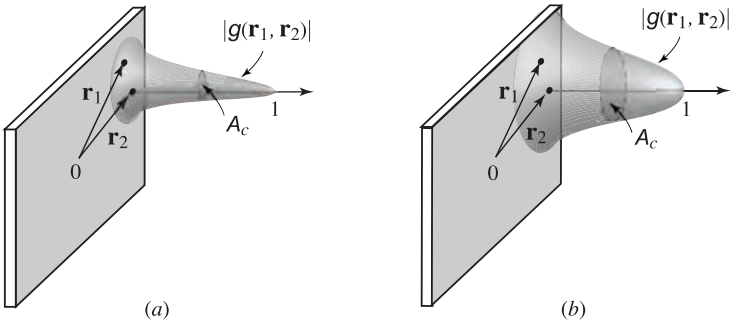


Figure 12.1-8 Two illustrative examples of the magnitude of the normalized mutual intensity as a function of $\mathbf{r}_1$ in the vicinity of a fixed point $\mathbf{r}_2$. The coherence area in (a) is smaller than that in (b).

Cross-Spectral Density

The mutual coherence function $G(\mathbf{r}_1, \mathbf{r}_2, \tau)$ describes the spatial correlation at each time delay τ. The time delay $\tau = 0$ is selected to define the mutual intensity $G(\mathbf{r}_1, \mathbf{r}_2) = G(\mathbf{r}_1, \mathbf{r}_2, 0)$, which is suitable for describing the spatial coherence of quasi-monochromatic light. A useful alternative is to describe coherence in the frequency domain by examining the spatial correlation at a fixed frequency. The **cross-spectral density** (or the cross-power spectrum) is defined as the Fourier transform of $G(\mathbf{r}_1, \mathbf{r}_2, \tau)$ with respect to τ:

$$S(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} G(\mathbf{r}_1, \mathbf{r}_2, \tau) \exp(-j2\pi\nu\tau)\, d\tau. \tag{12.1-28}$$

Cross-Spectral Density

When $\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}$, the cross-spectral density becomes the power-spectral density $S(\nu)$ at position $\mathbf{r}$, as defined in (12.1-17).

The **normalized cross-spectral density** is defined by

$$s(\mathbf{r}_1, \mathbf{r}_2, \nu) = \frac{S(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{S(\mathbf{r}_1, \mathbf{r}_1, \nu)\, S(\mathbf{r}_2, \mathbf{r}_2, \nu)}}. \tag{12.1-29}$$

Its magnitude can be shown to be bounded between zero and unity, so that it serves as a measure of the degree of spatial coherence at the frequency ν. It represents the degree of correlation of the fluctuation components of frequency ν at positions $\mathbf{r}_1$ and $\mathbf{r}_2$.

In certain cases, the cross-spectral density factors into a product of a function of position and another of frequency, $S(\mathbf{r}_1, \mathbf{r}_2, \nu) = G(\mathbf{r}_1, \mathbf{r}_2)s(\nu)$, so that the spatial and spectral properties are separable. The light is then said to be **cross-spectrally pure**. The mutual coherence function must then also factor into a product of a function of position and another of time, $G(\mathbf{r}_1, \mathbf{r}_2, \tau) = G(\mathbf{r}_1, \mathbf{r}_2)g(\tau)$, where $g(\tau)$ is the inverse Fourier transform of $s(\nu)$. If the factorization parts are selected such that $\int s(\nu)\, d\nu = 1$, then $G(\mathbf{r}_1, \mathbf{r}_2) = G(\mathbf{r}_1, \mathbf{r}_2, 0)$, so that $G(\mathbf{r}_1, \mathbf{r}_2)$ is nothing but the mutual intensity. Cross-spectrally pure light has two important properties:

1. At a single position $\mathbf{r}$, $S(\mathbf{r}, \mathbf{r}, \nu) = G(\mathbf{r}, \mathbf{r})s(\nu) = I(\mathbf{r})s(\nu)$. The spectrum has the same profile at all positions. If the light represents a visible image, it would appear to have the same color everywhere but the brightness would vary.

2. The normalized cross-spectral density

$$s(\mathbf{r}_1, \mathbf{r}_2, \nu) = G(\mathbf{r}_1, \mathbf{r}_2) \Big/ \sqrt{G(\mathbf{r}_1, \mathbf{r}_1)\, G(\mathbf{r}_2, \mathbf{r}_2)} = g(\mathbf{r}_1, \mathbf{r}_2) \tag{12.1-30}$$

is independent of frequency. In this case the normalized mutual intensity $g(\mathbf{r}_1, \mathbf{r}_2)$ describes the spatial coherence at all frequencies.

D. Longitudinal Coherence

In this section the concept of longitudinal coherence is introduced by taking examples of random waves with fixed wavefronts, such as plane and spherical waves.

Partially Coherent Plane Wave

Consider a plane wave

$$U(\mathbf{r}, t) = \mathfrak{a}\Big(t - \frac{z}{c}\Big) \exp\Big[j\omega_0 \Big(t - \frac{z}{c}\Big)\Big] \tag{12.1-31}$$

traveling in the z direction in a homogeneous medium with velocity c, as considered in Sec. 2.6A. The complex wavefunction $U(\mathbf{r}, t)$ satisfies the wave equation for an arbitrary function $\mathfrak{a}(t)$. If $\mathfrak{a}(t)$ is a random function, $U(\mathbf{r}, t)$ represents partially coherent light. The mutual coherence function defined in (12.1-21) is then

$$G(\mathbf{r}_1, \mathbf{r}_2, \tau) = G_{\mathfrak{a}}\left(\tau - \frac{z_2 - z_1}{c}\right) \exp\left[j\omega_0 \left(\tau - \frac{z_2 - z_1}{c}\right)\right], \tag{12.1-32}$$

where z_1 and z_2 are the z components of $\mathbf{r}_1$ and $\mathbf{r}_2$ and $G_{\mathfrak{a}}(\tau) = \langle \mathfrak{a}^*(t)\mathfrak{a}(t+\tau)\rangle$ is the autocorrelation function of $\mathfrak{a}(t)$, which is assumed to be independent of t [see Fig. 12.1-7(*b*)].

The intensity $I(\mathbf{r}) = G(\mathbf{r}, \mathbf{r}, 0) = G_{\mathfrak{a}}(0)$ is constant everywhere in space. Temporal coherence is characterized by the time function $G(\mathbf{r}, \mathbf{r}, \tau) = G_{\mathfrak{a}}(\tau) \exp(j\omega_0\tau)$, which is independent of position. The complex degree of coherence is $g(\mathbf{r}, \mathbf{r}, \tau) = g_{\mathfrak{a}}(\tau) \exp(j\omega_0\tau)$, where $g_{\mathfrak{a}}(\tau) = G_{\mathfrak{a}}(\tau)/G_{\mathfrak{a}}(0)$. The width of $|g_{\mathfrak{a}}(\tau)| = |g(\mathbf{r}, \mathbf{r}, \tau)|$, defined by an expression similar to (12.1-10), is the coherence time τ_c. It is the same at all positions.

The power spectral density is the Fourier transform of $G(\mathbf{r}, \mathbf{r}, \tau)$ with respect to τ. From (12.1-32), $S(\nu)$ is seen to be equal to the Fourier transform of $G_{\mathfrak{a}}(\tau)$ shifted by a frequency ν_0 (in accordance with the frequency shift property of the Fourier transform defined in Appendix A, Sec. A.1). The wave therefore has the same power spectral density everywhere in space.

The spatial coherence properties are described by

$$G(\mathbf{r}_1, \mathbf{r}_2, 0) = G_{\mathfrak{a}}\left(\frac{z_1 - z_2}{c}\right) \exp\left[j\omega_0 \frac{z_1 - z_2}{c}\right], \tag{12.1-33}$$

and its normalized version

$$g(\mathbf{r}_1, \mathbf{r}_2, 0) = g_{\mathfrak{a}}\left(\frac{z_1 - z_2}{c}\right) \exp\left[j\omega_0 \frac{z_1 - z_2}{c}\right]. \tag{12.1-34}$$

If the two points $\mathbf{r}_1$ and $\mathbf{r}_2$ lie in the same transverse plane, i.e., $z_1 = z_2$, then $|g(\mathbf{r}_1, \mathbf{r}_2, 0)| = |g_{\mathfrak{a}}(0)| = 1$. This means that fluctuations at points on a wavefront

(a plane normal to the z axis) are completely correlated; the coherence area in any transverse plane is infinite (see Fig. 12.1-9). On the other hand, fluctuations at two points separated by an axial distance z_2-z_1 such that $|z_2-z_1|/c > \tau_c$, or $|z_2-z_1| > l_c$ where $l_c = c\tau_c$ is the coherence length, are approximately uncorrelated.

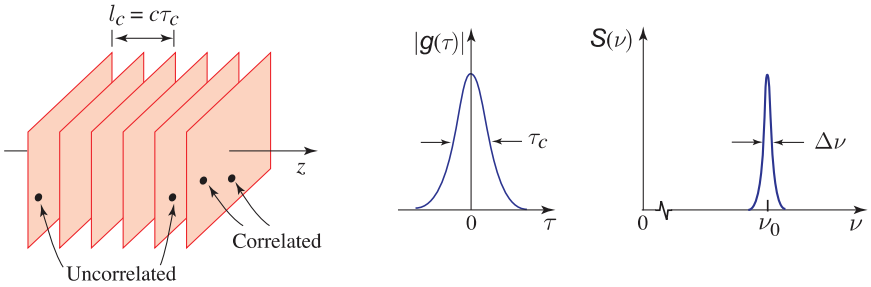


Figure 12.1-9 The fluctuations of a partially coherent plane wave at points on any wavefront (transverse plane) are completely correlated, whereas those at points on wavefronts separated by an axial distance greater than the coherence length $l_c = c\tau_c$ are approximately uncorrelated.

We conclude that the partially coherent plane wave is spatially coherent across each transverse plane, but only partially coherent in the axial direction. The axial (longitudinal) spatial coherence of the wave has a one-to-one correspondence with the temporal coherence. The relationship of the coherence length $l_c = c\tau_c$ to the maximum optical path difference in the system $l_{\max}$ governs the role played by coherence. If $l_c \gg l_{\max}$, the wave is effectively completely coherent. The coherence lengths of various light sources were provided in Table 12.1-2.

Partially Coherent Spherical Wave

A partially coherent spherical wave is described by the complex wavefunction (see Sec. 2.2B and Sec. 2.6A)

$$U(\mathbf{r},t) = \frac{1}{r}\,\mathfrak{a}\left(t-\frac{r}{c}\right)\exp\left[j\omega_0\left(t-\frac{r}{c}\right)\right], \tag{12.1-35}$$

where $\mathfrak{a}(t)$ is a random function. The corresponding mutual coherence function is

$$G(\mathbf{r}_1,\mathbf{r}_2,\tau) = \frac{1}{r_1 r_2}\,G_{\mathfrak{a}}\left(\tau-\frac{r_2-r_1}{c}\right)\exp\left[j\omega_0\left(\tau-\frac{r_2-r_1}{c}\right)\right], \tag{12.1-36}$$

with $G_{\mathfrak{a}}(\tau) = \langle \mathfrak{a}^*(t)\,\mathfrak{a}(t+\tau)\rangle$.

The intensity $I(\mathbf{r}) = G_{\mathfrak{a}}(0)/r^2$ varies in accordance with an inverse-square law. The coherence time τ_c is the width of the function $|g_{\mathfrak{a}}(\tau)| = |G_{\mathfrak{a}}(\tau)/G_{\mathfrak{a}}(0)|$. It is the same everywhere in space, as is the power spectral density. For $\tau = 0$, fluctuations at all points on a spherical wavefront are completely correlated, whereas fluctuations at points on two wavefronts separated by the radial distance $|r_2 - r_1| \gg l_c = c\tau_c$ are uncorrelated (see Fig. 12.1-10).

An arbitrary partially coherent wave transmitted through a pinhole generates a partially coherent spherical wave. This process therefore imparts spatial coherence to the incident wave (points on any sphere centered about the pinhole become completely correlated). However, the wave remains temporally partially coherent. Points at different distances from the pinhole are only partially correlated. The pinhole imparts spatial coherence, but not temporal coherence, to the wave.

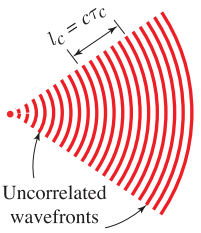


Figure 12.1-10 A partially coherent spherical wave exhibits complete spatial coherence at all points on a wavefront, while points on wavefronts separated by a distance greater than the coherence length $l_c = c\tau_c$ are approximately uncorrelated.

Suppose now that an optical filter of very narrow spectral width is placed at the pinhole, causing the transmitted wave to become approximately monochromatic. The wave will then have complete temporal as well as spatial coherence. Spatial coherence is imparted by the pinhole, which acts as a spatial filter, while temporal coherence is introduced by the narrowband filter. The price paid for obtaining such an ideally coherent wave is, of course, the loss of optical energy associated with the temporal and spatial filtering processes.

12.2 INTERFERENCE OF PARTIALLY COHERENT LIGHT

The interference of *coherent* light was discussed in Sec. 2.5. This section is devoted to the interference of *partially coherent* light.

A. Interference of Two Partially Coherent Waves

The statistical properties of two partially coherent waves U_1 and U_2 are characterized not only by their own mutual coherence functions but also by a measure of the degree to which their fluctuations are correlated. At a given position $\mathbf{r}$ and time t, the intensities of the two waves are $I_1 = \langle |U_1|^2 \rangle$ and $I_2 = \langle |U_2|^2 \rangle$, whereas their cross-correlation is described by the statistical average $G_{12} = \langle U_1^* U_2 \rangle$, along with its normalized version

$$g_{12} = \frac{\langle U_1^* U_2 \rangle}{\sqrt{I_1 I_2}} . \tag{12.2-1}$$

When the two waves are superposed, the average intensity of their sum is

$$\begin{aligned} I &= \langle |U_1 + U_2|^2 \rangle = \langle |U_1|^2 \rangle + \langle |U_2|^2 \rangle + \langle U_1^* U_2 \rangle + \langle U_1 U_2^* \rangle \\ &= I_1 + I_2 + G_{12} + G_{12}^* = I_1 + I_2 + 2\,\mathrm{Re}\{G_{12}\} \\ &= I_1 + I_2 + 2\sqrt{I_1 I_2}\ \mathrm{Re}\{g_{12}\}. \end{aligned} \tag{12.2-2}$$

We thus obtain

$$\boxed{I = I_1 + I_2 + 2\sqrt{I_1 I_2}\,|g_{12}| \cos\varphi\,,} \tag{12.2-3}$$

Interference Equation

where $\varphi = \arg\{g_{12}\}$ is the phase of g_{12}. The third term on the right-hand side of (12.2-3) represents optical interference.

It is useful to consider two limits of this equation:

1. For two *completely correlated* waves with $g_{12} = \exp(j\varphi)$ and $|g_{12}| = 1$, we recover the interference equation (2.5-4) for two coherent waves of phase difference φ.

2. For two *uncorrelated* waves with $g_{12} = 0$, the result is $I = I_1 + I_2$ and there is no interference.

In the general case, the normalized intensity versus the phase φ assumes the form of a sinusoidal pattern, as shown in Fig. 12.2-1. The strength of the interference is measured by the **visibility** $\mathcal{V}$ (also called the *modulation depth* or the *contrast* of the interference pattern):

$$\mathcal{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \tag{12.2-4}$$

where $I_{\max}$ and $I_{\min}$ are, respectively, the maximum and minimum values that I takes as φ is varied. Since $\cos\varphi$ stretches between 1 and -1, inserting (12.2-3) into (12.2-4) yields

$$\mathcal{V} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2} |g_{12}|. \tag{12.2-5}$$

The visibility is therefore proportional to the absolute value of the normalized cross-correlation $|g_{12}|$. In the special case when $I_1 = I_2$, this simplifies to

$$\boxed{\mathcal{V} = |g_{12}|.} \tag{12.2-6}$$

Visibility

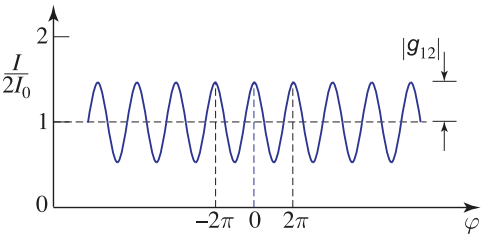


Figure 12.2-1 Normalized intensity $I/2I_0$ of the sum of two partially coherent waves of equal intensities ($I_1 = I_2 = I_0$), as a function of the phase φ of their normalized cross-correlation g_{12}. This sinusoidal pattern has visibility $\mathcal{V} = |g_{12}|$.

The interference equation (12.2-3) will now be considered in a number of specific contexts to highlight the effects that temporal and spatial coherence have on the interference of partially coherent light.

B. Interferometry and Temporal Coherence

Consider a partially coherent wave $U(t)$ with intensity I_0 and complex degree of temporal coherence $g(\tau) = \langle U^*(t)U(t+\tau)\rangle / I_0$. If $U(t)$ is simply added to a replica of itself that is delayed by the time τ, $U(t+\tau)$, what is the intensity I of the superposition?

Using the interference formula (12.2-2) with $U_1 = U(t)$, $U_2 = U(t+\tau)$, $I_1 = I_2 = I_0$, and $g_{12} = \langle U_1^* U_2\rangle / I_0 = \langle U^*(t)U(t+\tau)\rangle / I_0 = g(\tau)$, we obtain

$$I = 2I_0 [1 + \text{Re}\{g(\tau)\}] = 2I_0 [1 + |g(\tau)| \cos\varphi(\tau)], \tag{12.2-7}$$

where $\varphi(\tau) = \arg\{g(\tau)\}$. It is thus apparent that the ability of a wave to interfere with a time delayed replica of itself is governed by its complex degree of temporal coherence at that time delay.

Implementing the addition of a wave with a time-delayed replica of itself may be achieved by using a beamsplitter to generate two identical waves, one of which is made to traverse a longer optical path than the other, and then recombining them at another (or the same) beamsplitter. This can be effected, for example, with the help of a Mach–Zehnder or a Michelson interferometer (see Fig. 2.5-3).

Consider, as an example, the partially coherent plane wave introduced in Sec. 12.1D [see (12.1-31)], whose complex degree of temporal coherence is $g(\tau) = g_a(\tau) \exp(j\omega_0\tau)$. The spectral width of the wave is $\Delta\nu_c = 1/\tau_c$, where τ_c (the width of $|g_a(\tau)|$) is the coherence time. Substituting this into (12.2-7), we obtain

$$I = 2I_0 \{1 + |g_a(\tau)| \cos [\omega_0\tau + \varphi_a(\tau)]\}, \tag{12.2-8}$$

where $\varphi_a(\tau) = \arg\{g_a(\tau)\}$.

This relation between I and τ, which is known as an **interferogram**, is illustrated in Fig. 12.2-2. Assuming that $\Delta\nu_c = 1/\tau_c \ll \nu_0$, the functions $|g_a(\tau)|$ and $\varphi_a(\tau)$ vary slowly in comparison with the period $1/\nu_0$. The visibility of this interferogram in the vicinity of a particular time delay τ is $\mathcal{V} = |g(\tau)| = |g_a(\tau)|$. It has a peak value of unity near $\tau = 0$ and vanishes for $\tau \gg \tau_c$, i.e., when the optical path difference is much greater than the coherence length $l_c = c\tau_c$. For the Michelson interferometer illustrated in Fig. 12.2-2, $\tau = 2(d_2 - d_1)/c$. Interference occurs only when the optical path difference is smaller than the coherence length.

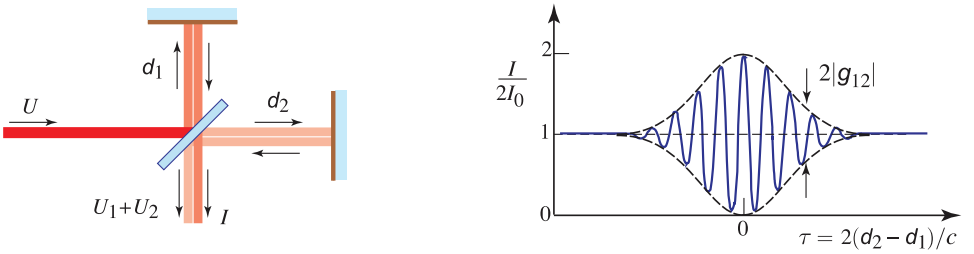


Figure 12.2-2 The normalized intensity $I/2I_0$, as a function of the time delay τ, when a partially coherent plane wave is introduced into a Michelson interferometer. The visibility is a measure of the magnitude of the complex degree of temporal coherence.

The magnitude of the complex degree of temporal coherence of a wave, $|g(\tau)|$, may therefore be measured by monitoring the visibility of the interference pattern as a function of time delay. The phase of $g(\tau)$ may be measured by observing the locations of the peaks of the pattern.

Fourier-Transform Spectroscopy

It is revealing to write (12.2-7) in terms of the power spectral density of the wave $S(\nu)$. Using the Fourier-transform relation between $G(\tau)$ and $S(\nu)$,

$$G(\tau) = I_0 g(\tau) = \int_0^\infty S(\nu) \exp(j2\pi\nu\tau)\, d\nu, \tag{12.2-9}$$

substituting into (12.2-7), and noting that $S(\nu)$ is real and that $\int_0^\infty S(\nu)\, d\nu = I_0$, we obtain

$$I = 2 \int_0^\infty S(\nu) [1 + \cos(2\pi\nu\tau)]\, d\nu. \tag{12.2-10}$$

This equation can be interpreted as representing a weighted superposition of interferograms produced by each of the monochromatic components of the wave. Each component ν produces an interferogram with period $1/\nu$ and unity visibility, but the composite interferogram exhibits reduced visibility by virtue of the different periods.

Equation (12.2-10) suggests that the spectral density $S(\nu)$ of a light source can be determined by measuring the interferogram I versus τ and then inverting the result by means of Fourier-transform methods. This technique is known as **Fourier-transform spectroscopy**.

Optical Coherence Tomography

Optical coherence tomography (OCT) is an interferometric technique for profiling a multilayered medium, i.e., for measuring the reflectance and depth of each of its boundaries. In its simplest form, known as **time-domain OCT**, it makes use of a partially coherent light source of short coherence length and a Michelson interferometer. As illustrated in Fig. 12.2-3, a replica of the original wave, delayed by a movable mirror, is superposed with a collection of waves reflected from the multiple boundaries of the sample. Information about the sample profile is carried by the interferogram, which is the intensity measured at the detector as the movable mirror is translated. By virtue of the short coherence length of the source, the interferogram comprises sets of fringes centered at path delays of the movable mirror that match those of the reflecting boundaries.

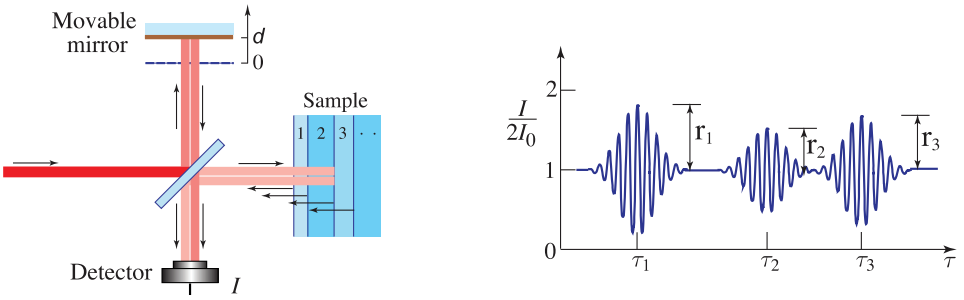


Figure 12.2-3 Optical coherence tomography.

Let $U(t-\tau)$ be the wave reflected from the movable mirror, with its associated time delay $\tau = d/c_o$, and let $\mathrm{r}_i U(t-\tau_i)$, $i = 1, 2, \ldots$, be the waves reflected from the boundaries of the sample, where r_i represents the amplitude reflectance at the ith boundary; the associated time delays are designated τ_i. For a symmetric beamsplitter, the average intensity is then $I(\tau) = \langle |U(t-\tau) + \sum_i \mathrm{r}_i U(t-\tau_i)|^2 \rangle$, which may be written in normalized form as

$$I/2I_0 = 1 + \sum_i \mathrm{r}_i \,\mathrm{Re}\{g(\tau - \tau_i)\} + \sum_{ij} \mathrm{r}_i \mathrm{r}_j^* \,\mathrm{Re}\{g(\tau_j - \tau_i)\}, \qquad (12.2\text{-}11)$$

since the complex degree of temporal coherence of the source is characterized by $g(\tau) = \langle U^*(t)\, U(t+\tau)\rangle \big/ \langle U^*(t) U(t)\rangle$.

The left-most summation on the right-hand side of (12.2-11) is of paramount importance since it represents interference between the reference wave from the movable mirror and each of the waves reflected from the sample boundaries. The right-most summation represents interference terms associated with pairs of reflections from the sample; since these terms are independent of the path delay of the movable mirror, $\tau = d/c$, they may be regarded as background contributions and ignored.

For a light source of central frequency ν_0, we have $g(\tau) = g_a(\tau)\exp(j\omega_0\tau)$, where the width of $g_a(\tau)$ is the coherence time τ_c. Equation (12.2-11) then becomes

$$I/2I_0 \approx 1 + \sum_i r_i \left| g_a(\tau - \tau_i) \right| \cos\left[\omega_0(\tau - \tau_i) + \varphi_a(\tau - \tau_i)\right], \qquad (12.2\text{-}12)$$

where $\varphi_a(\tau) = \arg\{g_a(\tau)\}$. If the source is of short coherence length, the function $g_a(\tau)$ is narrow. As illustrated in Fig. 12.2-3, the reflection from each sample boundary then generates a distinct set of interference fringes of brief duration τ_c, centered about its corresponding time delay. Measurement of the OCT interferogram therefore permits the reflectance at each boundary, as well as the width of each of the sample layers, to be determined.

Optical coherence tomography has proven to be an effective imaging technique in clinical medicine as well as in engineering. It can also be carried out in a **frequency-domain OCT** configuration in the spirit of Fourier-transform spectroscopy. A particularly useful configuration makes use of a narrowband optical source whose frequency is swept in time (e.g., a wavelength-swept laser); this approach offers improved detection sensitivity and data-acquisition rates.

C. Interferometry and Spatial Coherence

The effect of spatial coherence on interference is demonstrated by considering the Young's double-pinhole interference experiment discussed in Exercise 2.5-2 for coherent light. A partially coherent optical wave $U(\mathbf{r}, t)$ illuminates an opaque screen with two pinholes located at positions $\mathbf{r}_1$ and $\mathbf{r}_2$. The wave has mutual coherence function $G(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle U^*(\mathbf{r}_1, t)\, U(\mathbf{r}_2, t+\tau)\rangle$ and complex degree of coherence $g(\mathbf{r}_1, \mathbf{r}_2, \tau)$. The intensities at the pinholes are assumed to be equal.

Light is diffracted in the form of two spherical waves centered at the pinholes. The two waves interfere, and the intensity I of their sum is observed at a point $\mathbf{r}$ in the observation plane, at a distance d from the screen that is sufficiently large so that the paraboloidal approximation is applicable. In Cartesian coordinates, as shown in Fig. 12.2-4, $\mathbf{r}_1 = (-a, 0, 0)$, $\mathbf{r}_2 = (a, 0, 0)$, and $\mathbf{r} = (x, 0, d)$. The intensity is observed as a function of x. An important geometrical parameter is the angle $\theta \approx 2a/d$ subtended by the two pinholes.

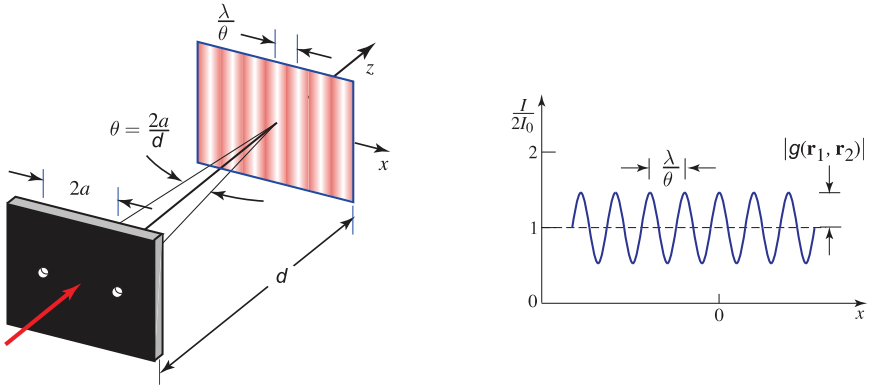


Figure 12.2-4 Young's double-pinhole interferometer illuminated by partially coherent light. The incident wave is quasi-monochromatic and has a normalized mutual intensity at the pinholes described by $g(\mathbf{r}_1, \mathbf{r}_2)$. The normalized intensity $I/2I_0$ in the observation plane, at a large distance from the pinhole plane, is a sinusoidal function of x with period λ/θ and visibility $\mathcal{V} = |g(\mathbf{r}_1, \mathbf{r}_2)|$.

In the paraboloidal (Fresnel) approximation [see (2.2-17)], the two diffracted spherical waves are approximately related to $U(\mathbf{r}, t)$ by

$$U_1(\mathbf{r}, t) \propto U\left(\mathbf{r}_1, t - \frac{|\mathbf{r} - \mathbf{r}_1|}{c}\right) \approx U\left(\mathbf{r}_1, t - \frac{d + (x + a)^2/2d}{c}\right) \tag{12.2-13a}$$

$$U_2(\mathbf{r}, t) \propto U\left(\mathbf{r}_2, t - \frac{|\mathbf{r} - \mathbf{r}_2|}{c}\right) \approx U\left(\mathbf{r}_2, t - \frac{d + (x - a)^2/2d}{c}\right), \tag{12.2-13b}$$

and have approximately equal intensities, $I_1 = I_2 = I_0$. The normalized cross-correlation between the two waves at $\mathbf{r}$ is

$$g_{12} = \frac{\langle U_1^*(\mathbf{r}, t) U_2(\mathbf{r}, t)\rangle}{I_0} = g(\mathbf{r}_1, \mathbf{r}_2, \tau_x), \tag{12.2-14}$$

where the difference in the time delays encountered by the two waves is given by

$$\tau_x = \frac{|\mathbf{r} - \mathbf{r}_1| - |\mathbf{r} - \mathbf{r}_2|}{c} = \frac{(x + a)^2 - (x - a)^2}{2dc} = \frac{2ax}{dc} = \frac{\theta}{c} x. \tag{12.2-15}$$

Substituting (12.2-14) into the interference formula (12.2-3) gives rise to an observed intensity $I \equiv I(x)$:

$$\boxed{I(x) = 2I_0 \left[1 + |g(\mathbf{r}_1, \mathbf{r}_2, \tau_x)| \cos \varphi_x\right],} \tag{12.2-16}$$

where $\varphi_x = \arg\{g(\mathbf{r}_1, \mathbf{r}_2, \tau_x)\}$. This equation describes the pattern of intensity observed as a function of position x in the observation plane, in terms of the magnitude and phase of the complex degree of coherence at the pinholes at time delay $\tau_x = \theta x/c$.

Quasi-Monochromatic Light

Moreover, if the light is quasi-monochromatic with central frequency $\nu_0 = \omega_0/2\pi$, i.e., if $g(\mathbf{r}_1, \mathbf{r}_2, \tau) \approx g(\mathbf{r}_1, \mathbf{r}_2) \exp(j\omega_0\tau)$, then (12.2-16) provides

$$\boxed{I(x) = 2I_0 \left[1 + \mathcal{V} \cos\left(\frac{2\pi\theta}{\lambda} x + \varphi\right)\right],} \tag{12.2-17}$$

where $\lambda = c/\nu_0$, $\mathcal{V} = |g(\mathbf{r}_1, \mathbf{r}_2)|$, $\tau_x = \theta x/c$, and $\varphi = \arg\{g(\mathbf{r}_1, \mathbf{r}_2)\}$. The interference-fringe pattern is then sinusoidal with spatial period λ/θ and visibility $\mathcal{V}$. In analogy with the temporal case, the visibility of the interference pattern is equal to the magnitude of the complex degree of spatial coherence at the two pinholes (Fig. 12.2-4). The locations of the peaks depend on the phase φ.

If the incident wave in a Young's interferometer is a coherent plane wave traveling in the z direction, $U(\mathbf{r}, t) = \exp(-jkz)\exp(j\omega_0 t)$, then $g(\mathbf{r}_1, \mathbf{r}_2) = 1$, whereupon $|g(\mathbf{r}_1, \mathbf{r}_2)| = 1$ and $\arg\{g(\mathbf{r}_1, \mathbf{r}_2)\} = 0$. The interference pattern then has unity visibility and a peak at $x = 0$. However, if the illumination is, instead, a tilted plane wave arriving from a direction in the x–z plane that makes a small angle θ_x with respect to the z axis, i.e., $U(\mathbf{r}, t) \approx \exp[-j(kz + k\theta_x x)]\exp(j\omega_0 t)$, then $g(\mathbf{r}_1, \mathbf{r}_2) = \exp(-jk\theta_x 2a)$. The visibility remains at $\mathcal{V} = 1$, but the tilt results in a phase shift $\varphi = -k\theta_x 2a = -2\pi\theta_x 2a/\lambda$, so that the interference pattern is shifted laterally by a fraction $(2a\theta_x/\lambda)$ of a period. When $\varphi = 2\pi$, the pattern is shifted by one period.

Interference with Light from an Extended Source

Suppose now that the incident light is a collection of independent plane waves arriving from a source that subtends an angle θ_s at the plane of the pinhole (Fig. 12.2-5). The phase shift φ then takes values in the range $\pm 2\pi(\theta_s/2)2a/\lambda = \pm 2\pi\theta_s a/\lambda$ and the fringe pattern is a superposition of displaced sinusoids. If $\theta_s = \lambda/2a$, then φ takes on values in the range $\pm\pi$, which is sufficient to wash out the interference pattern and thereby reduce its visibility to zero.

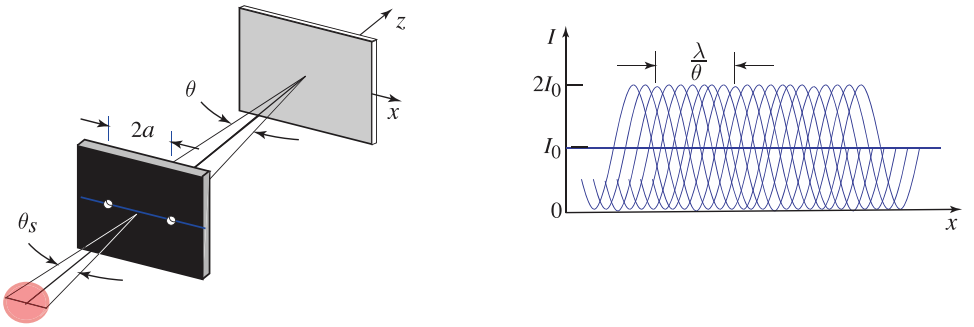


Figure 12.2-5 Young's interference fringes are washed out if the illumination emanates from a source of angular diameter $\theta_s > \lambda/2a$. If the distance $2a$ is smaller than λ/θ_s, the fringes become visible.

We conclude that the degree of spatial coherence at the two pinholes is very small when the angle subtended by the source is $\theta_s \geq \lambda/2a$. Consequently, a measure of the **coherence distance** in the plane of the screen is

$$\rho_c \approx \frac{\lambda}{\theta_s}, \tag{12.2-18}$$

Coherence Distance

and a measure of the coherence area of light emitted from a source subtending an angle θ_s is described by

$$A_c \approx \left(\frac{\lambda}{\theta_s}\right)^2. \tag{12.2-19}$$

The angle subtended by the sun, for example, is $0.5°$, so that the coherence distance for filtered sunlight of wavelength λ is $\rho_c \approx \lambda/\theta_s \approx 115\lambda$. At $\lambda = 0.5\ \mu$m, $\rho_c \approx 57.5\ \mu$m.

A more rigorous analysis (see Sec. 12.3C) reveals that the transverse coherence distance ρ_c for a circular incoherent light source of uniform intensity is

$$\rho_c = 1.22\frac{\lambda}{\theta_s}. \tag{12.2-20}$$

Effect of Spectral Width on Interference

Finally, we examine the effect of the spectral width of the light on interference in the Young's double-pinhole interferometer. The power spectral density of the incident

wave is assumed to be a narrow function of width $\Delta\nu_c$ centered about ν_0, and $\Delta\nu_c \ll \nu_0$. The complex degree of coherence then takes the form

$$g(\mathbf{r}_1, \mathbf{r}_2, \tau) = g_a(\mathbf{r}_1, \mathbf{r}_2, \tau) \exp(j\omega_0 \tau), \tag{12.2-21}$$

where $g_a(\mathbf{r}_1, \mathbf{r}_2, \tau)$ is a slowly varying function of τ (in comparison with the period $1/\nu_0$). Substituting (12.2-21) into (12.2-16), we obtain

$$\boxed{I(x) = 2I_0 \left[1 + \mathcal{V}_x \cos\left(\frac{2\pi\theta}{\bar{\lambda}} x + \varphi_x\right)\right],} \tag{12.2-22}$$

where $\mathcal{V}_x = |g_a(\mathbf{r}_1, \mathbf{r}_2, \tau_x)|$, $\varphi_x = \arg\{g_a(\mathbf{r}_1, \mathbf{r}_2, \tau_x)\}$, $\tau_x = \theta x/c$, and $\bar{\lambda} = c/\nu_0$.

The resulting interference pattern is sinusoidal with period $\bar{\lambda}/\theta$ but with varying visibility $\mathcal{V}_x$ and varying phase φ_x, specified by the magnitude and phase of the complex degree of coherence at the two pinholes, respectively, evaluated at the time delay $\tau_x = \theta x/c$. If $|g_a(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$ at $\tau = 0$ and decreases with increasing τ, vanishing for $\tau \gg \tau_c$, then the visibility $\mathcal{V}_x = 1$ at $x = 0$ and decreases with increasing x, vanishing for $x \gg x_c = c\tau_c/\theta$. The interference pattern is then visible over a distance

$$x_c = \frac{l_c}{\theta}, \tag{12.2-23}$$

where $l_c = c\tau_c$ is the coherence length and θ is the angle subtended by the two pinholes (Fig. 12.2-6).

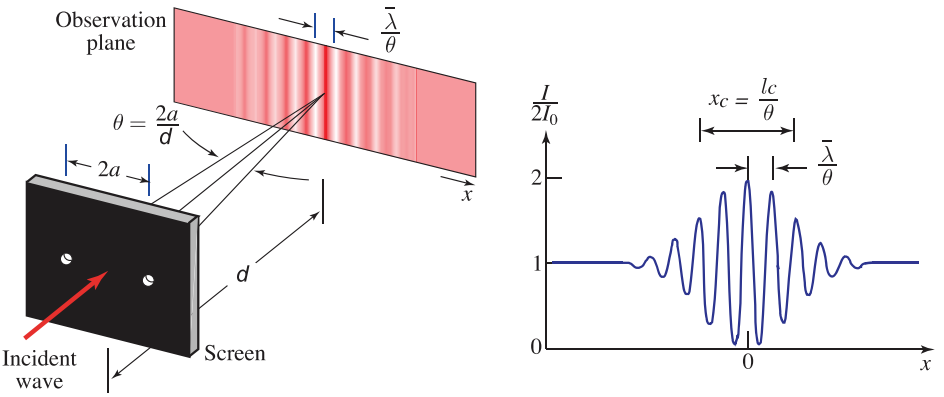


Figure 12.2-6 The visibility of Young's interference fringes at position x is the magnitude of the complex degree of coherence at the pinholes, at a time delay $\tau_x = \theta x/c$. For spatially coherent light the number of observable fringes is the ratio of the coherence length to the central wavelength, or, equivalently, the ratio of the central frequency to the spectral linewidth.

The number of observable fringes is thus $x_c/(\bar{\lambda}/\theta) = l_c/\bar{\lambda} = c\tau_c/\bar{\lambda} = \nu_0/\Delta\nu_c$. It is thus equal to the ratio of the coherence length to the central wavelength, $l_c/\bar{\lambda}$, or the ratio of the central frequency to the linewidth, $\nu_0/\Delta\nu_c$. Clearly, if $|g(\mathbf{r}_1, \mathbf{r}_2, 0)| < 1$, i.e., if the source is not spatially coherent, the visibility will be further reduced and even fewer fringes will be observable.

*12.3 TRANSMISSION OF PARTIALLY COHERENT LIGHT

The transmission of coherent light through thin optical components, apertures, and free space was discussed in Chapters 2 and 4. In this section we revisit this discussion for quasi-monochromatic partially coherent light. We assume that the spectral width of the light is sufficiently small so that the coherence length $l_c = c\tau_c = c/\Delta\nu_c$ is much greater than the differences of optical pathlengths in the system. The mutual coherence function may then be approximated by $G(\mathbf{r}_1, \mathbf{r}_2, \tau) \approx G(\mathbf{r}_1, \mathbf{r}_2) \exp(j2\pi\nu_0\tau)$, where $G(\mathbf{r}_1, \mathbf{r}_2)$ is the mutual intensity and ν_0 is the central frequency.

We observe at the outset that the laws of transmission applicable to the deterministic function $U(\mathbf{r})$ representing coherent light are also applicable to the random function $U(\mathbf{r})$ representing partially coherent light. However, for partially coherent light our interest is in the laws that govern statistical averages: the intensity $I(\mathbf{r})$ and the mutual intensity $G(\mathbf{r}_1, \mathbf{r}_2)$.

A. Propagation of Partially Coherent Light

Transmission Through Thin Optical Components

When a partially coherent wave is transmitted through a thin optical component characterized by an amplitude transmittance $\mathrm{t}(x, y)$, the incident and transmitted waves are related by $U_2(\mathbf{r}) = \mathrm{t}(\mathbf{r})U_1(\mathbf{r})$ where $\mathbf{r} = (x, y)$ is the position in the plane of the component (see Fig. 12.3-1). Using the definition of the mutual intensity, $G(\mathbf{r}_1, \mathbf{r}_2) = \langle U^*(\mathbf{r}_1)U(\mathbf{r}_2)\rangle$, we obtain

$$G_2(\mathbf{r}_1, \mathbf{r}_2) = \mathrm{t}^*(\mathbf{r}_1)\,\mathrm{t}(\mathbf{r}_2)\,G_1(\mathbf{r}_1, \mathbf{r}_2), \tag{12.3-1}$$

where $G_1(\mathbf{r}_1, \mathbf{r}_2)$ and $G_2(\mathbf{r}_1, \mathbf{r}_2)$ are the mutual intensities of the incident and transmitted light, respectively.

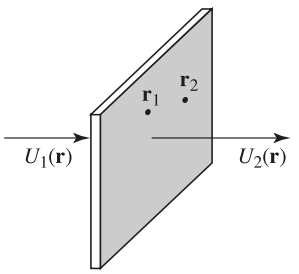


Figure 12.3-1 The absolute value of the degree of spatial coherence is not altered by transmission through a thin optical component.

Since the intensity at position $\mathbf{r}$ is equal to the mutual intensity at $\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}$, we have

$$I_2(\mathbf{r}) = |\mathrm{t}(\mathbf{r})|^2 I_1(\mathbf{r}). \tag{12.3-2}$$

The normalized mutual intensities defined by (12.1-27) therefore satisfy

$$|g_2(\mathbf{r}_1, \mathbf{r}_2)| = |g_1(\mathbf{r}_1, \mathbf{r}_2)|. \tag{12.3-3}$$

Although transmission through a thin optical component may change the intensity of partially coherent light, it does not alter the magnitude of its degree of spatial coherence. Of course, if the complex amplitude transmittance of the component itself were random, the coherence of the transmitted light would be altered accordingly.

Transmission Through an Arbitrary Optical System

We next consider transmission through an arbitrary optical system — one that includes propagation in free space and thick optical components. It was shown in Chapter 4 that the complex amplitude $U_2(\mathbf{r})$ at a point $\mathbf{r} = (x, y)$ in the output plane of such a system is generally a weighted superposition integral comprising contributions from the complex amplitudes $U_1(\mathbf{r})$ at points $\mathbf{r}' = (x', y')$ in the input plane,

$$U_2(\mathbf{r}) = \int h(\mathbf{r}; \mathbf{r}')\, U_1(\mathbf{r}')\, d\mathbf{r}', \tag{12.3-4}$$

where $h(\mathbf{r}; \mathbf{r}')$ is the impulse response function of the system (see Fig. 12.3-2). Equation (12.3-4) is a double integral with respect to $\mathbf{r}' = (x', y')$ that extends over the entire input plane.

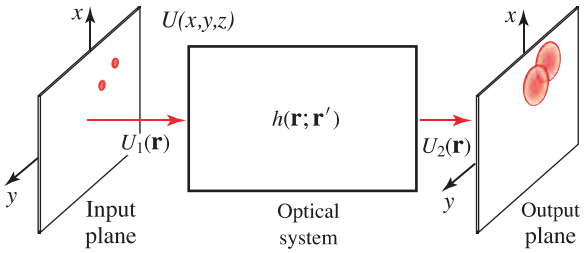


Figure 12.3-2 An optical system is characterized by its impulse response function $h(\mathbf{r}; \mathbf{r}')$.

To translate this relation between the random complex-amplitude functions $U_2(\mathbf{r})$ and $U_1(\mathbf{r})$ into a relation between their mutual intensities, we substitute (12.3-4) into the definition $G_2(\mathbf{r}_1, \mathbf{r}_2) = \langle U_2^*(\mathbf{r}_1)\, U_2(\mathbf{r}_2)\rangle$ and use the definition $G_1(\mathbf{r}_1, \mathbf{r}_2) = \langle U_1^*(\mathbf{r}_1)\, U_1(\mathbf{r}_2)\rangle$ to obtain

$$G_2(\mathbf{r}_1, \mathbf{r}_2) = \iint h^*(\mathbf{r}_1; \mathbf{r}_1')\, h(\mathbf{r}_2; \mathbf{r}_2')\, G_1(\mathbf{r}_1', \mathbf{r}_2')\, d\mathbf{r}_1'\, d\mathbf{r}_2'. \tag{12.3-5}$$

Mutual Intensity

If the mutual intensity $G_1(\mathbf{r}_1, \mathbf{r}_2)$ of the input light, and the impulse response function $h(\mathbf{r}; \mathbf{r}')$ of the system are known, the mutual intensity of the output light $G_2(\mathbf{r}_1, \mathbf{r}_2)$ is readily determined by computing the integrals in (12.3-5).

The intensity of the output light is obtained by using the definition $I_2(\mathbf{r}) = G_2(\mathbf{r}, \mathbf{r})$, which, with the help of (12.3-5), reduces to

$$I_2(\mathbf{r}) = \iint h^*(\mathbf{r}; \mathbf{r}_1')\, h(\mathbf{r}; \mathbf{r}_2')\, G_1(\mathbf{r}_1', \mathbf{r}_2')\, d\mathbf{r}_1'\, d\mathbf{r}_2'. \tag{12.3-6}$$

Image Intensity

Thus, to determine the intensity of the output light, we must know the mutual intensity of the input light. Knowledge of the input intensity $I_1(\mathbf{r})$ by itself is generally not sufficient to determine the output intensity $I_2(\mathbf{r})$.

B. Image Formation with Incoherent Light

We now consider the special case when the input light is incoherent. The mutual intensity $G_1(\mathbf{r}_1, \mathbf{r}_2)$ then vanishes when $\mathbf{r}_2$ is only slightly separated from $\mathbf{r}_1$, so that the coherence distance is much smaller than other pertinent dimensions in the system (for example, the resolution distance of an imaging system). According to (12.1-27), the mutual intensity may then be written in the form $G_1(\mathbf{r}_1, \mathbf{r}_2) = \sqrt{I_1(\mathbf{r}_1) I_1(\mathbf{r}_2)}\, g(\mathbf{r}_1 - \mathbf{r}_2)$, where $g(\mathbf{r}_1 - \mathbf{r}_2)$ is a very narrow function. When using this expression for $G_1(\mathbf{r}_1, \mathbf{r}_2)$ in the integrals in (12.3-5) and (12.3-6), it is convenient to approximate $g(\mathbf{r}_1 - \mathbf{r}_2)$ by a delta function, $g(\mathbf{r}_1 - \mathbf{r}_2) = \sigma \delta(\mathbf{r}_1 - \mathbf{r}_2)$, where $\sigma = \int g(\mathbf{r})\, d\mathbf{r}$ is the area under $g(\mathbf{r})$, whereupon

$$G_1(\mathbf{r}_1, \mathbf{r}_2) \approx \sigma \sqrt{I_1(\mathbf{r}_1) I_1(\mathbf{r}_2)}\; \delta(\mathbf{r}_1 - \mathbf{r}_2). \tag{12.3-7}$$

Since the mutual intensity must remain finite and $\delta(0) \to \infty$, this equation cannot be used in general. However, it is suitable for evaluating integrals such as that in (12.3-6). Substituting (12.3-7) into (12.3-6), the delta function reduces the double integral to a single integral, which provides

$$\boxed{I_2(\mathbf{r}) \approx \int I_1(\mathbf{r}')\, h_i(\mathbf{r}; \mathbf{r}')\, d\mathbf{r}',} \tag{12.3-8}$$

Imaging Equation
(Incoherent Illumination)

where

$$\boxed{h_i(\mathbf{r}; \mathbf{r}') = \sigma |h(\mathbf{r}; \mathbf{r}')|^2.} \tag{12.3-9}$$

Impulse Response Function
(Incoherent Illumination)

Under these conditions, the relation between the intensities at the input and output planes describes a linear system of impulse response function $h_i(\mathbf{r}; \mathbf{r}')$, which is also known as the **point-spread function**. When the input light is completely incoherent, therefore, the intensity of the light at each point $\mathbf{r}$ on the output plane is a weighted superposition of contributions from the intensities at many points $\mathbf{r}'$ on the input plane; there is no interference and the intensities simply add (Fig. 12.3-3). This is to be contrasted with the completely coherent system, in which the complex amplitudes, rather than intensities, are related by a superposition integral, as evidenced in (12.3-4).

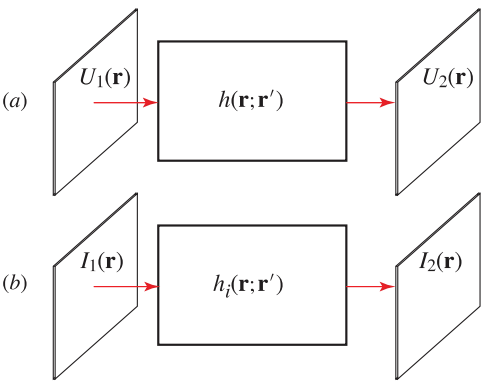


Figure 12.3-3 (*a*) The *complex amplitudes* of light at the input and output planes of an optical system illuminated by *coherent light* are related by a linear system with impulse response function $h(\mathbf{r}; \mathbf{r}')$. (*b*) The *intensities* of light at the input and output planes of an optical system illuminated by *incoherent light* are related by a linear system with impulse response function $h_i(\mathbf{r}; \mathbf{r}') = \sigma |h(\mathbf{r}; \mathbf{r}')|^2$.

In many optical systems, the impulse response function $h(\mathbf{r};\mathbf{r}')$ is a function of $\mathbf{r}-\mathbf{r}'$, say $h(\mathbf{r}-\mathbf{r}')$. The system is then said to be **shift invariant** or **isoplanatic** (see Appendix B, Sec. B.2). In this case $h_i(\mathbf{r};\mathbf{r}') = h_i(\mathbf{r}-\mathbf{r}')$. The integrals in (12.3-4) and (12.3-8) then represent two-dimensional convolutions and the systems can be described by transfer functions $H(\nu_x,\nu_y)$ and $H_i(\nu_x,\nu_y)$, which are the Fourier transforms of $h(\mathbf{r}) = h(x,y)$ and $h_i(\mathbf{r}) = h_i(x,y)$, respectively.

As an example, we apply the relations set forth above to an imaging system. The impulse response function of the single-lens focused imaging system illustrated in Fig. 12.3-4 was considered in Sec. 4.4C with coherent illumination. It was shown that, in the Fresnel approximation, it can be expressed as [see (4.4-12)]

$$h(\mathbf{r}) \propto P\left(\frac{x}{\lambda d_2}, \frac{y}{\lambda d_2}\right), \tag{12.3-10}$$

where $P(\nu_x,\nu_y)$ is the Fourier transform of the pupil function $p(x,y)$ and d_2 is the distance from the lens to the image plane. The pupil function is unity within the aperture and zero elsewhere.

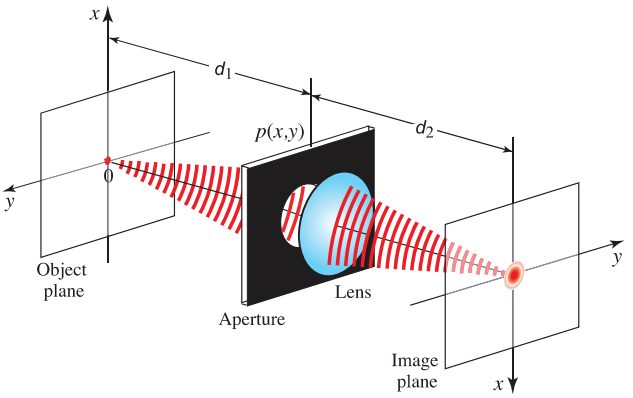


Figure 12.3-4 Single-lens imaging system.

When the illumination is quasi-monochromatic and spatially incoherent, the intensities of the light at the object and image planes are linearly related by a system with impulse response function

$$h_i(\mathbf{r}) = \sigma|h(\mathbf{r})|^2 \propto \left|P\left(\frac{x}{\lambda d_2}, \frac{y}{\lambda d_2}\right)\right|^2, \tag{12.3-11}$$

where λ is the wavelength corresponding to the central frequency ν_0.

EXAMPLE 12.3-1. ***Coherent and Incoherent Imaging Systems with Circular Apertures.*** If the aperture in Fig. 12.3-4 is a circle of radius a (diameter $D = 2a$), the pupil function $p(x,y) = 1$ for x,y inside the circle, and 0 elsewhere. Its Fourier transform is (see Appendix A, Sec. A.3)

$$P(\nu_x,\nu_y) = \frac{aJ_1(2\pi\nu_\rho a)}{\nu_\rho}, \qquad \nu_\rho = \sqrt{\nu_x^2+\nu_y^2}, \tag{12.3-12}$$

where $J_1(\cdot)$ is the Bessel function of order 1. The impulse response function of the coherent system is obtained by substituting into (12.3-10),

$$h(x, y) \propto \left[\frac{J_1(2\pi\nu_s\rho)}{\pi\nu_s\rho}\right], \qquad \rho = \sqrt{x^2 + y^2}, \tag{12.3-13}$$

where

$$\nu_s = \frac{\theta}{2\lambda}, \qquad \theta = \frac{2a}{d_2}, \tag{12.3-14}$$

which accords with the result reported in Example 4.4-1 [see (4.4-13)].

For incoherent illumination, the impulse response function is therefore

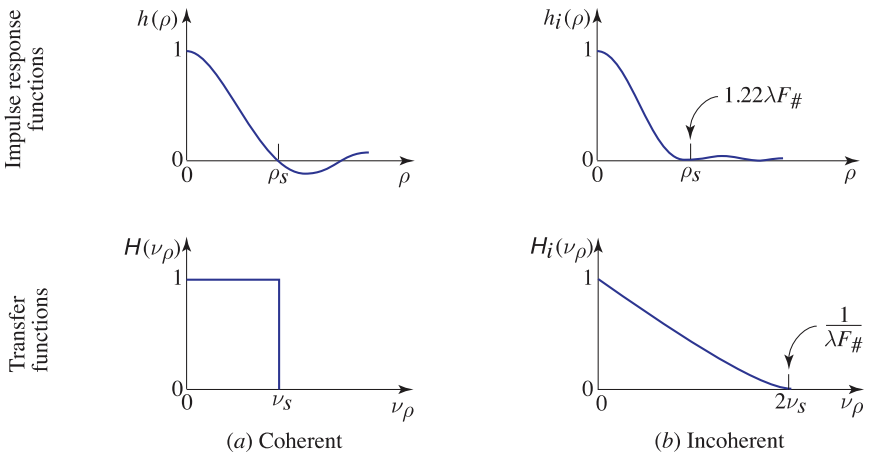
$$h_i(x, y) \propto \left[\frac{J_1(2\pi\nu_s\rho)}{\pi\nu_s\rho}\right]^2. \tag{12.3-15}$$

The impulse response functions $h(x, y)$ and $h_i(x, y)$ are illustrated in the top row of Fig. 12.3-5. Both functions reach their first zero when $2\pi\nu_s\rho = 3.832$, or $\rho = \rho_s \approx 3.832/2\pi\nu_s = 3.832\,\lambda/\pi\theta$, from which we obtain the two-point resolution

$$\boxed{\rho_s \approx 1.22\frac{\lambda}{\theta}.} \tag{12.3-16}$$

Two-Point Resolution

Thus, the image of a point (impulse) at the input plane is a patch of intensity $h_i(x, y)$ and radius ρ_s. When the input distribution comprises two points (impulses) separated by a distance ρ_s, the image of one point vanishes at the center of the image of the other point. The distance ρ_s is therefore a measure of the resolution of the imaging system.



(*a*) Coherent (*b*) Incoherent

Figure 12.3-5 Impulse response functions and transfer functions of a single-lens, focused, diffraction-limited imaging system with a circular aperture and lens $F_\#$ under (*a*) coherent, and (*b*) incoherent illumination.

The transfer functions of linear systems with the impulse response functions $h(x, y)$ and $h_i(x, y)$ are, respectively, the Fourier transforms (see Appendix A, Sec. A.3),

$$H(\nu_x, \nu_y) = \begin{cases} 1, & \nu_\rho < \nu_s \\ 0, & \text{otherwise,} \end{cases} \tag{12.3-17}$$

and

$$H_i(\nu_x, \nu_y) = \begin{cases} \dfrac{2}{\pi}\left[\left(\cos^{-1}\dfrac{\nu_\rho}{2\nu_s}\right) - \dfrac{\nu_\rho}{2\nu_s}\sqrt{1-\left(\dfrac{\nu_\rho}{2\nu_s}\right)^2}\right], & \nu_\rho < 2\nu_s \\ 0, & \text{otherwise,} \end{cases} \tag{12.3-18}$$

where $\nu_\rho = \sqrt{\nu_x^2 + \nu_y^2}$. Both functions have been normalized such that their values are unity at $\nu_\rho = 0$. These functions are illustrated in the bottom row of Fig. 12.3-5. For coherent illumination, the transfer function is flat and has a cutoff frequency $\nu_s = \theta/2\lambda$ lines/mm. For incoherent illumination, the transfer function decreases approximately linearly with the spatial frequency and has a cutoff frequency $2\nu_s = \theta/\lambda$ lines/mm.

If the object is placed at infinity so that $d_1 = \infty$, we have $d_2 = f$, where f is the focal length of the lens. The angle $\theta = 2a/f$ is then the inverse of the lens F-number since $F_\# = f/2a$. The cutoff frequencies ν_s and $2\nu_s$ are thus related to the lens F-number by [see (4.4-19)]

$$\text{Cutoff frequency (lines/mm)} = \begin{cases} \dfrac{1}{2\lambda F_\#} & \text{(coherent illumination)} \\ \dfrac{1}{\lambda F_\#} & \text{(incoherent illumination).} \end{cases} \tag{12.3-19}$$

It should not be concluded, however, that incoherent illumination is superior to coherent illumination because it has twice the spatial bandwidth. The transfer functions for the two systems should not be compared directly since one describes imaging of the complex amplitude while the other describes imaging of the intensity.

C. Gain of Spatial Coherence by Propagation

Equation (12.3-5) describes the change of the mutual intensity when the light propagates through an optical system of impulse response function $h(\mathbf{r}; \mathbf{r}')$. When the input light is incoherent, the mutual intensity $G_1(\mathbf{r}_1, \mathbf{r}_2)$ may be approximated by (12.3-7), whereupon the double integral in (12.3-5) reduces to

$$G_2(\mathbf{r}_1, \mathbf{r}_2) = \sigma \int h^*(\mathbf{r}_1; \mathbf{r})\, h(\mathbf{r}_2; \mathbf{r})\, I_1(\mathbf{r})\, d\mathbf{r}. \tag{12.3-20}$$

Mutual Intensity

It is evident that the transmitted light is no longer incoherent. In general, *light gains spatial coherence by the mere act of propagation*. This can be understood as follows. Although light fluctuations at different points of the input plane are uncorrelated, the radiation from each point spreads and overlaps with that from neighboring points. The light reaching two points at the output plane originates from many points at the input plane, some of which are common (see Fig. 12.3-6). These common contributions engender partial correlation between fluctuations at the output points.

This phenomenon is not unlike the transmission of an uncorrelated time signal (white noise) through a low-pass filter. The filter smoothes the function and reduces its spectral bandwidth, so that its coherence time increases and it is no longer uncorrelated. The propagation of light through an optical system is a form of spatial filtering that reduces the spatial bandwidth and therefore increases the coherence area.

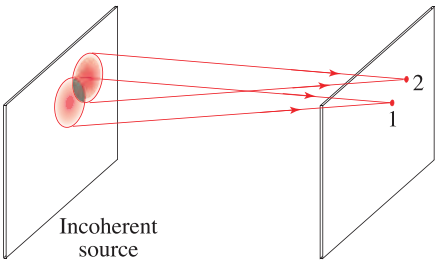


Figure 12.3-6 Gain of coherence by propagation is a result of the spreading of light. Although the light is completely uncorrelated at the source, the light fluctuations at points 1 and 2 share a common origin, the dark shaded area, and are therefore partially correlated.

van Cittert–Zernike Theorem

There is a mathematical identity between the expressions for the gain of *coherence of initially incoherent light* propagating through an optical system, and the change of the *amplitude of coherent light* traveling through the same system. With respect to (12.3-20), if the observation point $\mathbf{r}_1$ is fixed at the origin $\mathbf{0}$, for example, and the mutual intensity $G_2(\mathbf{0}, \mathbf{r}_2)$ is examined as a function of $\mathbf{r}_2$, then

$$G_2(\mathbf{0}, \mathbf{r}_2) = \sigma \int h^*(\mathbf{0}; \mathbf{r})\, h(\mathbf{r}_2; \mathbf{r})\, I_1(\mathbf{r})\, d\mathbf{r}. \tag{12.3-21}$$

Defining $U_2(\mathbf{r}_2) = G_2(\mathbf{0}, \mathbf{r}_2)$ and $U_1(\mathbf{r}) = \sigma h^*(\mathbf{0}; \mathbf{r}) I_1(\mathbf{r})$, (12.3-21) may be written in the familiar form

$$U_2(\mathbf{r}_2) = \int h(\mathbf{r}_2; \mathbf{r})\, U_1(\mathbf{r})\, d\mathbf{r}, \tag{12.3-22}$$

which is nothing other than the integral (12.3-4) that governs the propagation of coherent light. Thus, the observed mutual intensity $G(\mathbf{0}, \mathbf{r}_2)$ at the output of an optical system whose input is incoherent is mathematically identical to the observed complex amplitude of a coherent wave $U_1(\mathbf{r}) = \sigma h^*(\mathbf{0}; \mathbf{r}) I_1(\mathbf{r})$ presented at the input of the same system.

As an example, assume that the incoherent input wave has uniform intensity and extends over an aperture $p(\mathbf{r})$ [$p(\mathbf{r}) = 1$ within the aperture and zero elsewhere] so that $I_1(\mathbf{r}) = p(\mathbf{r})$; and assume also that the optical system is free space so that $h(\mathbf{r}'; \mathbf{r}) = \exp(-jk|\mathbf{r}' - \mathbf{r}|)/|\mathbf{r}' - \mathbf{r}|$. The mutual intensity $G_2(\mathbf{0}, \mathbf{r}_2)$ is then identical to the amplitude $U_2(\mathbf{r}_2)$ obtained when a coherent wave with input amplitude $U_1(\mathbf{r}) = \sigma h^*(\mathbf{0}; \mathbf{r}) p(\mathbf{r}) = \sigma p(\mathbf{r}) \exp(jkr)/r$ is transmitted through the same system. This is a spherical wave converging to the point $\mathbf{0}$ at the output plane and transmitted through the aperture.

This connection between the gain of spatial coherence of incoherent light, and the diffraction of coherent light, traveling through the same system is known as the **Van Cittert–Zernike theorem**.

Gain of Coherence in Free Space

Consider an optical system comprising free-space propagation between two parallel planes separated by a distance d (Fig. 12.3-7). Light at the input plane is quasi-monochromatic and spatially incoherent, and has intensity $I(x, y)$ extending over a finite area. The impulse response function is then described by the Fresnel diffraction formula [see (4.1-18)]:

$$h(\mathbf{r}; \mathbf{r}') = h_0 \exp\left[-j\pi \frac{(x - x')^2 + (y - y')^2}{\lambda d}\right], \tag{12.3-23}$$

where $\mathbf{r} = (x, y, d)$ and $\mathbf{r}' = (x', y', 0)$ are the coordinates of points at the output and input planes, respectively, and $h_0 = (j/\lambda d)\exp(-j2\pi d/\lambda)$ is a constant.

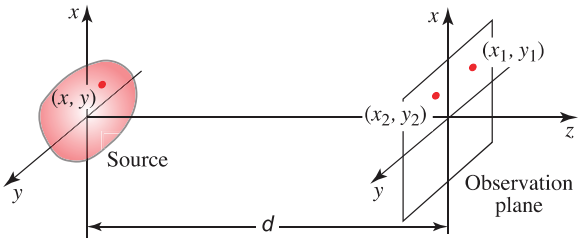


Figure 12.3-7 Radiation from an incoherent source in free space.

We determine the mutual coherence function $G(x_1, y_1, x_2, y_2)$ at two points (x_1, y_1) and (x_2, y_2) in the output plane, by substituting (12.3-23) into (12.3-20) to obtain

$$|G(x_1, y_1, x_2, y_2)| = \sigma_1 \left| \iint_{-\infty}^{\infty} \exp\left\{ j\frac{2\pi}{\lambda d}\left[(x_2 - x_1)x + (y_2 - y_1)y\right]\right\} I(x, y)\, dx\, dy \right|, \tag{12.3-24}$$

where $\sigma_1 = \sigma|h_0|^2 = \sigma/\lambda^2 d^2$ is another constant. Given $I(x, y)$, one can easily determine $|G(x_1, y_1, x_2, y_2)|$ in terms of the two-dimensional Fourier transform of $I(x, y)$,

$$\mathcal{I}(\nu_x, \nu_y) = \iint_{-\infty}^{\infty} \exp[j2\pi(\nu_x x + \nu_y y)]\, I(x, y)\, dx\, dy, \tag{12.3-25}$$

evaluated at $\nu_x = (x_2 - x_1)/\lambda d$ and $\nu_y = (y_2 - y_1)/\lambda d$. The magnitude of the corresponding normalized mutual intensity is then

$$\boxed{|g(x_1, y_1, x_2, y_2)| = \left| \mathcal{I}\left(\frac{x_2 - x_1}{\lambda d}, \frac{y_2 - y_1}{\lambda d}\right)\right| \Big/ \mathcal{I}(0, 0).} \tag{12.3-26}$$

This Fourier transform-relation between the intensity profile of an incoherent source and the degree of spatial coherence of its far field is similar to the Fourier-transform relation between the amplitude of coherent light at the input and output planes (see Sec. 4.2A). The similarity is expected in view of the Van Cittert–Zernike theorem.

The implications of (12.3-26) are profound. If the area of the source, i.e., the spatial extent of $I(x, y)$, is small, its Fourier transform $\mathcal{I}(\nu_x, \nu_y)$ is wide, so that the mutual intensity at the output plane extends over a wide area and the area of coherence at the output plane is large. In the extreme limit in which light at the input plane originates from a point, the area of coherence is infinite and the radiated field is spatially completely coherent. This confirms our earlier discussions in Sec. 12.1D with respect to the coherence of spherical waves. On the other hand, if the input light originates from a large extended source, the propagated light has a small area of coherence.

EXAMPLE 12.3-2. ***Radiation from an Incoherent Circular Source.*** For input light with uniform intensity $I(x, y) = I_0$ confined to a circular aperture of radius a, (12.3-26) yields

$$|g(x_1, y_1, x_2, y_2)| = \left| \frac{2J_1(\pi\rho\theta_s/\lambda)}{\pi\rho\theta_s/\lambda} \right|, \tag{12.3-27}$$

where $\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ is the distance between the two points, $\theta_s = 2a/d$ is the angle subtended by the source, and $J_1(\cdot)$ is the Bessel function. This relation is plotted in Fig. 12.3-8. The Bessel function reaches its first zero when its argument is 3.832. We can therefore define the area of coherence as a circle of radius $\rho_c = 3.832(\lambda/\pi\theta_s)$, so that

$$\boxed{\rho_c = 1.22\frac{\lambda}{\theta_s}.} \tag{12.3-28}$$

Coherence Distance

A similar result, (12.2-18), was obtained using a less rigorous analysis. The area of coherence is inversely proportional to θ_s^2. An incoherent light source of wavelength $\lambda = 0.6$ μm and radius 1 cm observed at a distance $d = 100$ m, for example, has a coherence distance $\rho_c \approx 3.7$ mm.

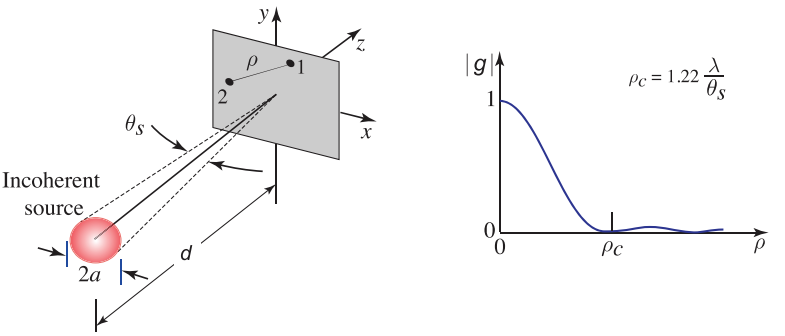


Figure 12.3-8 The magnitude of the degree of spatial coherence of light radiated from an incoherent circular light source subtending an angle θ_s, as a function of the separation ρ.

Measurement of the Angular Diameter of Stars: The Michelson Stellar Interferometer

Equation (12.3-28) is the basis of a method for measuring the angular diameters of stars. If the star is regarded as an incoherent disk of diameter $2a$ with uniform brilliance, then at an observation plane a distance d away from the star, the coherence function decreases to 0 when the separation between the two observation points reaches $\rho_c = 1.22\lambda/\theta_s$. Measuring ρ_c at a given value of λ permits us to determine the angular diameter $\theta_s = 2a/d$.

As an example, taking the angular diameter of the sun to be $0.5°$, $\theta_s = 8.7 \times 10^{-3}$ radians, and assuming that the intensity is uniform, we obtain $\rho_c = 140\lambda$. For $\lambda = 0.5$ μm, we have $\rho_c = 70$ μm. To observe interference fringes in a Young double-slit apparatus, the slits would have to be separated by a distance smaller than 70 μm. Stars of smaller angular diameter have correspondingly larger areas of coherence. For example, the first star whose angular diameter was measured using this technique (α-Orion) has an angular diameter $\theta_s = 22.6 \times 10^{-8}$, so that at $\lambda = 0.57$ μm, we have $\rho_c = 3.1$ m. A Young interferometer can be modified to accommodate such large slit separations by using movable mirrors, as shown in Fig. 12.3-9.

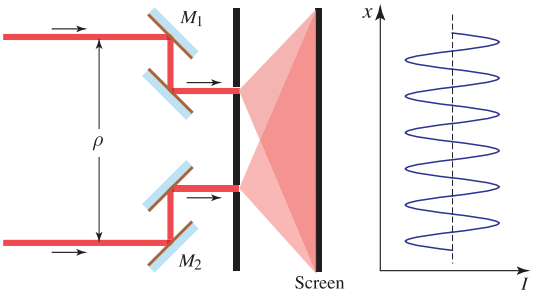


Figure 12.3-9 The Michelson stellar interferometer. The angular diameter of a star is estimated by measuring the mutual intensity at two points with variable separation ρ using a Young double-slit interferometer. The distance ρ between mirrors M_1 and M_2 is varied and the visibility of the interference fringes is measured. When $\rho = \rho_c = 1.22\lambda/\theta_s$, the visibility is 0.

12.4 PARTIAL POLARIZATION

As we have seen in Chapter 6, the scalar wave theory of light is often inadequate and a vector theory that includes the polarization of light is required. This section provides a brief discussion of the statistical theory of *random* light, including the effects of polarization. The **theory of partial polarization** is based on characterizing the components of the optical field vector by correlations and cross-correlations similar to those defined earlier in this chapter.

To simplify the presentation, we shall not be concerned with spatial effects. We therefore limit ourselves to light described by a transverse electromagnetic (TEM) plane wave traveling in the z direction. The electric-field vector has two components, in the x and y directions, with complex wavefunctions $U_x(t)$ and $U_y(t)$ that are generally random. Each function is characterized by its autocorrelation function (the temporal coherence function),

$$G_{xx}(\tau) = \langle U_x^*(t)U_x(t+\tau)\rangle \tag{12.4-1}$$

$$G_{yy}(\tau) = \langle U_y^*(t)U_y(t+\tau)\rangle. \tag{12.4-2}$$

An additional descriptor of the wave is the cross-correlation function of $U_x(t)$ and $U_y(t)$,

$$G_{xy}(\tau) = \langle U_x^*(t)U_y(t+\tau)\rangle. \tag{12.4-3}$$

The normalized function

$$g_{xy}(\tau) = \frac{G_{xy}(\tau)}{\sqrt{G_{xx}(0)G_{yy}(0)}} \tag{12.4-4}$$

is the cross-correlation coefficient of $U_x(t)$ and $U_y(t+\tau)$. It satisfies the inequality $0 \le |g_{xy}(\tau)| \le 1$. When the two components are uncorrelated at all times, $|g_{xy}(\tau)| = 0$; when they are completely correlated at all times, $|g_{xy}(\tau)| = 1$.

The spectral properties are, in general, tied to the polarization properties and the autocorrelation and cross-correlation functions can have different dependencies on τ. However, for quasi-monochromatic light, all dependencies on τ in (12.4-1)–(12.4-4) are approximately of the form $\exp(j2\pi\nu_0\tau)$, so that the polarization properties are described by the values at $\tau = 0$. The three numbers $G_{xx}(0)$, $G_{yy}(0)$, and $G_{xy}(0)$, hereafter denoted G_{xx}, G_{yy}, and G_{xy}, are then used to describe the polarization of the wave. Note that $G_{xx} = I_x$ and $G_{yy} = I_y$ are real numbers that represent the intensities of the x and y components, but G_{xy} is complex and $G_{yx} = G_{xy}^*$, as can easily be verified from the definition.

Coherency Matrix

It is convenient to write the four variables G_{xx}, G_{xy}, G_{yx}, and G_{yy} in the form of a 2×2 Hermitian matrix

$$\mathbf{G} = \begin{bmatrix} G_{xx} & G_{xy} \\ G_{yx} & G_{yy} \end{bmatrix} \tag{12.4-5}$$

called the **coherency matrix**. The diagonal elements are the intensities I_x and I_y, whereas the off-diagonal elements are the cross-correlations. The trace of the matrix, given by $\text{Tr}\,\mathbf{G} = I_x + I_y \equiv \bar{I}$, is the total intensity.

The coherency matrix may also be written in terms of the Jones vector, $\mathbf{J} = \begin{bmatrix} U_x \\ U_y \end{bmatrix}$ defined in terms of the complex wavefunctions and complex amplitudes (instead of in terms of the complex envelopes as in Sec. 6.1B),

$$\langle \mathbf{J}^* \mathbf{J}^\mathsf{T} \rangle = \left\langle \begin{bmatrix} U_x^* \\ U_y^* \end{bmatrix} \begin{bmatrix} U_x & U_y \end{bmatrix} \right\rangle = \begin{bmatrix} \langle U_x^* U_x \rangle & \langle U_x^* U_y \rangle \\ \langle U_y^* U_x \rangle & \langle U_y^* U_y \rangle \end{bmatrix} = \mathbf{G}, \tag{12.4-6}$$

where the superscript T denotes the transpose of a matrix, and U_x and U_y denote $U_x(t)$ and $U_y(t)$, respectively.

The Jones vector is transformed by polarization devices, such as polarizers and retarders, in accordance with the rule $\mathbf{J}' = \mathbf{T}\mathbf{J}$ [see (6.1-17)], where $\mathbf{T}$ is the Jones matrix representing the device [see (6.1-18) to (6.1-25)]. The coherency matrix is therefore transformed in accordance with $\mathbf{G}' = \langle \mathbf{T}^* \mathbf{J}^* (\mathbf{T}\mathbf{J})^\mathsf{T} \rangle = \langle \mathbf{T}^* \mathbf{J}^* \mathbf{J}^\mathsf{T} \mathbf{T}^\mathsf{T} \rangle = \mathbf{T}^* \langle \mathbf{J}^* \mathbf{J}^\mathsf{T} \rangle \mathbf{T}^\mathsf{T}$, so that

$$\boxed{\mathbf{G}' = \mathbf{T}^* \mathbf{G} \mathbf{T}^\mathsf{T}.} \tag{12.4-7}$$

We thus have a formalism for determining the effect of polarization devices on the coherency matrix of partially polarized light.

Stokes Parameters and Poincaré-Sphere Representation

The Stokes parameters were defined in Sec. 6.1A for coherent light as a set of four real parameters related to the products of the x and y components of the complex envelope [see (6.1-9)]. This definition is readily generalized to partially coherent light as an average of these products:

$$\boxed{\begin{aligned} \mathrm{S}_0 &= \langle |U_x|^2 \rangle + \langle |U_y|^2 \rangle = G_{xx} + G_{yy} && \text{(12.4-8a)} \\ \mathrm{S}_1 &= \langle |U_x|^2 \rangle - \langle |U_y|^2 \rangle = G_{xx} - G_{yy} && \text{(12.4-8b)} \\ \mathrm{S}_2 &= 2\,\text{Re}\{\langle U_x^* U_y \rangle\} = 2\,\text{Re}\{G_{xy}\} && \text{(12.4-8c)} \\ \mathrm{S}_3 &= 2\,\text{Im}\{\langle U_x^* U_y \rangle\} = 2\,\text{Im}\{G_{xy}\}. && \text{(12.4-8d)} \end{aligned}}$$

Stokes Parameters

Thus, the Stokes parameters are directly related to elements of the coherency matrix $\mathbf{G}$. The first parameter, S_0, is simply the sum of the diagonal elements, which is the total intensity $\bar{I}$. The second, S_1, is the difference of the diagonal elements, i.e., the difference between the intensities of the two polarization components. The third and

fourth, S_2 and S_3, respectively, are proportional to the real and imaginary parts of the off-diagonal element, i.e., the cross-correlation function. Using these relations, it can be readily shown that the inequality $|G_{xy}|^2 \leq G_{xx}G_{yy}$ leads to the condition $\mathrm{S}_1^2 + \mathrm{S}_2^2 + \mathrm{S}_3^2 \leq \mathrm{S}_0^2$. For coherent light, these inequalities become equalities.

The state of polarization of partially polarized light may be represented geometrically on the Poincaré sphere (Fig. 6.1-5) as a point whose Cartesian coordinates are $(\mathrm{S}_1/\mathrm{S}_0, \mathrm{S}_2/\mathrm{S}_0, \mathrm{S}_3/\mathrm{S}_0)$. Since $\mathrm{S}_1^2 + \mathrm{S}_2^2 + \mathrm{S}_3^2 \leq \mathrm{S}_0^2$, such a point lies inside, or on, the surface of the sphere.

To understand the significance of the coherency matrix and the Stokes parameters, we next examine two limiting cases.

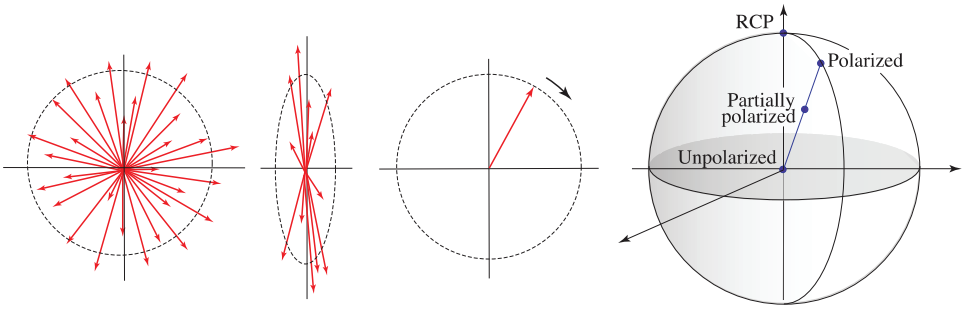
Unpolarized Light

Light of intensity $\bar{I}$ is said to be **unpolarized** if its two components have the same intensity and are uncorrelated, i.e., $I_x = I_y \equiv \frac{1}{2}\bar{I}$ and $G_{xy} = 0$. The coherency matrix is then

$$\mathbf{G} = \tfrac{1}{2}\bar{I}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \qquad \text{(12.4-9)}$$

Unpolarized Light

By use of (12.4-7) and (6.1-22), it can be shown that (12.4-9) is invariant to rotation of the coordinate system, so that the two components always have equal intensities and are uncorrelated. Unpolarized light therefore has an electric-field vector that is statistically isotropic; it is equally likely to have any direction in the x–y plane, as illustrated in Fig. 12.4-1(a). Analogous results for partially polarized and right-circularly polarized (RCP) light are displayed in Figs. 12.4-1(b) and 12.4-1(c), respectively.



(a) Unpolarized (b) Partially polarized (c) Polarized (RCP) (d) Poincaré sphere

Figure 12.4-1 Fluctuations of the electric-field vector for (a) unpolarized light, (b) partially polarized light, and (c) polarized light with circular polarization. (d) Poincaré-sphere representation for unpolarized light (at the origin), partially polarized light (in the interior), and elliptically polarized light (on the surface).

When passed through a polarizer, unpolarized light becomes linearly polarized, but it remains random with an average intensity $\frac{1}{2}\bar{I}$. A wave retarder has no effect on unpolarized light since it only introduces a phase shift between two components that have a totally random phase to begin with. Similarly, unpolarized light transmitted through a polarization rotator remains unpolarized. These effects may be formally derived by use of (12.4-7) and (12.4-9) together with (6.1-18), (6.1-19), and (6.1-20), respectively.

The Stokes parameters describing unpolarized light are $(\mathrm{S}_0, \mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3) = (\bar{I}, 0, 0, 0)$ as can be readily shown by use of (12.4-8) and (12.4-9). The corresponding representation on the Poincaré sphere is a point with Cartesian coordinates $(\mathrm{S}_1/\mathrm{S}_0, \mathrm{S}_2/\mathrm{S}_0, \mathrm{S}_3/\mathrm{S}_0)$ = $(0, 0, 0)$ so that the point is located at the origin of the sphere [Fig. 12.4-1(d)].

Polarized Light

If the cross-correlation coefficient $g_{xy} = G_{xy}/\sqrt{I_x I_y}$ has unity magnitude, $|g_{xy}| = 1$, the two components of the optical field are perfectly correlated and the light is said to be completely polarized (or simply **polarized**). The coherency matrix then takes the form

$$\mathbf{G} = \begin{bmatrix} I_x & \sqrt{I_x I_y}\, e^{j\varphi} \\ \sqrt{I_x I_y}\, e^{-j\varphi} & I_y \end{bmatrix}, \tag{12.4-10}$$

where φ is the argument of g_{xy}. Defining $U_x = \sqrt{I_x}$ and $U_y = \sqrt{I_y}\, e^{j\varphi}$, we have

$$\mathbf{G} = \begin{bmatrix} U_x^* U_x & U_x^* U_y \\ U_y^* U_x & U_y^* U_y \end{bmatrix} = \mathbf{J}^* \mathbf{J}^{\mathsf{T}}, \tag{12.4-11}$$

where $\mathbf{J}$ is a Jones matrix with components U_x and U_y. Thus, $\mathbf{G}$ has the same form as the coherency matrix of a coherent wave. Using the Jones vectors provided in Table 6.1-1, we can determine the coherency matrices for different states of polarization. Two examples are:

Linearly polarized in the x direction:	$\mathbf{G} = \bar{I} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$	Right-circularly polarized:	$\mathbf{G} = \frac{1}{2}\bar{I} \begin{bmatrix} 1 & j \\ -j & 1 \end{bmatrix}$

The Stokes parameters corresponding to (12.4-11) satisfy the relation $\mathrm{S}_1^2 + \mathrm{S}_2^2 + \mathrm{S}_3^2 = \mathrm{S}_0^2$, so that polarized light is represented by a point on the surface of, rather than inside, the Poincaré sphere [Fig. 12.4-1(d)].

It is instructive to examine the distinction between unpolarized light and circularly polarized light. In both cases the intensities of the x and y components are equal ($I_x = I_y$). For circularly polarized light the two components are completely correlated, but for unpolarized light they are uncorrelated. Circularly polarized light may be transformed into linearly polarized light by the use of a wave retarder, but unpolarized light remains unpolarized upon passage through such a device. Circularly polarized light is represented by a point at the north or south pole of the Poincaré sphere, while unpolarized light is represented by a point at the origin [Fig. 12.4-1(d)].

Degree of Polarization

Partial polarization is a general state of random polarization that lies between the two ideal limits of unpolarized and polarized light. One measure of the **degree of polarization** $\mathbb{P}$ is defined in terms of the determinant and the trace of the coherency matrix:

$$\mathbb{P} = \sqrt{1 - \frac{4 \det \mathbf{G}}{(\mathrm{Tr}\, \mathbf{G})^2}} \tag{12.4-12}$$

$$= \sqrt{1 - 4\left[\frac{I_x I_y}{(I_x + I_y)^2}\right](1 - |g_{xy}|^2)}\,. \tag{12.4-13}$$

This measure is meaningful because of the following considerations:

- It satisfies the inequality $0 \leq \mathbb{P} \leq 1$.
- For polarized light, $\mathbb{P}$ has its highest value of 1, as can readily be seen by substituting $|g_{xy}| = 1$ into (12.4-13). For unpolarized light it has its lowest value $\mathbb{P} = 0$, since $I_x = I_y$ and $g_{xy} = 0$.
- It is invariant to rotation of the coordinate system (since the determinant and the trace of a matrix are invariant to unitary transformations).
- The degree of polarization in (12.4-13) may also be expressed in terms of the Stokes parameters as:

$$\mathbb{P} = \frac{\sqrt{\mathrm{S}_1^2 + \mathrm{S}_2^2 + \mathrm{S}_3^2}}{\mathrm{S}_0}, \tag{12.4-14}$$

  so that in the Poincaré-sphere representation [Fig. 12.4-1(d)], it is equal to the distance from the origin of the sphere.
- It can be shown (Exercise 12.4-1) that a partially polarized wave can always be regarded as a mixture of two uncorrelated waves: a completely polarized wave and an unpolarized wave, with the ratio of the intensity of the polarized component to the total intensity equal to the degree of polarization $\mathbb{P}$.

EXERCISE 12.4-1

Partially Polarized Light. Demonstrate that the superposition of unpolarized light of intensity $(I_x + I_y)(1 - \mathbb{P})$, and linearly polarized light with intensity $(I_x + I_y)\mathbb{P}$, where $\mathbb{P}$ is given by (12.4-13), yields light whose x and y components have intensities I_x and I_y and normalized cross-correlation $|g_{xy}|$.

READING LIST

Statistical Optics, Coherence, and Partial Polarization

J. J. Gil Pérez and R. Ossikovski, *Polarized Light and the Mueller Matrix Approach*, CRC Press/Taylor & Francis, 2016.

J. W. Goodman, *Statistical Optics*, Wiley, 2nd ed. 2015.

E. Wolf, *Introduction to the Theory of Coherence and Polarization of Light*, Cambridge University Press, 2007.

O. Marchenko, S. Kazantsev, and L. Windholz, *Demonstrational Optics: Part 2, Coherent and Statistical Optics*, Springer-Verlag, 2007.

M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th expanded and corrected ed. 2002, Chapter 10.

B. R. Frieden, *Probability, Statistical Optics, and Data Testing: A Problem Solving Approach*, Springer-Verlag, 1983, 3rd ed. 2001.

C. Brosseau, *Fundamentals of Polarized Light: A Statistical Optics Approach*, Wiley, 1998.

L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge University Press, 1995.

L. Mandel and E. Wolf, eds., *Selected Papers on Coherence and Fluctuations of Light (1850–1966)*, SPIE Optical Engineering Press (Milestone Series Volume 19), 1990.

M. C. Teich and B. E. A. Saleh, Photon Bunching and Antibunching, in E. Wolf, ed., *Progress in Optics*, North-Holland, 1988, vol. 26, pp. 1–104.

J. Peřina, *Coherence of Light*, Reidel, 1971, 2nd ed. 1985.

B. E. A. Saleh, *Photoelectron Statistics with Applications to Spectroscopy and Optical Communication*, Springer-Verlag, 1978.

B. Crosignani, P. Di Porto, and M. Bertolotti, *Statistical Properties of Scattered Light*, Academic Press, 1975.

R. Hanbury-Brown, *The Intensity Interferometer: Its Application to Astronomy*, Taylor & Francis, 1974.

L. Mandel and E. Wolf, eds., *Selected Papers on Coherence and Fluctuations of Light*, Volumes 1 and 2, Dover, 1970.

M. J. Beran and G. B. Parrent, Jr., *Theory of Partial Coherence*, Prentice Hall, 1964; SPIE Optical Engineering Press, reissued 1974.

E. L. O'Neill, *Introduction to Statistical Optics*, Addison–Wesley, 1963; Dover, reissued 2003.

Imaging

D. F. Buscher, *Practical Optical Interferometry: Imaging at Visible and Infrared Wavelengths*, Cambridge University Press, 2015.

A. Donges and R. Noll, *Laser Measurement Technology: Fundamentals and Applications*, Springer-Verlag, 2015.

W. Drexler and J. G. Fujimoto, eds., *Optical Coherence Tomography: Technology and Applications*, Volumes. 1–3, Springer-Verlag, 2nd ed. 2015.

M. C. Teich, B. E. A. Saleh, F. N. C. Wong, and J. H. Shapiro, Variations on the Theme of Quantum Optical Coherence Tomography: A Review, *Quantum Information Processing*, vol. 11, pp. 903–923, 2012.

B. E. A. Saleh, *Introduction to Subsurface Imaging*, Cambridge University Press, 2011.

D. J. Brady, *Optical Imaging and Spectroscopy*, Wiley, 2009.

M. E. Brezinski, *Optical Coherence Tomography: Principles and Applications*, Academic Press/Elsevier, 2006.

H. H. Barrett and K. J. Myers, *Foundations of Image Science*, Wiley, 2004.

Random Processes, Random Media, and Speckle

B. Hajek, *Random Processes for Engineers*, Cambridge University Press, 2015.

V. Krishnan and K. Chandra, *Probability and Random Processes*, Wiley, 2nd ed. 2015.

R. G. Gallager, *Stochastic Processes: Theory for Applications*, Cambridge University Press, 2014.

H. Pishro-Nik, *Introduction to Probability, Statistics, and Random Processes*, Kappa Research, 2014.

O. Korotkova, *Random Light Beams: Theory and Applications*, CRC Press/Taylor & Francis, 2014.

N. Pinel and C. Bourlier, *Electromagnetic Wave Scattering from Random Rough Surfaces: Asymptotic Models*, Wiley–ISTE, 2013.

P. Bajorski, *Statistics for Imaging, Optics, and Photonics*, Wiley, 2012.

J. W. Goodman, *Speckle Phenomena in Optics: Theory and Applications*, Roberts, 2007, paperback ed. 2010.

L. C. Andrews and R. L. Phillips, *Laser Beam Propagation through Random Media*, SPIE Optical Engineering Press, 2nd ed. 2005.

A. A. Kokhanovsky, *Polarization Optics of Random Media*, Springer-Verlag/Praxis, 2003.

A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw–Hill, 1965, 4th ed. 2002.

B. R. Frieden, *Probability, Statistical Optics, and Data Testing: A Problem Solving Approach*, Springer-Verlag, 1983, 3rd ed. 2001.

H. E. Rowe, *Electromagnetic Propagation in Multi-Mode Random Media*, Wiley, 1999.

P. Meinlschmidt, K. D. Hinsch, and R. S. Sirohi, eds., *Selected Papers on Electronic Speckle Pattern Interferometry: Principles and Practice*, SPIE Optical Engineering Press (Milestone Series Volume 132), 1996.

R. S. Sirohi, ed., *Selected Papers on Speckle Metrology*, SPIE Optical Engineering Press (Milestone Series Volume 35), 1991.

M. Françon, *Laser Speckle and Application in Optics*, Academic Press, 1979.

A. Ishimaru, *Wave Propagation and Scattering in Random Media*, Volume 1 and Volume 2, Academic Press, 1978; IEEE Press/Oxford University Press, reissued 1997.

PROBLEMS

12.1-4 **Lorentzian Spectrum.** A light-emitting diode (LED) emits light of Lorentzian spectrum with a linewidth $\Delta\nu$ (FWHM) $= 10^{13}$ Hz centered about a frequency corresponding to a wavelength $\lambda_o = 0.7$ μm. Determine the linewidth $\Delta\lambda_o$ (in units of nm), the coherence time τ_c, and the coherence length l_c. What is the maximum time delay within which the magnitude of the complex degree of temporal coherence $|g(\tau)|$ is greater than 0.5?

12.1-5 **Proof of the Wiener–Khinchin Theorem.** Use the definitions in (12.1-4), (12.1-14), and (12.1-15) to prove that the spectral density $S(\nu)$ is the Fourier transform of the autocorrelation function $G(\tau)$. Prove that the intensity I is the integral of the power spectral density $S(\nu)$.

12.1-6 **Mutual Intensity.** The mutual intensity of an optical wave at points on the x axis is given by

$$G(x_1, x_2) = I_0 \exp\left[-\frac{(x_1^2 + x_2^2)}{W_0^2}\right] \exp\left[\frac{-(x_1 - x_2)^2}{\rho_c^2}\right],$$

where I_0, W_0, and ρ_c are constants. Sketch the intensity distribution as a function of x. Derive an expression for the normalized mutual intensity $g(x_1, x_2)$ and sketch it as a function of $x_1 - x_2$. What is the physical meaning of the parameters I_0, W_0, and ρ_c?

12.1-7 **Mutual Coherence Function.** An optical wave has a mutual coherence function at points on the x axis,

$$G(x_1, x_2, \tau) = \exp\left(-\frac{\pi\tau^2}{2\tau_c^2}\right) \exp[j2\pi u(x_1, x_2)\tau] \exp\left[-\frac{(x_1 - x_2)^2}{\rho_c^2}\right],$$

where $u(x_1, x_2) = 5 \times 10^{14}\text{s}^{-1}$ for $x_1 + x_2 > 0$, and $6 \times 10^{14}\text{s}^{-1}$ for $x_1 + x_2 < 0$, $\rho_c = 1$ mm, and $\tau_c = 1$ μs. Determine the intensity, the power spectral density, the coherence length, and the coherence distance in the transverse plane. Which of these quantities is position dependent? If this wave were recorded on color film, what would the recorded image look like?

12.1-8 **Coherence Length.** Show that light of narrow spectral width has a coherence length $l_c \approx \lambda^2/\Delta\lambda$, where $\Delta\lambda$ is the linewidth in wavelength units. Show that for light of broad uniform spectrum extending between the wavelengths $\lambda_{\min}$ and $\lambda_{\max} = 2\lambda_{\min}$, the coherence length $l_c = \lambda_{\max}$.

12.1-9 **Effect of Spectral Width on Spatial Coherence.** A point source at the origin $(0, 0, 0)$ of a Cartesian coordinate system emits light with a Lorentzian spectrum and coherence time $\tau_c = 10$ ps. Determine an expression for the normalized mutual intensity of the light at the points $(0, 0, d)$ and $(x, 0, d)$, where $d = 10$ cm. Sketch the magnitude of the normalized mutual intensity as a function of x.

12.1-10 **Gaussian Mutual Intensity.** An optical wave in free space has a mutual coherence function $G(\mathbf{r}_1, \mathbf{r}_2, \tau) = J(\mathbf{r}_1 - \mathbf{r}_2)\exp(j2\pi\nu_0\tau)$.

(a) Show that the function $J(\mathbf{r})$ must satisfy the Helmholtz equation $\nabla^2 J + k_o^2 J = 0$, where $k_o = 2\pi\nu_0/c$.

(b) An approximate solution of the Helmholtz equation is the Gaussian-beam solution

$$J(\mathbf{r}) = \frac{1}{q(z)} \exp\left[-\frac{jk_o(x^2 + y^2)}{2q(z)}\right] \exp(-jk_o z),$$

where $q(z) = z + jz_0$ and z_0 is a constant. This solution has been studied extensively in Chapter 3 in connection with Gaussian beams. Determine an expression for the coherence area near the z axis and show that it increases with $|z|$, so that the wave gains coherence with propagation away from the origin.

12.2-1 **Effect of Spectral Width on Fringe Visibility.** Light from a sodium lamp of Lorentzian spectral linewidth $\Delta\nu = 5 \times 10^{11}$ Hz is used in a Michelson interferometer. Determine the maximum pathlength difference for which the visibility of the interferogram $\mathcal{V} > \frac{1}{2}$.

12.2-2 **Number of Observable Fringes in Young's Interferometer.** Determine the number of observable fringes in Young's interferometer if each of the sources in Table 12.1-2 is used. Assume full spatial coherence in all cases.

12.2-3 **Spectrum of a Superposition of Two Waves.** An optical wave is a superposition of two waves $U_1(t)$ and $U_2(t)$ with identical spectra $S_1(\nu) = S_2(\nu)$, which are Gaussian with spectral width $\Delta\nu$ and central frequency ν_0. The waves are not necessarily uncorrelated. Determine an expression for the power spectral density $S(\nu)$ of the superposition $U(t) = U_1(t) + U_2(t)$. Explore the possibility that $S(\nu)$ is also Gaussian, with a shifted central frequency $\nu_1 \neq \nu_0$. If this were possible, our faith in using the Doppler shift as a method to determine the velocity of stars would be shaken, since frequency shifts could originate from something other than the Doppler effect.

*12.3-1 **Partially Coherent Gaussian Beam.** A quasi-monochromatic light wave of wavelength λ travels in free space in the z direction. Its intensity in the $z = 0$ plane is a Gaussian function $I(x) = I_0 \exp(-2x^2/W_0^2)$ and its normalized mutual intensity is also a Gaussian function $g(x_1, x_2) = \exp[-(x_1 - x_2)^2/\rho_c^2]$. Show that the intensity at a distance z satisfying conditions of the Fraunhofer approximation is also a Gaussian function $I_z(x) \propto \exp[-2x^2/W^2(z)]$ and derive an expression for the beam width $W(z)$ as a function of z and the parameters W_0, ρ_c, and λ. Discuss the effect of spatial coherence on beam divergence.

*12.3-2 **Fourier-Transform Lens.** Quasi-monochromatic spatially incoherent light of uniform intensity illuminates a transparency of intensity transmittance $f(x, y)$ and the emerging light is transmitted between the front and back focal planes of a lens. Determine an expression for the intensity of the observed light. Compare your results with the case of coherent light in which the lens performs the Fourier transform (see Sec. 4.2).

*12.3-3 **Light from Two-Point Incoherent Source.** A spatially incoherent quasi-monochromatic source of light emits only at two points separated by a distance $2a$. Determine an expression for the normalized mutual intensity at a distance d from the source (use the Fraunhofer approximation).

*12.3-4 **Coherence of Light Transmitted Through a Fourier-Transform Optical System.** Light from a quasi-monochromatic spatially incoherent source with uniform intensity is transmitted through a thin slit of width $2a$ and travels between the front and back focal planes of a lens. Determine an expression for the normalized mutual intensity in the back focal plane.

12.4-2 **Partially Polarized Light.** The intensities of the two components of a partially polarized wave are $I_x = I_y = \frac{1}{2}$, and the argument of the cross-correlation coefficient g_{xy} is $\pi/2$.

(a) Plot the degree of polarization $\mathbb{P}$ versus the magnitude of the cross-correlation coefficient $|g_{xy}|$.

(b) Determine the coherency matrix if $\mathbb{P} = 0$, 0.5, and 1, and describe the nature of the light in each case.

(c) If the light is transmitted through a polarizer with its axis in the x direction, what is the intensity of the light transmitted?

CHAPTER
13

PHOTON OPTICS

13.1 THE PHOTON 516
- A. Photon Energy
- B. Photon Polarization
- C. Photon Position
- D. Photon Momentum
- E. Photon Interference
- F. Photon Time

13.2 PHOTON STREAMS 529
- A. Photon Flow
- B. Randomness of Photon Flow
- C. Photon-Number Statistics
- D. Random Partitioning of Photon Streams

*13.3 QUANTUM STATES OF LIGHT 541
- A. Coherent States
- B. Quadrature-Squeezed States
- C. Photon-Number-Squeezed States
- D. Two-Photon Light

Max Planck (1858–1947) suggested that the emission and absorption of light by matter takes the form of quanta of energy.

Albert Einstein (1879–1955) advanced the hypothesis that light itself comprises quanta of energy.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Electromagnetic optics, introduced in Chapter 5, provides the most complete treatment of light within the confines of **classical optics**. It encompasses wave optics, which in turn encompasses ray optics (Fig. 13.0-1). Though classical electromagnetic theory is capable of providing explanations for the preponderance of effects in optics, as attested to by the earlier chapters of this book, it nevertheless fails to account for certain optical phenomena. This failure, which became evident in about 1900, ultimately led to the formulation of a quantum electromagnetic theory known as **quantum electrodynamics** (QED). When applied to optical phenomena, QED is usually called **quantum optics**. Quantum optics properly describes almost all known optical phenomena.

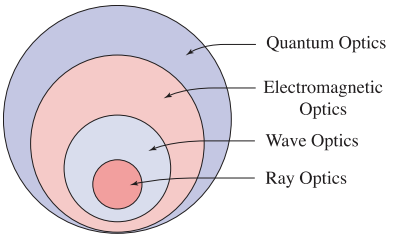


Figure 13.0-1 The theory of quantum optics explains virtually all optical phenomena. It is more general than electromagnetic optics, which was shown earlier to encompass wave optics and ray optics.

In the mathematical framework of quantum electrodynamics, the vectors $\mathbf{E}$ and $\mathbf{H}$ that are used to describe the electric and magnetic fields of classical electromagnetic optics, respectively, are promoted to operators in a Hilbert space. These operators are assumed to satisfy certain operator equations and commutation relations that govern their time dynamics and interdependence. Though the equations of QED describe the interactions of electromagnetic fields with matter in much the same way as Maxwell's equations, QED leads to results that are uniquely quantum in nature. In spite of its vast successes, quantum optics is nevertheless not the final arbiter of *all* optical effects. That distinction currently belongs to the **electroweak theory**, which combines quantum electrodynamics with the theory of weak interactions.[†] Continuing efforts are underway to combine electroweak theory with the theories of strong and gravitational interactions in an attempt to forge a general unified theory that accommodates all four fundamental forces of nature, as they are currently understood.

This Chapter

A formal treatment of quantum optics is beyond the scope of this book. Nevertheless, it is possible to describe many of the quantum properties of light, and its interaction with matter, by supplementing electromagnetic optics with several simple relationships drawn from quantum optics; these embody the corpuscularity, localization, and fluctuations of quantum fields and energy. This set of rules, which we call **photon optics**, permits us to deal with optical phenomena that lie beyond the reach of classical theory, while retaining classical optics as a limiting case. However, photon optics is not capable of accommodating all of the optical effects that can be explained by quantum optics.

In Sec. 13.1 we introduce the concept of the photon and examine its properties. Using electromagnetic optics as a point of departure, we impose a number of rules that

[†] For example, parity-nonconserving small rotations of the plane of polarization of light upon passage through certain materials cannot be accommodated by quantum electrodynamics but are successfully explained by electroweak theory; see, e.g., P. A. Vetter, D. M. Meekhof, P. K. Majumder, S. K. Lamoreaux, and E. N. Fortson, Precise Test of Electroweak Theory from a New Measurement of Parity Nonconservation in Atomic Thallium, *Physical Review Letters*, vol. 74, pp. 2658–2661, 1995.

govern the behavior of photon energy, polarization, position, momentum, interference, and time. These rules, which are deceptively simple, form the basis of photon optics and have far-reaching implications. This is followed, in Sec. 13.2, by a discussion of the properties of collections of photons and photon streams. The number of photons emitted by a fixed-intensity light source in a sequence of fixed time intervals is almost always random, with statistical properties that depend on the nature of the source. The photon-number statistics for commonly encountered optical sources, such as lasers and thermal radiators, are set forth. The effect of simple optical components, such as beamsplitters and filters, on the randomness of photon streams is also examined. In Sec. 13.3, we study the random fluctuations of the magnitude, phase, and photon number associated with the electromagnetic field from the perspective of quantum optics. We provide a brief introduction to coherent, quadrature-squeezed, photon-number-squeezed, and entangled-photon states of light, and indicate several generation mechanisms and applications. The interactions of photons with atoms and semiconductors are described in Secs. 14.3 and 17.2, respectively.

13.1 THE PHOTON

From a quantum perspective, light consists of particles called **photons**. A photon carries electromagnetic energy and momentum, as well as intrinsic angular momentum (or spin) associated with its polarization properties. It can also carry orbital angular momentum. The photon has zero rest mass and travels at c_o, the speed of light in vacuum; its speed in dielectric materials is reduced to $c < c_o$. A photon concomitantly has a wavelike character that determines its localization properties in space and time, and governs how it interferes and diffracts.

The notion of the photon initially grew out of an attempt by Max Planck in 1900 to resolve a long-standing conundrum concerning the spectrum of blackbody radiation emanating from a cavity held at a fixed temperature T (this topic is discussed in Sec. 14.4B). Planck ultimately resolved the problem by assuming that the allowed energies of the atoms in the walls of the cavity were quantized to discrete values. In 1905, Albert Einstein proposed that the quantization be imposed directly on the energy of the electromagnetic radiation, rather than on the atoms, which led to the concept of the photon. This enabled Einstein to successfully explain the photoelectric effect (this topic is discussed in Sec. 19.1A). The term "photon" was introduced by Gilbert Lewis in 1926.

The concept of the photon and the rules of photon optics are introduced by considering light inside an optical resonator (cavity). This is a convenient choice because it restricts the space under consideration to a simple geometry. However, the presence of the resonator turns out not to be an important feature of the argument; the results can be shown to be independent of the form of the resonator, and even of its presence.

Electromagnetic-Optics Theory of Light in a Resonator

In accordance with electromagnetic optics, light inside a lossless resonator of volume V is completely characterized by an electromagnetic field that takes the form of a superposition of discrete orthogonal modes of different spatial distributions, different frequencies, and different polarizations. The electric-field vector, $\mathcal{E}(\mathbf{r}, t) = \text{Re}\{\mathbf{E}(\mathbf{r}, t)\}$, can therefore be expressed in terms of the complex electric field $\mathbf{E}(\mathbf{r}, t)$ via

$$\mathbf{E}(\mathbf{r}, t) = \sum_{\mathbf{q}} A_{\mathbf{q}}\, U_{\mathbf{q}}(\mathbf{r})\, \exp(j2\pi\nu_{\mathbf{q}} t)\, \widehat{\mathbf{e}}_{\mathbf{q}}. \tag{13.1-1}$$

The $\mathbf{q}$th mode has complex envelope $A_{\mathbf{q}}$, frequency $\nu_{\mathbf{q}}$, polarization along the direction of the unit vector $\widehat{\mathbf{e}}_{\mathbf{q}}$, and a spatial distribution characterized by the complex function

$U_{\mathbf{q}}(\mathbf{r})$, which is normalized such that $\int_V |U_{\mathbf{q}}(\mathbf{r})|^2\, d\mathbf{r} = 1$. The expansion functions $U_{\mathbf{q}}(\mathbf{r})$, $\exp(j2\pi\nu_{\mathbf{q}}t)$, and $\hat{\mathbf{e}}_{\mathbf{q}}$ are not unique; other choices are available, including those comprising polychromatic modes.

In a cubic resonator of dimension d, a convenient choice for the spatial expansion functions is the set of standing waves

$$U_{\mathbf{q}}(\mathbf{r}) = \left(\frac{2}{d}\right)^{3/2} \sin\left(q_x \frac{\pi}{d} x\right) \sin\left(q_y \frac{\pi}{d} y\right) \sin\left(q_z \frac{\pi}{d} z\right), \tag{13.1-2}$$

where the integers q_x, q_y, and q_z are usually specified in the form (q_x, q_y, q_z) [Sec. 11.3C and Fig. 13.1-1(a)]. In accordance with (5.4-9), the energy density associated with mode $\mathbf{q}$ is $\frac{1}{2}\epsilon|A_{\mathbf{q}}|^2\,|U_{\mathbf{q}}(\mathbf{r})|^2$, so that the energy contained in mode $\mathbf{q}$ is

$$E_{\mathbf{q}} = \tfrac{1}{2}\epsilon \int_V |A_{\mathbf{q}}|^2\,|U_{\mathbf{q}}(\mathbf{r})|^2\, d\mathbf{r} = \tfrac{1}{2}\epsilon|A_{\mathbf{q}}|^2, \tag{13.1-3}$$

where V is the modal volume. In classical electromagnetic theory, the energy $E_{\mathbf{q}}$ can assume any nonnegative value, no matter how small, and the total energy is the sum of the energies in all modes.

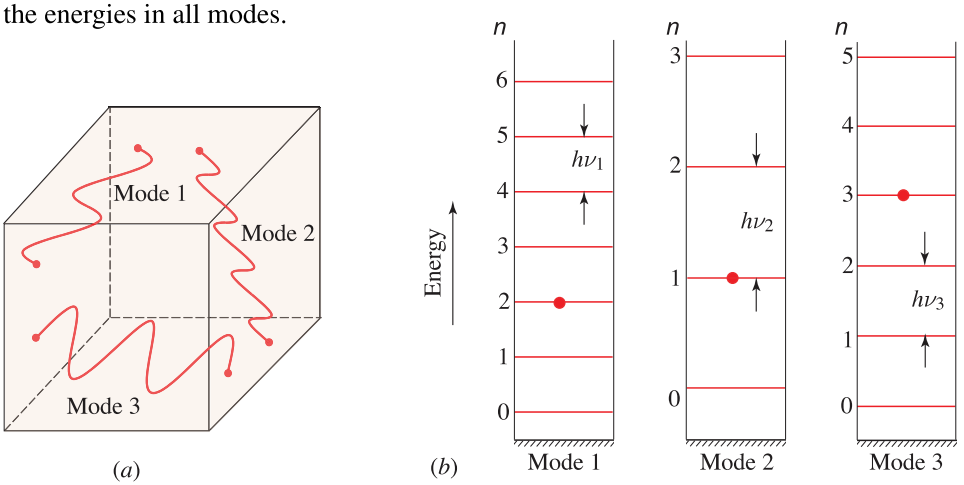


Figure 13.1-1 (a) Schematic of three electromagnetic modes of different frequencies and directions in a cubic resonator. (b) Allowed energy levels of three modes in the context of photon optics. Modes 1, 2, and 3 have frequencies ν_1, ν_2, and ν_3, respectively. In the example presented in the figure, modes 1, 2, and 3 contain $n = 2$, 1, and 3 photons, respectively, as represented by the filled circles.

Photon-Optics Theory of Light in a Resonator

The electromagnetic-optics theory described above is maintained in photon optics, but a restriction is placed on the energy that each mode is permitted to carry. Rather than assuming a continuous range, with no minimum allowed value, the modal energy is restricted to discrete values separated by a fixed energy $h\nu$, where ν is the frequency of the mode [Fig. 13.1-1(b)]. The energy of a mode is thus quantized, with only integral units of this fixed energy permitted. Each unit of energy is carried by a single photon and the mode may carry an arbitrary number of photons.

Light in a resonator comprises a set of modes, each of which contains an integral number of identical photons. Characteristics of the mode, such as its frequency, spatial distribution, direction of propagation, and polarization, are assigned to the photons.

A. Photon Energy

Photon optics provides that the energy of an electromagnetic mode is quantized to discrete levels separated by the energy of a photon [Fig. 13.1-1(*b*)]. The energy of a photon in a mode of frequency ν is

$$\boxed{E = h\nu = \hbar\omega,} \qquad (13.1\text{-}4)$$

Photon Energy

where $h = 6.6261 \times 10^{-34}$ J·s is **Planck's constant** and $\hbar \equiv h/2\pi$. Energy may be added to, or taken from this mode only in units of $h\nu$.

A mode containing zero photons nevertheless carries energy $E_0 = \frac{1}{2}h\nu$, which is called the **zero-point energy** and is associated with the fluctuations of the **vacuum state** (Fig. 13.3-3). When it carries n photons, a mode therefore has total energy

$$E_n = (n + \tfrac{1}{2})\, h\nu, \qquad n = 0, 1, 2, \ldots . \qquad (13.1\text{-}5)$$

This expression is identical to that for the energy levels of a quantum-mechanical harmonic oscillator, as provided in (13.3-4); the connection will be established in Sec. 13.3. In most experiments, the zero-point energy is not directly observable because only energy differences [e.g., $E_2 - E_1$ in (13.1-5)] are measured. However, the zero-point energy is not innocuous since it constitutes a source of noise ("shot noise") that limits the sensitivity of certain precision measurements. This will become evident in Sec. 13.3B, where we will demonstrate how the vacuum state can be manipulated ("squeezed") by configuring an experiment in a particular way so as to reduce the deleterious effects of vacuum-state fluctuations. Zero-point fluctuations are also responsible for the process of spontaneous emission from an atom, as discussed in Sec. 14.3. Moreover, are the origin of the **Casimir effect**, a small attractive force that acts between two parallel uncharged conducting plates located in close proximity.

The order of magnitude of the photon energy is readily estimated. An infrared photon of wavelength $\lambda_o = 1\ \mu$m in free space has a frequency $\nu \approx 3 \times 10^{14}$ Hz, by virtue of the relation $\lambda_o \nu = c_o$, and a period $T = 1/\nu$. Its energy is thus $h\nu \approx 1.99 \times 10^{-19}$ J. In units of electron volts, the photon energy becomes $h\nu/e = (1.99 \times 10^{-19})/(1.6 \times 10^{-19}) = 1.24$ eV; this is equivalent to the kinetic energy imparted to an electron when it is accelerated through a potential difference of 1.24 V. Another example is provided by a microwave photon with a wavelength of 1 cm; the photon energy is then 10^4 times smaller, namely $h\nu = 1.24 \times 10^{-4}$ eV. A convenient conversion formula between wavelength (μm) and photon energy (eV) is therefore expressible as

$$\boxed{E\ (\text{eV}) \approx \frac{1.24}{\lambda_o\ (\mu\text{m})}\,.} \qquad (13.1\text{-}6)$$

The reciprocal wavelength is also frequently used as a unit of energy, often in chemistry. It is specified in cm^{-1} and is determined by expressing the wavelength in cm and simply taking the inverse. Thus, 1 cm^{-1} corresponds to 1.24/10 000 eV and 1 eV corresponds to 8065 cm^{-1}. Conversions among photon wavelength, frequency, period, and energy are illustrated in Fig. 13.1-2.

Because the photon energy increases with frequency, the particle nature of light becomes increasingly prevalent as the frequency of the radiation increases. X-rays and gamma-rays almost always behave like particles, and wavelike effects such as diffraction and interference are difficult to discern. In contrast, radio waves almost always behave like waves. The frequency of light in the optical region is such that both

particle-like and wavelike behavior are readily observed, thus spurring the need for photon optics.

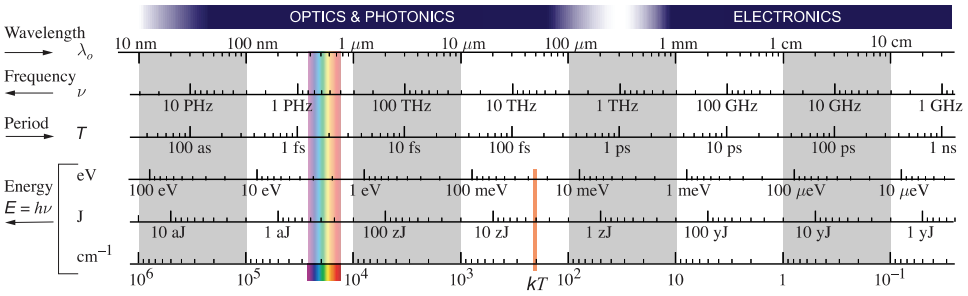


Figure 13.1-2 Relationships among photon wavelength λ_o, frequency ν, period T, and energy E (specified in units of eV, J, and reciprocal wavelength $1/\lambda_o$ in cm^{-1}). A photon of free-space wavelength $\lambda_o = 1$ μm has frequency $\nu = 300$ THz, period $T = 3.33$ fs, and energy $E = 1.24$ eV = 199 zJ = 10^4 cm^{-1}. At room temperature ($T = 300°$ K), the thermal energy $kT = 26$ meV = 4.17 zJ = 210 cm^{-1}. Two spectral domains are indicated: 1) optics & photonics, and 2) electronics.

B. Photon Polarization

As indicated earlier, light is characterized by a set of modes of different frequencies, directions, and polarizations, each occupied by an integral number of photons. For each monochromatic plane wave traveling in a particular direction, there are two polarization modes. The polarization of a photon is that of the mode it occupies. For example, the photon may be linearly polarized in the x direction, or right circularly polarized. Since the polarization modes of free space are degenerate, they are not unique. One may use modes with linear polarization in the x and y directions, linear polarization in two other orthogonal directions, say x' and y', or right- and left-circular polarizations. The choice of a particular set is a matter of convenience. A problem arises when a photon occupying a given mode (say linear polarization in the x direction) is to be observed in a different set of modes (say linear polarization in the x' and y' directions). Since the photon energy cannot be split between the two modes, a probabilistic interpretation is called for.

In classical electromagnetic optics, the state of polarization of a plane wave is described by a Jones vector, whose components (A_x, A_y) are the components of the complex envelope in the x and y directions, respectively (Sec. 6.1A). The same wave may also be represented in a different coordinate system (x', y'), e.g., one that makes a 45° angle with the initial coordinate system, by a Jones vector with components

$$A_{x'} = \tfrac{1}{\sqrt{2}}(A_x - A_y), \qquad A_{y'} = \tfrac{1}{\sqrt{2}}(A_x + A_y), \tag{13.1-7}$$

as described in Sec. 6.1B. Therefore, a wave that is linearly polarized in the x direction is described by a Jones vector with components $(A_0, 0)$ in the x–y coordinate system, where A_0 is the complex envelope. In the (x', y') coordinate system, the Jones vector has components $(\frac{1}{\sqrt{2}}A_0, \frac{1}{\sqrt{2}}A_0)$.

The state of polarization of a single photon is described by a Jones vector with complex components (A_x, A_y), normalized such that $|A_x|^2 + |A_y|^2 = 1$. The coefficients A_x and A_y are interpreted as complex probability amplitudes, and their squared magnitudes, $|A_x|^2$ and $|A_y|^2$, represent the probabilities that the photon is observed in the x and y linear polarization modes, respectively.

The components (A_x, A_y) are transformed from one coordinate system to another in the same manner as ordinary Jones vectors, and the new components represent complex probability amplitudes in the new modes. Thus, a single photon may exist, probabilistically, in more than one mode. This concept is illustrated by the following examples.

Linearly Polarized Photon

A photon is linearly polarized in the x direction. In terms of the x–y linearly polarized modes, the photon is described by a Jones vector with components $(1, 0)$. In a set of linearly polarized modes in the x' and y' directions at $45°$, these components are $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ so that the probabilities of observing the photon in a linear polarization mode along the x' or y' directions are both $1/2$. This is illustrated schematically in Fig. 13.1-3.

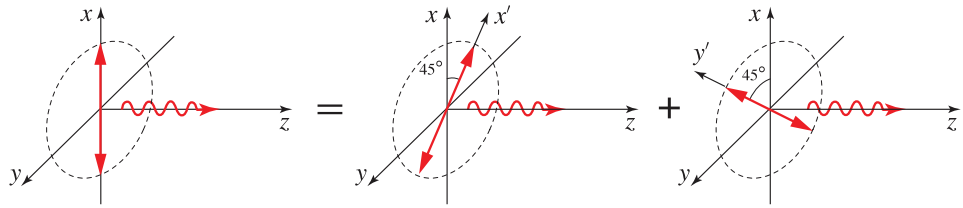


Figure 13.1-3 A photon in the x linear polarization mode is the same as a photon in a superposition of the x' and y' linear polarization modes, each with probability $1/2$.

EXAMPLE 13.1-1. ***Transmission of a Linearly Polarized Photon Through a Polarizer.*** Consider the transmission of a photon that is linearly polarized in the x direction through a linear polarizer whose transmission axis is along the x' direction at an angle θ, as illustrated in Fig. 13.1-4. The polarizer transmits light that is linearly polarized in the x' direction but blocks light in the orthogonal y' direction. The probability that the photon is transmitted through the polarizer is determined by writing the Jones vector of the photon polarization state in the x'–y' coordinate system as $(\cos\theta, -\sin\theta)$ [see (6.1-21) and (6.1-22)]. The probability of observing the photon in the mode with x' linear polarization is therefore $\cos^2\theta$, which represents the probability of passage of the photon through the polarizer: $p(\theta) = \cos^2\theta$. The probability that the photon is blocked is therefore $1 - p(\theta) = \sin^2\theta$. Classical polarization optics reveals that the intensity transmittance of a polarizer in this same configuration is $\cos^2\theta$ (Prob. 6.1-7). We conclude that the probability of transmission of a single photon is identical to the classical transmittance, i.e., $p(\theta) = \mathcal{T}(\theta)$.

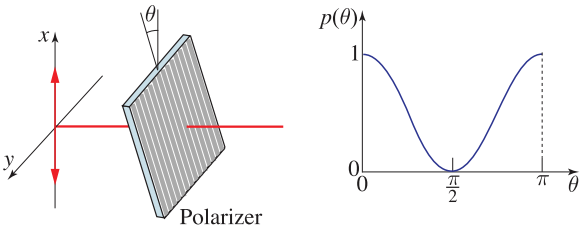


Figure 13.1-4 Probability of a linearly polarized photon passing through a polarizer. The axis of the polarizer is at an angle θ with respect to the photon polarization.

Circularly Polarized Photon

A circularly polarized photon is described by a Jones vector that has components $\frac{1}{\sqrt{2}}(1, \pm j)$, where the $+$ and $-$ signs correspond to right- and left-handed polarization, respectively. This description is based on an x–y coordinate system, i.e., linearly polarized modes. Therefore, the probability of the photon passing through a linear polarizer pointing in either the x or y direction is $1/2$. It can also be shown that this result prevails whatever the direction of the linear polarizer. The circularly polarized photon may be regarded as equivalent to the probabilistic superposition of one photon with linear polarization in the x direction and another in the y direction, each with probability $1/2$.

Right- and left-circular polarizations may also be used as modes (as a coordinate system). In that description, a linearly polarized photon may be regarded as a probabilistic superposition of right- and left-circularly polarized photons, each with probability $1/2$, as illustrated in Fig. 13.1-5.

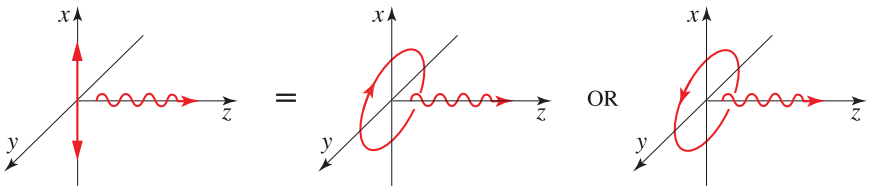


Figure 13.1-5 A linearly polarized photon is equivalent to the superposition of a right- and a left-circularly polarized photon, each with probability $1/2$.

C. Photon Position

Associated with each photon of frequency ν is a wave described by the complex wavefunction $U(\mathbf{r})\exp(j2\pi\nu t)$ of the mode. However, when a photon impinges on a detector of small area dA located normal to the direction of propagation, at the position $\mathbf{r}$, its indivisibility causes it to be either wholly detected or not detected at all. The location at which the photon is registered is not precisely determined. Rather, it is governed by the optical intensity $I(\mathbf{r}) \propto |U(\mathbf{r})|^2$, in accordance with the following probabilistic law:

The probability $p(\mathbf{r})\,dA$ of observing a photon at the position $\mathbf{r}$ within an incremental area dA, at any time, is proportional to the local optical intensity $I(\mathbf{r}) \propto |U(\mathbf{r})|^2$, so that

$$p(\mathbf{r})\,dA \propto I(\mathbf{r})\,dA. \tag{13.1-8}$$

Photon Position

The photon is therefore more likely to be found at those locations where the intensity is high. A photon in a mode described by a standing wave with the intensity distribution $I(x, y, z) \propto \sin^2(\pi z/d)$, where $0 \le z \le d$, for example, is most likely to be detected at $z = d/2$, but will never be detected at $z = 0$ or $z = d$. In contrast to waves, which are extended in space, and particles, which are localized in space, optical photons behave as extended *and* localized entities. This behavior is called **wave–particle duality**. The localized nature of photons becomes evident when they are detected.

EXERCISE 13.1-1

Photon in a Gaussian Beam.

(a) Consider a single photon described by a Gaussian beam (the $\mathrm{TEM}_{0,0}$ mode of a spherical-mirror resonator; see Secs. 3.1B, 5.4A, and 11.2B). What is the probability of detecting the photon at a point within a circle whose radius is the waist radius of the beam W_0? Recall from (3.1-12) that at the waist, $I(\rho, z = 0) \propto \exp(-2\rho^2/W_0^2)$, where ρ is the radial coordinate.

(b) If the beam carries a large number n of independent photons, estimate the average number of photons that lie within this circle.

Transmission of a Single Photon Through a Beamsplitter

An ideal beamsplitter is an optical device that losslessly splits a beam of light into two beams that emerge at right angles. It is characterized by an intensity transmittance $\mathcal{T}$ and an intensity reflectance $\mathcal{R} = 1 - \mathcal{T}$. The intensity of the transmitted wave I_t and the intensity of the reflected wave I_r can be calculated from the intensity of the incident wave I using the electromagnetic relations $I_t = \mathcal{T}I$ and $I_r = (1 - \mathcal{T})I$.

Because a photon is indivisible, it must choose between the two possible directions permitted by the beamsplitter. A single photon incident on the device will follow these directions in accordance with the probabilistic photon-position rule (13.1-8). The probability that the photon is transmitted is proportional to I_t and is therefore equal to the transmittance $\mathcal{T} = I_t/I$. The probability that it is reflected is $1 - \mathcal{T} = I_r/I$. From the point of view of probability, the problem is identical to that of flipping a biased coin. Figure 13.1-6 illustrates the process.

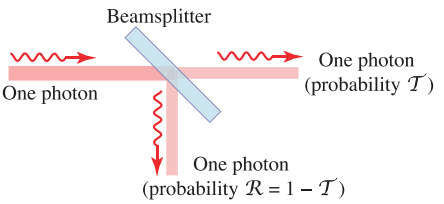


Figure 13.1-6 Probabilistic reflection or transmission of a photon at a lossless beamsplitter.

Single-Photon Imaging

As described in Sec. 4.4 and in (A.3-3) of Appendix A, a coherent imaging system is characterized by an impulse response function $h(x, y; x', y')$ that links its output and input fields, $U_o(x, y)$ and $U_i(x, y)$, respectively, via the two-dimensional convolution

$$U_o(x, y) = \int_{-\infty}^{\infty}\int U_i(x', y')\, h(x, y; x', y')\, dx'\, dy'. \tag{13.1-9}$$

The very same relationship characterizes the single-photon wavefunctions at the output and input of a single-photon imaging system, where $|U_o(x)|^2$ represents the probability density function of the photon position in the image plane.

D. Photon Momentum

In classical electromagnetic optics, as discussed in Sec. 5.4A, an electromagnetic plane wave carries a linear momentum density $(W/c)\widehat{\mathbf{k}}$, where W is the energy density (per unit volume) and $\widehat{\mathbf{k}}$ is a unit vector in the direction of the wavevector $\mathbf{k}$.

In photon optics, the linear momentum of a photon is $\mathbf{p} = (E/c)\widehat{\mathbf{k}}$ where $E = \hbar\omega = \hbar ck$ is the photon energy, so that:

The linear momentum associated with a photon in a plane-wave mode of wavevector $\mathbf{k}$ is

$$\mathbf{p} = \hbar\mathbf{k}. \tag{13.1-10}$$

Photon Momentum

The magnitude of the momentum is $p = \hbar k = \hbar\omega/c = \hbar 2\pi/\lambda$, so that

$$p = E/c = h/\lambda. \tag{13.1-11}$$

*Momentum of a Localized Wave

A wave more general than a plane wave, with a complex wavefunction of the form $U(\mathbf{r})\ \exp(j2\pi\nu t)$, can be expanded as a sum of plane waves of different wavevectors by using the techniques of Fourier optics (Chapter 4). The component with wavevector $\mathbf{k}$ may be written in the form $A(\mathbf{k})\exp(-j\mathbf{k}\cdot\mathbf{r})\exp(j2\pi\nu t)$, where $A(\mathbf{k})$ is its amplitude.

The momentum of a photon described by an arbitrary complex wavefunction $U(\mathbf{r})\exp(j2\pi\nu t)$ is uncertain. It assumes the value

$$\mathbf{p} = \hbar\mathbf{k}, \tag{13.1-12}$$

with probability proportional to $|A(\mathbf{k})|^2$, where $A(\mathbf{k})$ is the amplitude of the plane-wave Fourier component of $U(\mathbf{r})$ with wavevector $\mathbf{k}$.

If $f(x, y) = U(x, y, 0)$ is the complex amplitude at the $z = 0$ plane, the plane-wave Fourier component with wavevector $\mathbf{k} = (k_x, k_y, k_z)$ has an amplitude $A(\mathbf{k}) = F(k_x/2\pi, k_y/2\pi)$, where $F(\nu_x, \nu_y)$ is the two-dimensional Fourier transform of $f(x, y)$, as described in Chapter 4. Because the functions $f(x, y)$ and $F(\nu_x, \nu_y)$ form a Fourier transform pair, their widths are inversely related and satisfy the position–direction relation provided in (A.2-6) of Appendix A. The uncertainty relation between the position of the photon and the direction of its momentum is established because the position of the photon at the $z = 0$ plane is probabilistically determined by $|U(\mathbf{r})|^2 = |f(x, y)|^2$, and the direction of its momentum is probabilistically determined by $|A(\mathbf{k})|^2 = |F(k_x/2\pi, k_y/2\pi)|^2$. Thus if, at the plane $z = 0$, σ_x is the positional uncertainty in the x direction, and $\sigma_\theta = \sin^{-1}(\sigma_{kx}/k) \approx (\lambda/2\pi)\sigma_{kx}$ is the angular uncertainty about the z axis (which is assumed to be $\ll 1$), then the uncertainty relation $\sigma_x\sigma_{kx} \geq 1/2$ is equivalent to $\sigma_x\sigma_\theta \geq \lambda/4\pi$.

A plane-wave photon has a known momentum (fixed direction and magnitude), so that $\sigma_\theta = 0$, but its position is totally uncertain ($\sigma_x = \infty$); it is equally likely to be detected anywhere in the $z = 0$ plane. When a plane-wave photon passes through an aperture, its position becomes localized at the expense of a spread in the direction of its momentum. The position–direction uncertainty of the photon therefore parallels the theory of diffraction described in Sec. 4.3. At the other extreme from the plane wave is the spherical-wave photon. It is well localized in position (at the center of the wave), but the direction of its momentum is totally uncertain.

Radiation Pressure

Because a photon carries momentum, and momentum is conserved, the atom emitting the photon experiences a recoil of magnitude $h\nu/c$. The momentum associated with a photon can also be transferred to an object, giving rise to a force that results in mechanical motion. As an example, light beams can be used to deflect atomic beams or to contain collections of atoms or small dielectric particles (Sec. 14.3F). The term **radiation pressure** is often used to describe this phenomenon (pressure = force/area).

EXERCISE 13.1-2

Photon-Momentum Recoil. Determine the recoil velocity imparted to a ^{198}Hg atom that emits a photon of energy 4.88 eV. Compare this with the root-mean-square thermal velocity v of the atom at a temperature $T = 300°$ K (this is obtained by setting the average kinetic energy equal to the average thermal energy, $\frac{1}{2}mv^2 = \frac{3}{2}kT$, where $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant).

Photon Spin Angular Momentum

The intrinsic **spin angular momentum** of a photon associated with circular polarization is characterized by an electric-field vector that rotates in a circle normal to the direction of propagation. Because the photon travels at the speed of light, the projection of its spin vector lies either parallel or antiparallel to the wavevector, so that its **helicity** is quantized to two values,

$$\boxed{\mathbb{S} = \pm\hbar\,,} \qquad \text{(13.1-13)}$$
Photon Spin

where the plus (minus) signs are associated with right-handed (left-handed) circular polarization, respectively. Linearly polarized photons have equal probability of exhibiting parallel and antiparallel spin. Just as photons can transfer linear momentum to an object, so too can circularly polarized photons exert a torque on an object. A circularly polarized photon will, for example, exert a torque on a half-wave plate.

Bosons and fermions. Fundamental particles are divided into two broad classes: **Bosons**, such as photons and other force-carrier particles, have a spin that is an integer multiple of $\hbar$, as do quasiparticles such as plasmons, polaritons, and phonons. In contrast **fermions**, such as electrons, protons, neutrons, and other material particles, have a spin that is a half-integer multiple of $\hbar$.

Photon Orbital Angular Momentum

Aside from the spin angular momentum associated with polarization, an electromagnetic wave may carry angular momentum by virtue of the twisting of its wavefront about the axis of propagation. For example, the Laguerre–Gaussian beam described by the complex wavefunction $U_{l,m}(\rho, \phi, z)$ in (3.4-1), which has an azimuthal phase dependence $\exp(-jl\phi)$ and therefore a helical wavefront, has an angular momentum (for $l \neq 0$) that is independent of its state of polarization. To distinguish this from spin angular momentum, it is referred to as **orbital angular momentum**. A photon in such a spatial mode possesses an orbital angular momentum $\mathsf{L} = l\hbar$.

Another example is provided by a photon in a whispering-gallery mode (WGM) of a cylindrical resonator (Sec. 11.3B). In the context of ray optics, the mode is described by a ray tracing the cross-sectional circular boundary of the resonator. In the context of wave optics, the wavelength satisfies the resonance condition $2\pi a = q\lambda$, where a is the

radius of the circle and $q = 1, 2, \ldots$. The linear momentum of the photon is $p = \hbar k = \hbar 2\pi/\lambda = q\hbar/a$, and its angular momentum is therefore $ap = q\hbar$. Similarly, a photon in a WGM mode of a microsphere resonator (Sec. 11.4C) of radius a has an angular momentum $L = \ell\hbar$, where the integer ℓ is associated with the resonance wavelength for an optical path that traces a great circle. The quantity ℓ may be regarded as an angular-momentum quantum number similar to that used to describe the hydrogen atom (Sec. 14.1A).

E. Photon Interference

Young's double-pinhole interference experiment is generally invoked to demonstrate the wave nature of light (Exercise 2.5-2). In fact, Young's experiment can be carried out even when there is only a single photon in the apparatus at a given time. The outcome of this experiment can be understood in the context of photon optics by using the photon-position rule. The intensity at the observation plane is calculated using wave optics and the result is converted to a probability density function that specifies the random position of the detected photon. The interference arises from phase differences associated with the two possible paths.

Consider a plane wave illuminating a screen with two pinholes separated by a distance $2a$, as portrayed in Fig. 13.1-7 (see also Fig. 2.5-6). The line joining the holes defines the x axis. Two spherical waves are generated that interfere at the observation plane. As is understood from Exercise 2.5-2, in the paraboloidal-wave approximation, these give rise to a sinusoidal intensity that behaves in accordance with

$$I(x) \approx 2I_0 \left(1 + \cos \frac{2\pi x\theta}{\lambda}\right). \tag{13.1-14}$$

Here I_0 is the intensity of each of the waves individually at the observation plane, λ is the wavelength, and θ is the angle subtended by the two pinholes at the observation plane (Fig. 13.1-7). The result provided in (13.1-14) describes the intensity pattern that is experimentally observed for classical light.

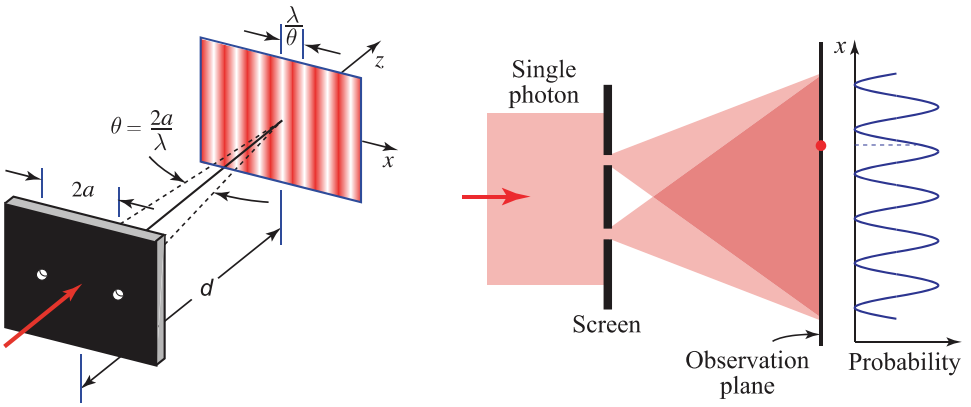


Figure 13.1-7 Young's double-pinhole experiment with a single photon. The interference pattern $I(x)$ determines the probability density of detecting the photon at position x.

Now, if only a single photon is present in the apparatus, in accordance with (13.1-8) the probability of detecting it at position x is proportional to $I(x)$. It is most likely to be detected at those values of x for which $I(x)$ is a maximum and it will never

be detected at values of x for which $I(x) = 0$. If a histogram of the locations of the detected photon is constructed by repeating the experiment many times, as Taylor did in 1909, the classical interference pattern obtained by carrying out the experiment only once with a strong beam of light emerges. The classical interference pattern does indeed represent the probability density of the position at which a single photon is observed.

The occurrence of the interference is a result of the extended nature of the photon, which permits it to pass through *both* holes of the apparatus. This endows the photon with knowledge of the geometry of the entire experiment when it reaches the observation plane, where it is detected as a single entity. If one of the holes were to be covered, the interference pattern would disappear.

EXERCISE 13.1-3

Single Photon in a Mach–Zehnder Interferometer. Consider a plane wave of light of wavelength λ that is split into two parts at a beamsplitter (Sec. 13.1C) and recombined in a Mach–Zehnder interferometer, as shown in Fig. 13.1-8 [see also Fig. 2.5-3(a)]. If the wave contains only a single photon, plot the probability of finding it at the detector as a function of d/λ (for $0 \le d/\lambda \le 1$), where d is the difference between the two optical paths of the light. Assume that the mirrors and beamsplitters are perfectly flat and lossless, and that the beamsplitters have $\mathcal{T} = \mathcal{R} = 1/2$. Where might the photon be located when the probability of finding it at the detector is not unity?

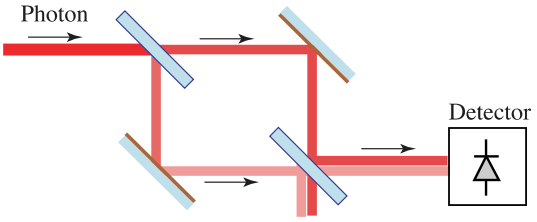


Figure 13.1-8 Single photon in a Mach–Zehnder interferometer.

F. Photon Time

The modal expansion provided in (13.1-1) comprises monochromatic modes that are "eternal" harmonic functions of time; a photon in a monochromatic mode is equally likely to be detected at any time. As indicated earlier, however, a modal expansion of the radiation inside (or outside) a resonator is not unique. A more general expansion comprises polychromatic modes such as time-localized wavepackets (Sec. 2.6A). The probability of detecting a photon described by the complex wavefunction $U(\mathbf{r}, t)$, at any position and in the incremental time interval between t and $t + dt$, is proportional to $I(\mathbf{r}, t)\, dt \propto |U(\mathbf{r}, t)|^2\, dt$. The photon-position rule of photon optics presented in (13.1-8) may therefore be generalized to include photon time localization:

The probability $p(\mathbf{r}, t)\, dA\, dt$ of observing a photon at position $\mathbf{r}$ within an incremental area dA, and during an incremental time interval dt following time t, is proportional to the optical intensity of the mode at $\mathbf{r}$ and t, i.e.,

$$p(\mathbf{r}, t)\, dA\, dt \propto I(\mathbf{r}, t)\, dA\, dt \propto |U(\mathbf{r}, t)|^2\, dA\, dt. \tag{13.1-15}$$

Photon Position and Time

Time–Energy Uncertainty

The time during which a photon in a monochromatic mode of frequency ν may be detected is totally uncertain, whereas the value of its frequency ν (and its energy $h\nu$) is absolutely certain. In contrast, a photon in a wavepacket mode with an intensity function $I(t)$ of duration σ_t must be localized within this time. Bounding the photon time in this way engenders an uncertainty in its frequency (and energy) by virtue of the properties of the Fourier transform, and the result is a polychromatic photon. Suppressing the $\mathbf{r}$ dependence for simplicity, the frequency uncertainty is readily determined by carrying out a Fourier-transform expansion of $U(t)$ in terms of its harmonic components,

$$U(t) = \int_{-\infty}^{\infty} V(\nu) \exp(j2\pi\nu t)\, d\nu, \tag{13.1-16}$$

where $V(\nu)$ is the Fourier transform of $U(t)$ (Sec. A.1 of Appendix A). The width σ_ν of $|V(\nu)|^2$ represents the spectral width. If σ_t is the power-RMS width of the function $|U(t)|^2$, then σ_t and σ_ν must satisfy the duration–bandwidth reciprocity relation $\sigma_\nu \sigma_t \geq 1/4\pi$ or, equivalently, $\sigma_\omega \sigma_t \geq 1/2$ (the definitions of σ_t and σ_ν that lead to this uncertainty relation are provided in Sec. A.2 of Appendix A).

The energy of the photon $\hbar\omega$ cannot then be specified to an accuracy better than $\sigma_E = \hbar\sigma_\omega$. It follows that the energy uncertainty of a photon, and the time during which it may be detected, must satisfy

$$\boxed{\sigma_E \sigma_t \geq \frac{\hbar}{2},} \tag{13.1-17}$$

Time–Energy Uncertainty

which is known as the **time–energy uncertainty relation**. This relation is analogous to that between position and wavenumber (momentum), which sets a limit on the precision with which the position and momentum of a photon can be simultaneously specified. The average energy $\overline{E}$ of this polychromatic photon is $\overline{E} = h\overline{\nu} = \hbar\overline{\omega}$.

To summarize: a monochromatic photon ($\sigma_\nu \to 0$) has an eternal duration within which it can be observed ($\sigma_t \to \infty$). In contrast, a photon associated with an optical wavepacket is localized in time and is thus polychromatic with a corresponding energy uncertainty. Thus, a *wavepacket photon* can be viewed as a confined traveling packet of energy.

EXERCISE 13.1-4

Single Photon in a Gaussian Wavepacket. Consider a plane-wave wavepacket (Sec. 2.6A) containing a single photon traveling in the z direction, with complex wavefunction

$$U(\mathbf{r}, t) = \mathfrak{a}\left(t - \frac{z}{c}\right) \tag{13.1-18}$$

where

$$\mathfrak{a}(t) = \exp\left(-\frac{t^2}{4\tau^2}\right) \exp(j2\pi\nu_0 t). \tag{13.1-19}$$

(a) Show that the uncertainties in its time and z position are $\sigma_t = \tau$ and $\sigma_z = c\sigma_t$, respectively.

(b) Show that the uncertainties in its energy and momentum satisfy the minimum uncertainty relations:

$$\sigma_E \sigma_t = \hbar/2 \tag{13.1-20}$$

$$\sigma_z \sigma_p = \hbar/2. \tag{13.1-21}$$

Equation (13.1-21) is the minimum-uncertainty limit of the **Heisenberg position–momentum uncertainty relation** provided in (A.2-7) of Appendix A.

Summary

Electromagnetic radiation may be described in terms of a sum of modes, e.g., monochromatic uniform plane waves of the form

$$\mathbf{E}(\mathbf{r}, t) = \sum_{\mathbf{q}} A_{\mathbf{q}} \exp(-j\mathbf{k}_{\mathbf{q}} \cdot \mathbf{r}) \exp(j2\pi\nu_{\mathbf{q}} t)\, \widehat{\mathbf{e}}_{\mathbf{q}}. \tag{13.1-22}$$

Each plane wave has two orthogonal polarization states (e.g., vertical/horizontal linearly polarized, right/left circularly polarized) represented by the vectors $\widehat{\mathbf{e}}_{\mathbf{q}}$. When the energy of a mode is measured, the result is an integer (in general, random) number of energy quanta (photons). Each of the photons associated with the mode $\mathbf{q}$ has the following properties:

- Energy $E = h\nu_{\mathbf{q}}$.
- Momentum $\mathbf{p} = \hbar\mathbf{k}_{\mathbf{q}}$.
- Helicity (spin angular momentum) $\mathbb{S} = \pm\hbar$, if it is circularly polarized.
- The photon is equally likely to be found anywhere in space, and at any time, since the wavefunction of the mode is a monochromatic plane wave.

The choice of modes is not unique. A modal expansion in terms of non-monochromatic (quasi-monochromatic), non-plane waves, is also possible:

$$\mathbf{E}(\mathbf{r}, t) = \sum_{\mathbf{q}} A_{\mathbf{q}}\, U_{\mathbf{q}}(\mathbf{r}, t)\, \widehat{\mathbf{e}}_{\mathbf{q}}. \tag{13.1-23}$$

Each of the photons associated with the mode $\mathbf{q}$ then has the following properties:

- The photon position and time are governed by the complex wavefunction $U_{\mathbf{q}}(\mathbf{r}, t)$. The probability of observing the photon at position $\mathbf{r}$ within an incremental area dA, and during an incremental time interval dt following time t, is proportional to $|U_{\mathbf{q}}(\mathbf{r}, t)|^2\, dA\, dt$.
- If $U_{\mathbf{q}}(\mathbf{r}, t)$ has a finite time duration σ_t, i.e., if the photon is localized in time, then the photon energy $h\nu_{\mathbf{q}}$ has an uncertainty $h\sigma_\nu \geq h/4\pi\sigma_t$.
- If $U_{\mathbf{q}}(\mathbf{r}, t)$ has a finite spatial extent in the transverse $(z = 0)$ plane, i.e., if the photon is localized in the x direction, for example, then the direction of photon momentum is uncertain. The spread in photon momentum can be determined by analyzing $U_{\mathbf{q}}(\mathbf{r}, t)$ as a sum of plane waves, the wave with wavevector $\mathbf{k}$ corresponding to photon momentum $\hbar\mathbf{k}$. Spatial localization of the photon in the transverse plane results in an increase in the uncertainty of the photon-momentum direction.

13.2 PHOTON STREAMS

In Sec. 13.1 we concentrated on the properties and behavior of single photons. We now consider the properties of collections of photons. As a result of the processes by which photons are created (e.g., atomic emissions, as discussed in Chapter 14), the number of photons occupying any mode is generally random. Photon streams often contain numerous propagating modes, each carrying a random number of photons. If an experiment is carried out in which a weak stream of photons falls on a photosensitive surface, the individual photons are registered (detected) at random localized instants of time and at random points in space, in accordance with (13.1-15). (This space–time process can be discerned by viewing a barely illuminated object with the naked, dark-adapted eye.)

The temporal and spatial behavior of the photon registrations can be examined individually. The *temporal pattern* can be highlighted by making use of a detector that integrates light over a finite area A, but has good temporal resolution (e.g., a photodiode), as illustrated in Fig. 13.2-1. According to (13.1-15), the probability of detecting a photon in the incremental time interval between t and $t + dt$ is proportional to $P(t) = \int_A I(\mathbf{r}, t)\, dA$, which is the optical power at time t. The photons are registered at random times.

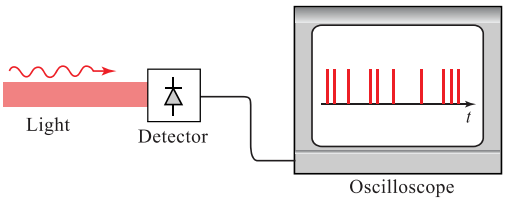


Figure 13.2-1 Photon registrations at random localized instants of time for a detector that integrates light over an area A.

The *spatial pattern* of photon registrations, on the other hand, is readily manifested by making use of a detector that integrates light over a fixed exposure time T but has good spatial resolution (e.g., photographic film). In accordance with (13.1-15), the probability of detecting a photon in an incremental area dA surrounding the point $\mathbf{r}$ is then proportional to $\int_0^T I(\mathbf{r}, t)\, dt$, which is the integrated local intensity. The photons are registered at random locations, as illustrated by the grainy image of Max Planck provided in Fig. 13.2-2. This image was obtained by rephotographing, under very low light conditions, the image of Max Planck presented on page 514. Each white dot in the photograph represents a random photon registration; the density of these registrations follows the local intensity.

Figure 13.2-2 Random photon registrations exhibit a spatial density that follows the local optical intensity. This image of Max Planck, illuminated by a sparse stream of photons, should be compared with that on page 514, illuminated with high-intensity light.

A. Photon Flow

We begin by introducing a number of definitions that relate the mean flow of photons to classical electromagnetic intensity, power, and energy. These definitions are inspired by (13.1-15), which dictates the position and time at which a single-photon detection occurs. We then turn to randomness in the photon flow and consider the photon-number statistics for different sources of light. Finally, we consider the random partitioning of a photon stream by a beamsplitter or photodetector.

Mean Photon-Flux Density

Monochromatic light of frequency ν and constant classical intensity $I(\mathbf{r})$ (watts/cm^2) carries a mean **photon-flux density**

$$\boxed{\phi(\mathbf{r}) = \frac{I(\mathbf{r})}{h\nu}.} \qquad (13.2\text{-}1)$$

Mean Photon-Flux Density

Since each photon carries energy $h\nu$, this equation provides a straightforward conversion from a classical measure (units of energy/s-cm^2) into a quantum measure (units of photons/s-cm^2). For quasi-monochromatic light of central frequency $\overline{\nu}$, all photons have approximately the same energy $h\overline{\nu}$, so that the mean photon-flux density can be approximately expressed as

$$\phi(\mathbf{r}) \approx \frac{I(\mathbf{r})}{h\overline{\nu}}. \qquad (13.2\text{-}2)$$

Typical values of $\phi(\mathbf{r})$ for some commonly encountered sources of light are provided in Table 13.2-1. It is clear from these numbers that trillions of photons rain down on each square centimeter of us each second.

Table 13.2-1 Mean photon-flux density for various sources of light.

Source	Mean Photon-Flux Density (photons/s-cm^2)
Starlight	10^6
Moonlight	10^8
Twilight	10^{10}
Indoor light	10^{12}
Sunlight	10^{14}
Laser light[a]	10^{22}

[a] A 10-mW He–Ne laser beam at $\lambda_o = 633$ nm focused to a 20-μm-diameter spot.

Mean Photon Flux

The mean photon flux Φ (units of photons/s) is obtained by integrating the mean photon-flux density over a specified area,

$$\boxed{\Phi = \int_A \phi(\mathbf{r})\, dA = \frac{P}{h\overline{\nu}},} \qquad (13.2\text{-}3)$$

Mean Photon Flux

where the optical power (watts) is defined as

$$P = \int_A I(\mathbf{r})\, dA \qquad (13.2\text{-}4)$$

and $h\overline{\nu}$ is again the average energy of a photon. As an example, 1 nW of optical power at a wavelength $\lambda_o = 0.2\ \mu\text{m}$ delivers to an object an average photon flux $\Phi \approx 10^9$ photons/s. Roughly speaking, one photon thus strikes the object every nanosecond, i.e.,

$$\boxed{1\text{ nW at }\lambda_o = 0.2\ \mu\text{m} \Rightarrow 1\text{ photon/ns.}} \tag{13.2-5}$$

A photon of wavelength $\lambda_o = 1\ \mu\text{m}$ carries one-fifth as much energy, in which case 1 nW corresponds to an average of 5 photons/ns.

Mean Number of Photons

The mean number of photons $\overline{n}$ detected in the area A and in the time interval T is obtained by multiplying the mean photon flux Φ in (13.2-3) by the time duration, whereupon

$$\overline{n} = \Phi T = \frac{E}{h\overline{\nu}}, \tag{13.2-6}$$

Mean Photon Number

where $E = PT$ is the optical energy (joules).

To summarize, the relations between the classical and quantum measures of photon flow are:

Classical		Quantum	
Optical intensity	$I(\mathbf{r})$	Photon-flux density	$\phi(\mathbf{r}) = I(\mathbf{r})/h\overline{\nu}$
Optical power	P	Photon flux	$\Phi = P/h\overline{\nu}$
Optical energy	E	Photon number	$\overline{n} = E/h\overline{\nu}$

Spectral Densities

For polychromatic light of nonnegligible bandwidth, it is useful to define spectral densities of the classical intensity, power, and energy, and their respective quantum counterparts: spectral photon-flux density, spectral photon flux, and spectral photon number:

Classical		Quantum	
I_ν	(W/cm^2-Hz)	$\phi_\nu = I_\nu/h\nu$	(photons/s-cm^2-Hz)
P_ν	(W/Hz)	$\Phi_\nu = P_\nu/h\nu$	(photons/s-Hz)
E_ν	(J/Hz)	$\overline{n}_\nu = E_\nu/h\nu$	(photons/Hz)

As an example, $P_\nu\, d\nu$ represents the optical power in the frequency range between ν and $\nu + d\nu$ while $\Phi_\nu\, d\nu$ represents the flux of photons in that frequency range.

Time-Varying Light

If the light intensity is time varying, it follows that the mean photon-flux density given in (13.2-1) is a function of time, i.e.,

$$\phi(\mathbf{r},t) = \frac{I(\mathbf{r},t)}{h\overline{\nu}}\,. \tag{13.2-7}$$

Mean Photon-Flux Density

The mean photon flux and optical power are then also functions of time,

$$\Phi(t) = \int_A \phi(\mathbf{r},t)\,dA = \frac{P(t)}{h\overline{\nu}}\,, \tag{13.2-8}$$

Mean Photon Flux

where

$$P(t) = \int_A I(\mathbf{r},t)\,dA\,. \tag{13.2-9}$$

Consequently, the mean number of photons registered in a time interval between $t = 0$ and T, which is obtained by integrating the photon flux, also varies with time,

$$\overline{n} = \int_0^T \Phi(t)\,dt = \frac{E}{h\overline{\nu}}\,, \tag{13.2-10}$$

Mean Photon Number

where the mean optical energy (intensity integrated over time and area) is given by

$$E = \int_0^T P(t)\,dt = \int_0^T \int_A I(\mathbf{r},t)\,dA\,dt\,. \tag{13.2-11}$$

B. Randomness of Photon Flow

When the classical intensity $I(\mathbf{r},t)$ is constant, the time and position at which a single photon is detected are governed by (13.1-15), which provides that the probability density of detecting that photon at the space–time point $(\mathbf{r},t)$ is proportional to $I(\mathbf{r},t)$. The classical electromagnetic intensity $I(\mathbf{r},t)$ governs the behavior of photon streams as well as single photons, but the interpretation ascribed to $I(\mathbf{r},t)$ differs:

> *For photon streams, the classical intensity $I(\mathbf{r},t)$ determines the mean photon-flux density $\phi(\mathbf{r},t)$. The fluctuations of $\phi(\mathbf{r},t)$ are determined by the properties of the light source that emits the photons.*

Consider a detector that integrates over space, such as that illustrated in Fig. 13.2-1. If the intensity I is constant in time, then so too is the power P. The mean photon-flux density is then $\phi = I/h\overline{\nu}$ and the mean photon flux is $\Phi = P/h\overline{\nu}$. Nevertheless, the times at which the photons arrive are random, as illustrated schematically in Fig. 13.2-3(a); the statistical properties are determined by the nature of the source emitting the photons, as set forth in Sec. 13.2C. An example of how random photon arrivals might arise may be understood from the following example. Consider a source of optical power $P = 1$ nW that emits at a wavelength $\lambda_o = 1\ \mu$m so it delivers an *average* photon flux of $\Phi = 5$ photons/ns or 0.005 photons/ps. Since only integral numbers of photons may be detected, this signifies that if 10^5 time intervals are examined, each of duration $T = 1$ ps, then most intervals will be empty (zero photons will be registered), about 500 intervals will contain one photon, and very few intervals will contain two or more photons.

If the optical power $P(t)$ does vary with time, the mean density of photon detections will follow the function $P(t)$, as schematically illustrated in Fig. 13.2-3(b). The mean

photon flux $\Phi(t) = P(t)/h\overline{\nu}$ accommodates the fact that there are more photon arrivals when the power is large than when it is small. This variation is in addition to the fluctuations in photon occurrence times associated with the character of the source.

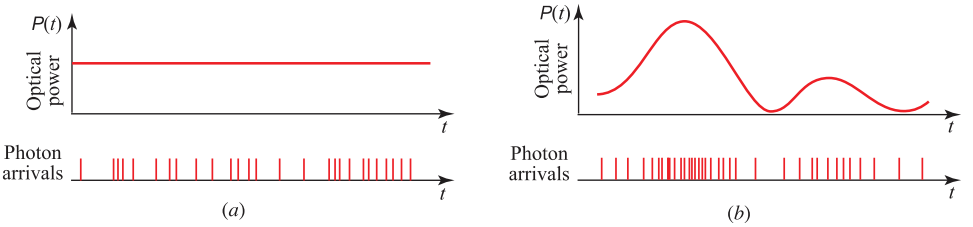


Figure 13.2-3 (a) Constant optical power and a sample function of the randomly arriving photons, whose statistical properties are determined by the nature of the source. (b) Time-varying optical power and a sample function of the randomly arriving photons, whose statistical properties are determined both by the fluctuations of the optical power and by the nature of the source, as considered in Sec. 13.2C.

The image of Max Planck in Fig. 13.2-2 illustrates analogous behavior in the spatial domain. The locations of the detected photons generally follow the classical intensity distribution, with a high photon density where the intensity is large and a low photon density where the intensity is small. But there is considerable graininess (noise) in the image corresponding to the fluctuations in photon occurrence positions associated with the source of illumination. These fluctuations are most discernible when the mean photon-flux density is small, as is the case in Fig. 13.2-2. When the mean photon-flux density becomes large everywhere, as it is in the image of Max Planck on page 514, the graininess disappears and the classical intensity distribution is recovered.

C. Photon-Number Statistics

An understanding of photon-number statistics is important for applications such as the reduction of noise in weak images and the optimization of optical information transmission. In an optical fiber communications system, for example, information is carried in the form of pulses of light (Chapter 25). Only the mean number of photons per pulse is controllable at the source; the actual number of photons emitted is unpredictable and varies from pulse to pulse, resulting in errors in the transmission of information.

The statistical distribution of the number of photons depends on the nature of the light source and must generally be treated in the context of quantum optics, as described briefly in Sec. 13.3. Under certain conditions, however, the arrival of photons may be regarded as the independent occurrences of a sequence of random events at a rate equal to the photon flux, which is proportional to the optical power. The optical power may be deterministic (as with coherent light) or a time-varying random process (as with partially coherent light). For partially coherent light (Chapter 12), the power fluctuations are correlated, so that the photon arrivals will no longer form a sequence of independent events and the photon statistics are significantly altered.

Coherent Light

Coherent light has constant optical power P. The corresponding mean photon flux $\Phi = P/h\overline{\nu}$ (photons/s) is also constant, but the actual photon registration times are random, as portrayed schematically in Fig. 13.2-4. Given a time interval of duration T, called the **counting time**, let the n signify the number of detected photons, called the **photon number**. We already know from (13.2-6) that the **mean photon number** is $\overline{n} = \Phi T = PT/h\overline{\nu}$. We now seek to establish the **photon-number distribution** $p(n)$, i.e., the probability $p(0)$ of detecting zero photons, the probability $p(1)$ of detecting one photon, etc., in the counting time T.

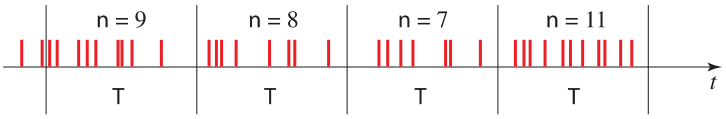


Figure 13.2-4 Random arrival of photons for a coherent light source of power P. Consecutive counting times of duration T are indicated. Though the optical power is constant, the photon number n observed in each counting time is random.

An expression for the photon-number distribution, $p(n)$ vs. n, suitable for coherent light can be derived under the assumption that the photon registrations are statistically independent events, as derived below. The result, known as the **Poisson distribution**, takes the form

$$p(n) = \frac{\bar{n}^{\,n} \exp(-\bar{n}\,)}{n!}, \qquad n = 0, 1, 2, \ldots . \tag{13.2-12}$$

Poisson Distribution

Equation (13.2-12) is displayed on a semilogarithmic plot in Fig. 13.2-5 for several values of the mean photon number $\bar{n}$.

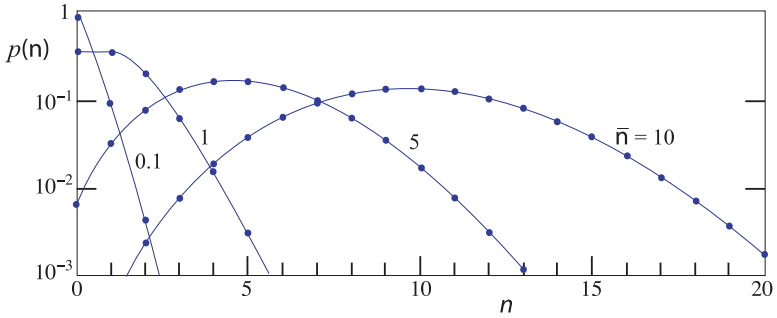


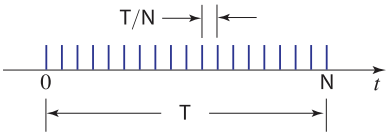
Figure 13.2-5 Semilogarithmic plot of the Poisson photon-number distribution, $p(n)$ vs. n, for several values of the mean photon number $\bar{n}$. The curves become progressively broader as $\bar{n}$ increases.

The Poisson photon-number distribution is suitable for describing the photon statistics of the coherent light emitted by an ideal, amplitude-stabilized, single-mode laser operated well above its threshold of oscillation (Chapter 16). This same result emerges in the context of quantum optics (Sec. 13.3A). Moreover, the Poisson distribution provides a good approximation for describing the photon statistics of a number of other sources of light, such as multimode thermal light.

□ **Derivation of the Poisson Distribution.** Divide the time interval T shown in Fig. 13.2-4 into a large number N of subintervals, each of sufficiently short duration T/N such that each subinterval carries one photon with probability $p = \bar{n}/N$ and zero photons with probability $1-p$. The probability of finding n independent photons in the N subintervals, like the flips of a biased coin, then follows the binomial distribution:

$$p(n) = \frac{N!}{n!\,(N-n)!}\, p^{\,n} (1-p)^{N-n}$$

$$= \frac{N!}{n!\,(N-n)!} \left(\frac{\bar{n}}{N}\right)^{n} \left(1 - \frac{\bar{n}}{N}\right)^{N-n}.$$



In the limit as the number of subintervals $N \to \infty$, we have $N!/(N-n)!\,N^{\,n} \to 1$, $(1-\bar{n}/N)^{-n} \to 1$, and $(1-\bar{n}/N)^{N} \to \exp(-\bar{n}\,)$, which yields (13.2-12). ■

Photon-Number Mean and Variance

The mean and variance are typically used to characterize a random variable. The mean photon number is expressed as

$$\overline{n} = \sum_{n=0}^{\infty} n\, p(n), \tag{13.2-13}$$

while the variance, which is the average of the squared deviation from the mean, is given by

$$\sigma_n^2 = \sum_{n=0}^{\infty} (n - \overline{n})^2\, p(n). \tag{13.2-14}$$

The standard deviation σ_n, which is the square root of the variance, is a measure of the width of the distribution. The quantities $p(n)$, $\overline{n}$, and σ_n are collectively called **photon-number statistics**. Though the distribution $p(n)$ contains information beyond the mean and variance, these two parameters provide a rough outline of its nature.

It is not difficult to show, by inserting (13.2-12) into (13.2-13) and (13.2-14), that the mean of the Poisson distribution is indeed $\overline{n}$ and that its variance is equal to its mean:

$$\boxed{\sigma_n^2 = \overline{n}\,.} \tag{13.2-15}$$

Variance
Poisson Distribution

Taking $\overline{n} = 100$, for example, the standard deviation is $\sigma_n = 10$, which signifies that the observation of 100 photons on average is accompanied by an uncertainty of roughly ± 10 photons.

Signal-to-Noise Ratio

Photon-number randomness constitutes a fundamental source of noise that must be contended with when using light to transmit a signal. A useful measure of the performance of an information transmission system is the photon-number-based signal-to-noise ratio (SNR). Representing the signal by its mean $\overline{n}$, and the noise by its standard deviation σ_n, the SNR is defined as

$$\text{SNR} = \frac{(\text{mean})^2}{\text{variance}} = \frac{\overline{n}^2}{\sigma_n^2}\,. \tag{13.2-16}$$

If the light obeys Poisson photon-number statistics, then $\sigma_n^2 = \overline{n}$ from (13.2-15), so that

$$\boxed{\text{SNR} = \overline{n}\,.} \tag{13.2-17}$$

Signal-to-Noise Ratio
Poisson Distribution

The Poisson signal-to-noise ratio increases linearly with the mean photon number.

Though the SNR is a useful measure of the randomness of a signal, in some applications it is necessary to make use of the full probability distribution. For example, an important measure of the performance of a digital fiber optic communications system is the probability of error. If such a system is used to transmit information using bits with a mean photon number of, say $\overline{n} = 20$, (13.2-12) dictates that the probability of no photons being received is $p(0) \approx 2 \times 10^{-9}$. Since the receipt of zero photons represents an error, as discussed in Sec. 25.2B, the full probability distribution is required to calculate system performance.

Thermal Light

When the photon arrival times are not independent, as is the case for thermal light, the photon-number statistics differ from Poisson. Thermal light is generated in an optical resonator whose walls are maintained at a fixed temperature T and whose atoms emit photons into the modes of the resonator. In accordance with the laws of statistical mechanics under conditions of thermal equilibrium, the probability of occupancy of energy level E_n in a system satisfies the **Boltzmann probability distribution**

$$P(E_n) \propto \exp\left(-\frac{E_n}{kT}\right). \tag{13.2-18}$$

Boltzmann Distribution

This exponential distribution is sketched in Fig. 13.2-6 with $P(E_n)$ plotted along the abscissa. The occupancy of each energy level is random and higher energies are relatively less probable than lower energies. The distribution is parameterized by kT, where k is **Boltzmann's constant** ($k = 1.38 \times 10^{-23}$ J/°K). At $T = 300°$ K (room temperature), $kT = 26$ meV = 4.14 zJ = 209 cm^{-1}, as illustrated in Fig. 13.1-2. The origin of the Boltzmann distribution is discussed in Sec. 14.2.

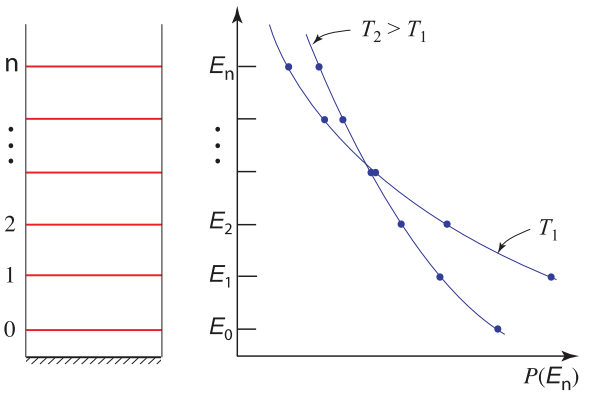


Figure 13.2-6 Boltzmann probability distribution $P(E_n)$ (plotted along the abscissa) versus energy E_n (plotted along the ordinate) for two values of the temperature T. The lower the temperature, the less likely that higher energy levels are occupied. The allowed energy levels of a collection of photons in a mode of frequency ν are illustrated at left.

We now assume that the collection of photons in a mode of frequency ν inside the resonator behaves as a gas in thermal equilibrium at temperature T that obeys the Boltzmann distribution, and that the mode has allowed energy levels given by $E_n = (n + \frac{1}{2})h\nu$, as provided in (13.1-5). It then follows that the probability of finding n photons in the mode is given by

$$p(n) \propto \exp\left(-\frac{nh\nu}{kT}\right) = \left[\exp\left(-\frac{h\nu}{kT}\right)\right]^n, \qquad n = 0, 1, 2, \ldots. \tag{13.2-19}$$

Normalizing (13.2-19) so that it sums to unity, i.e., imposing the condition $\sum_{n=0}^{\infty} p(n) = 1$, provides the normalization constant $[1 - \exp(-h\nu/kT)]$. The zero-point energy $E_0 = \frac{1}{2}h\nu$ disappears into the normalization and does not affect the results.

The probability distribution for the number of photons n in a resonator mode of frequency ν given in (13.2-19) is most simply written in terms of the mean photon number $\bar{n}$ as

$$p(n) = \frac{1}{\bar{n}+1}\left(\frac{\bar{n}}{\bar{n}+1}\right)^n, \tag{13.2-20}$$

Bose–Einstein Distribution

where

$$\overline{n} = \frac{1}{\exp(h\nu/kT) - 1}\,, \tag{13.2-21}$$

which has been determined from (13.2-13). It is reassuring that (13.2-21) accords with the mean photon number calculated in (14.4-7) for a collection of photons interacting with atoms in thermal equilibrium, as will be seen in Sec. 14.4A. In the parlance of probability theory, the distribution displayed in (13.2-20) is known as the **geometric distribution** since $p(n)$ is a geometrically decreasing function of n. In the physics literature, it is generally referred to as the **Bose–Einstein distribution** since it was first set forth by Bose based on a statistical argument for counting the states of indistinguishable particles such as photons. Einstein recognized that (13.2-20) was also applicable for describing bosons whose numbers are conserved, and he predicted the possibility of a condensation to the lowest energy state in a bosonic atomic gas cooled below a critical temperature (Sec. 14.3F).

The Bose–Einstein distribution is displayed on a semilogarithmic plot in Fig. 13.2-7 for several values of the mean photon number $\overline{n}$ [or, equivalently, several values of the temperature T via (13.2-21)]. Its exponential character is apparent from the straight-line behavior on this semilogarithmic plot. Comparing Fig. 13.2-7 with Fig. 13.2-5 for the Poisson distribution demonstrates that the photon-number distributions for thermal light decrease monotonically from $n = 0$ and are far broader than those for coherent light.

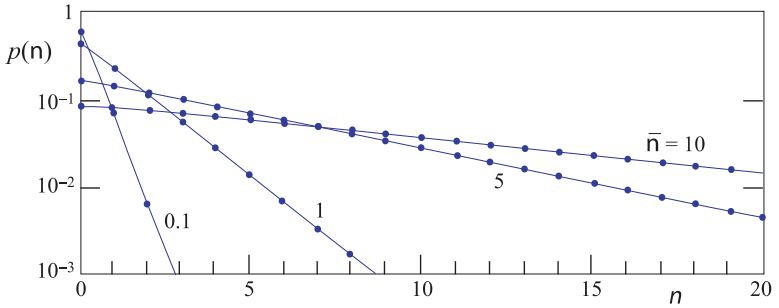


Figure 13.2-7 Semilogarithmic plot of the Bose–Einstein photon-number distribution, $p(n)$ vs. n, for several values of the mean photon number $\overline{n}$. The curves broaden substantially as $\overline{n}$ increases.

The photon-number variance of the Bose–Einstein distribution, which is readily calculated via (13.2-14), turns out to be

$$\boxed{\sigma_n^2 = \overline{n} + \overline{n}^2,} \tag{13.2-22}$$

Variance
Bose–Einstein Distribution

where $\overline{n}$ is the photon-number mean. Comparing the Bose–Einstein and Poisson variances given in (13.2-22) and (13.2-15), respectively, reveals that, for $\overline{n} > 1$, the former grows quadratically with $\overline{n}$ while the latter grows linearly. The photon-number fluctuations of the Bose–Einstein distribution are clearly far greater than those of the Poisson distribution, as is apparent in the comparison of Figs. 13.2-7 and 13.2-5. This large variability is consistent with the random nature of thermal light, as described in Sec. 12.1. The noisiness of the Bose–Einstein distribution is crisply highlighted in its signal-to-noise ratio, which, in accordance with (13.2-16), is given by

$$\text{SNR} = \frac{\overline{n}}{\overline{n} + 1}\,. \tag{13.2-23}$$

The Bose–Einstein SNR always remains smaller than unity no matter how large $\overline{n}$, confirming that thermal light is generally too noisy to be used for the transmission of information.

EXERCISE 13.2-1

Average Energy of a Resonator Mode in Thermal Equilibrium. Show that the average energy of a resonator mode of frequency ν, under conditions of thermal equilibrium at temperature T, is given by

$$\overline{E} = kT\frac{h\nu/kT}{\exp(h\nu/kT)-1}. \tag{13.2-24}$$

Sketch the dependence of $\overline{E}$ on $h\nu$ for several values of kT. Use a Taylor-series expansion of the denominator to obtain an expression for $\overline{E}$ in the limit $h\nu/kT \ll 1$. Explain the result on a physical basis.

*Doubly Stochastic Photon-Number Statistics

As indicated earlier, coherent light has constant intensity $I(\mathbf{r},t)$, constant optical power P, and constant photon flux $\Phi = P/h\overline{\nu}$. The arriving photons behave as independent events with a Poisson photon-number distribution $p(n) = \overline{n}^n e^{-\overline{n}}/n!$, where the mean photon number $\overline{n} = \Phi T = PT/h\overline{\nu}$ is constant. However, if the light is partially coherent and its intensity varies in time, then so too does the optical power [as portrayed in Fig. 13.2-3(*b*)], the photon flux, and the mean photon number $\overline{n}$. In accordance with (13.2-10) and (13.2-11), in that case the mean photon number, which we denote as w rather than $\overline{n}$ for notational convenience, can be expressed as

$$w \equiv \overline{n} = \frac{1}{h\overline{\nu}}\int_0^T P(t)\,dt = \frac{1}{h\overline{\nu}}\int_0^T\int_A I(\mathbf{r},t)\,dA\,dt. \tag{13.2-25}$$

The integrated intensity w, which has units of mean photon number and is thus dimensionless, varies in time for partially coherent light.

Variations in the mean photon number arising from intensity fluctuations cause the photon-number distribution to depart from Poisson behavior, as we now show. If the fluctuations of w are described by a probability density function $p(w)$, the new photon-number probability distribution is obtained by averaging the Poisson distribution conditioned on w being constant, $p(n|w) = w^n e^{-w}/n!$, over the permitted values of w dictated by $p(w)$. The alert reader will notice that we have co-opted the symbol n for the new photon number. The resultant photon-number distribution then takes the form

$$\boxed{p(n) = \int_0^\infty \frac{w^n e^{-w}}{n!}\,p(w)\,dw,} \tag{13.2-26}$$

Mandel's Formula

which is known as **Mandel's formula**. Equation (13.2-26) is a **doubly stochastic photon-number distribution** by virtue of its two contributing sources of randomness: 1) the random arrivals of the photons themselves, which locally behave in Poisson fashion; and 2) the integrated-intensity fluctuations arising from the partially coherent nature of the illumination. An underlying sequence of random photon arrivals that leads to doubly stochastic photon-number statistics is known as a **doubly stochastic Poisson process (DSPP)**.

The photon-number mean and variance for the doubly stochastic photon-number distribution are obtained by using (13.2-13) and (13.2-14) in conjunction with (13.2-26); the results are

$$\bar{n} = \overline{w} \tag{13.2-27}$$

and

$$\sigma_n^2 = \bar{n} + \sigma_w^2 , \tag{13.2-28}$$

respectively. Here σ_w^2 signifies the variance of w. The photon-number variance thus comprises two contributions: 1) the basic Poisson contribution $\bar{n}$; and 2) a (positive) contribution arising from the intensity fluctuations. It is worth noting that the theory of photon statistics presented here is applicable only for **classical light**; a more general theory encompassing **nonclassical light** requires a quantum approach (Sec. 13.3).

An instructive example makes use of an integrated intensity that obeys the exponential probability density function

$$p(w) = \begin{cases} \dfrac{1}{\overline{w}} \exp\left(-\dfrac{w}{\overline{w}}\right), & w \geq 0 \\ 0, & w < 0. \end{cases} \tag{13.2-29}$$

Equation (13.2-29) is appropriate for quasi-monochromatic light whose independent and identically distributed real and imaginary complex-field-amplitude components are Gaussian. It is applicable for a source of light whose spectral width is sufficiently small such that its coherence time τ_c is much greater than the counting time T, and whose coherence area A_c is much greater than the detector integration area A (see Chapter 12). The photon-number distribution $p(n)$ that corresponds to (13.2-29) is determined by substituting this result into (13.2-26) and evaluating the integral. The outcome turns out to be the Bose–Einstein distribution given in (13.2-20). Hence, the photon statistics of the Gaussian-distributed optical field described above are identical to those of single-mode thermal light. Multimode thermal light emerges when the counting time T and photodetector integration area A are not small (Probs. 13.2-6–13.2-8).

D. Random Partitioning of Photon Streams

A photon stream is said to be partitioned when it is subjected to the removal of some of its photons. The process is called **random partitioning** when the removed photons are diverted and **random selection** when they are annihilated. There are numerous ways in which this can occur. Perhaps the simplest example of random partitioning is provided by an ideal lossless beamsplitter. Photons are randomly selected to exit either of the two output ports (Fig. 13.2-8). An example of random selection is provided by the action of a photodetector. Photons incident on the photosensitive material are selected either to be absorbed and create photoelectrons or to pass through it and be lost.

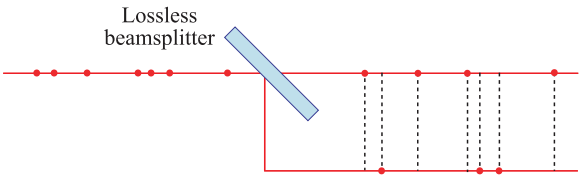


Figure 13.2-8 Random partitioning of a stream of photons by a beamsplitter.

We restrict the treatment presented here to situations in which the probability of each photon being partitioned behaves in accordance with an independent **Bernoulli trial** (coin toss). In terms of the beamsplitter, this is satisfied if a photon stream impinges on only one of its input ports (Fig. 13.2-8). This eliminates the possibility of interference, which in general invalidates the independent-trial assumption (see, e.g., Example 13.3-3). The results derived below apply to both random partitioning and random selection.

Consider a lossless beamsplitter of transmittance $\mathcal{T}$ and reflectance $\mathcal{R} = 1 - \mathcal{T}$. The result of a single photon impinging on this device was examined in Sec. 13.1C, where it was shown that the probability of a photon being transmitted is equal to the transmittance of the beamsplitter $\mathcal{T}$ and the probability of the photon being reflected is equal to $1 - \mathcal{T}$ (Fig. 13.1-6). We now consider a photon stream of mean flux Φ is incident on the beamsplitter, so that a mean number of photons $\overline{n} = \Phi T$ impinges on it in the time interval T. The mean number of transmitted and reflected photons is then $\mathcal{T}\overline{n}$ and $(1 - \mathcal{T})\overline{n}$, respectively.

We proceed to determine how the photon-number statistics of the incident photon stream are modified upon partitioning. If the incident stream consists of precisely n photons, the probability $p(m)$ that m photons are transmitted is the same as that of flipping a biased coin n times and obtaining m heads, when the probability of obtaining a head is $\mathcal{T}$. From elementary probability theory we know that $p(m)$ is characterized by the binomial distribution

$$p(m) = \binom{n}{m} \mathcal{T}^m (1 - \mathcal{T})^{n-m}, \qquad m = 0, 1, \ldots, n, \tag{13.2-30}$$

where $\binom{n}{m} = n!/m!\,(n - m)!$. By symmetry, the results for the reflected photons are identical, with $1 - \mathcal{T}$ replacing $\mathcal{T}$. The statistics of the binomial distribution yield the mean number of transmitted photons

$$\overline{m} = \mathcal{T} n, \tag{13.2-31}$$

and its variance

$$\sigma_m^2 = \mathcal{T}(1 - \mathcal{T})n = (1 - \mathcal{T})\,\overline{m}. \tag{13.2-32}$$

The signal-to-noise ratio specified in (13.2-16) is thus $\text{SNR} = \overline{m}^2/\sigma_m^2 = \overline{m}/(1 - \mathcal{T})$, which increases linearly with the mean number of transmitted photons $\overline{m}$. When the incident wave is strong, the photons will therefore be partitioned between the transmitted and reflected beams in good agreement with $\mathcal{T}$ and $(1 - \mathcal{T})$, respectively, confirming that the classical-optics result is recovered.

The expressions provided above for a fixed number of incident photons permit the photon-number statistics of the partitioned stream to be determined. The calculation proceeds by recognizing that in the general case the number of photons n at the input to the beamsplitter is random rather than fixed. Let the probability be $p_0(n)$ that in a specified time interval there are n photons present at the beamsplitter input. The photon-number probability distribution for the transmitted stream will then be a weighted sum of binomial distributions, with the weighting established by the probability of n photons being present. The photon-number distribution $p(m)$ at the output of the beamsplitter, for an input photon-number distribution $p_0(n)$, is therefore $p(m) = \sum_n p(m|n)\,p_0(n)$, where the observation of m photons conditioned on n being fixed is the binomial distribution $p(m|n) = \binom{n}{m} \mathcal{T}^m (1-\mathcal{T})^{n-m}$. Finally, then, we obtain a formula that yields $p(m)$ in terms of $p_0(n)$ and $\mathcal{T}$:

$$p(m) = \sum_{n=m}^{\infty} \binom{n}{m} \mathcal{T}^m (1 - \mathcal{T})^{n-m} p_0(n). \tag{13.2-33}$$

Photon-Number Distribution Under Random Partitioning

This formula is also applicable for the *detection* of photons, as considered in Sec. 19.6A.

When the input photon-number distribution $p_0(n)$ is Poisson (for coherent light) or Bose–Einstein (for single-mode thermal light), the results turn out to be quite simple: the form of the partitioned photon-number distribution $p(m)$ exactly matches that of the incident photon-number distribution $p_0(n)$. Thus, single-mode laser light transmitted through a beamsplitter remains Poisson and thermal light remains Bose–Einstein, although of course the photon-number mean is reduced by the factor $\mathcal{T}$. In contrast, photon-number-squeezed light (Sec. 13.3C) does not retain its form under random partitioning, a property that is responsible for its lack of robustness. In particular, number-state light, which comprises a deterministic photon number, obeys the binomial photon-number distribution after partitioning.

The signal-to-noise ratio for m is readily determined for partitioned photon streams. For coherent light and single-mode thermal light, the results are, respectively,

$$\text{SNR} = \begin{cases} \mathcal{T}\,\overline{n} & \text{coherent light} \qquad (13.2\text{-}34) \\ \dfrac{\mathcal{T}\,\overline{n}}{\mathcal{T}\,\overline{n}+1} & \text{thermal light.} \qquad (13.2\text{-}35) \end{cases}$$

Since $\mathcal{T} \leq 1$ it is clear that random partitioning decreases the signal-to-noise ratio. Another way of stating this is to say that random partitioning introduces noise.

*13.3 QUANTUM STATES OF LIGHT

The number of photons in an electromagnetic mode is generally a random quantity. In this section it will be shown that in the context of quantum optics the electric field itself is also generally random. Consider a monochromatic plane-wave electromagnetic mode in a volume V, described by the electric field $\mathcal{E}(\mathbf{r},t) = \text{Re}\{\mathbf{E}(\mathbf{r},t)\}$, where

$$\mathbf{E}(\mathbf{r},t) = A\exp(-j\mathbf{k}\cdot\mathbf{r})\,\exp(j2\pi\nu t)\,\widehat{\mathbf{e}}. \qquad (13.3\text{-}1)$$

According to classical electromagnetic optics, as provided in (13.1-3), the energy of the mode is fixed at $\frac{1}{2}\epsilon|A|^2V$. We define a complex variable a, such that $\frac{1}{2}\epsilon|A|^2V = h\nu|a|^2$, thereby allowing $|a|^2$ to be interpreted as the energy of the mode in units of photon number. The electric field may then be written as

$$\mathbf{E}(\mathbf{r},t) = \sqrt{\frac{2h\nu}{\epsilon V}}\;a\,\exp(-j\mathbf{k}\cdot\mathbf{r})\,\exp(j2\pi\nu t)\,\widehat{\mathbf{e}}, \qquad (13.3\text{-}2)$$

where the complex variable a determines the complex amplitude of the field.

In classical electromagnetic optics, $a\,\exp(j2\pi\nu t)$ is a rotating phasor whose projection on the real axis determines the sinusoidal electric field, as portrayed in Fig. 13.3-1. The real and imaginary parts of $a = x + jp$, which are $x = \text{Re}\{a\}$ and $p = \text{Im}\{a\}$, respectively, are termed the quadrature components of the phasor a because they are a quarter cycle $(90°)$ out of phase with each other. They determine the amplitude and phase of the sine wave that represents the temporal variation of the electric field. The rotating phasor $a\,\exp(j2\pi\nu t)$ also describes the motion of a harmonic oscillator; the real component x is proportional to position and the imaginary component p to momentum. From a mathematical point of view, then, a *classical* monochromatic mode of the electromagnetic field and a classical harmonic oscillator behave identically. A parallel argument can be constructed to show that a *quantum* monochromatic electromagnetic mode and a one-dimensional quantum-mechanical harmonic oscillator also have identical behavior. To facilitate the comparison, we review the quantum theory of a simple harmonic oscillator before proceeding.

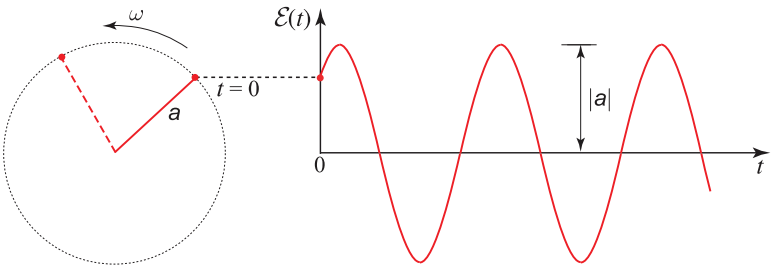


Figure 13.3-1 The real and imaginary parts of the variable $a\exp(j2\pi\nu t)$, which govern the complex amplitude of a classical electromagnetic mode of frequency ν. The time dynamics are identical to those of a classical harmonic oscillator with angular frequency $\omega = 2\pi\nu$.

Quantum Theory of the Harmonic Oscillator

A particle of mass m, position x, momentum p, and potential energy $V(x) = \frac{1}{2}\kappa x^2$, where κ is the elastic constant, represents a one-dimensional harmonic oscillator of total energy $\frac{1}{2}p^2/m + \frac{1}{2}\kappa x^2$ and oscillation frequency $\omega = \sqrt{\kappa/m}$. Without loss of generality, we take $m = 1$ so that the energy is $\frac{1}{2}(p^2 + \omega^2 x^2)$.

In accordance with the laws of quantum mechanics, the behavior of this quantum system in a stationary state is described by a complex wavefunction $\psi(x)$ that satisfies the time-independent Schrödinger equation [see (14.1-3)],

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = E\psi(x), \tag{13.3-3}$$

where E is the energy of the particle. The solutions of the Schrödinger equation for the harmonic oscillator, for which $V(x) = \frac{1}{2}\omega^2 x^2$, give rise to discrete energy values given by

$$E_n = \left(n + \tfrac{1}{2}\right)h\nu, \qquad n = 0, 1, 2, \ldots . \tag{13.3-4}$$

Adjacent energy levels are seen to be separated by a quantum of energy $h\nu = \hbar\omega$. The corresponding wavefunctions $\psi_n(x)$ are normalized Hermite–Gaussian functions,

$$\psi_n(x) = \frac{1}{\sqrt{2^n\, n!}}\left(\frac{\omega}{\pi\hbar}\right)^{1/4}\mathbb{H}_n\left[\sqrt{\frac{\omega}{\hbar}}\,x\right]\exp\left(-\frac{\omega x^2}{2\hbar}\right), \tag{13.3-5}$$

where $\mathbb{H}_n(x)$ is the Hermite polynomial of order n [see (3.3-6)–(3.3-8) and (3.3-11)].

An arbitrary wavefunction $\psi(x)$ may be expanded in terms of the set of orthonormal eigenfunctions $\{\psi_n(x)\}$ in terms of the superposition $\psi(x) = \sum_n c_n\psi_n(x)$. Given the wavefunction $\psi(x)$, which governs the state of the system, the behavior of the particle is determined as follows:

- The probability $p(n)$ that the harmonic oscillator carries n quanta of energy is given by the coefficient $|c_n|^2$.
- The probability density of finding the particle at the position x is given by $|\psi(x)|^2$.
- The probability density that the momentum of the particle is p is given by $|\phi(p)|^2$, where $\phi(p)$ is proportional to the inverse Fourier transform of $\psi(x)$ evaluated at the (spatial) frequency p/h,

$$\phi(p) = \frac{1}{\sqrt{h}}\int_{-\infty}^{\infty}\psi(x)\exp\left(j2\pi\frac{p}{h}x\right)dx. \tag{13.3-6}$$

As shown in Sec. A.2 of Appendix A, the Fourier-transform relation between

$\psi(x)$ and $\phi(p)$ indicates that there is an uncertainty relation between the power-RMS widths of x and p/h given by

$$\frac{\sigma_x \sigma_p}{h} \geq \frac{1}{4\pi} \qquad \text{or} \qquad \sigma_x \sigma_p \geq \frac{\hbar}{2}. \tag{13.3-7}$$

This is the well-known Heisenberg position–momentum uncertainty relation provided in (A.2-7) of Appendix A.

Analogy Between an Optical Mode and a Harmonic Oscillator

The energy of an electromagnetic mode is $h\nu|a|^2 = h\nu(x^2 + p^2)$. The analogy with a harmonic oscillator of energy $\frac{1}{2}(\omega^2 x^2 + p^2)$ is established by effecting the connections

$$x = \left(1/\sqrt{2\hbar\omega}\right)\omega x \qquad \text{and} \qquad p = \left(1/\sqrt{2\hbar\omega}\right)p. \tag{13.3-8}$$

The modal energy then becomes $h\nu(x^2 + p^2) = \frac{1}{2}(\omega^2 x^2 + p^2)$, which leads to the harmonic-oscillator energy levels provided in (13.3-4). Because the analogy is complete, we conclude that the energy of a quantum, monochromatic electromagnetic mode, like that of a one-dimensional, quantum-mechanical harmonic oscillator, is quantized to the values $E_n = (n + \frac{1}{2})\,h\nu$, as initially suggested in (13.1-5). With proper scaling normalization factors, the behavior of the position and momentum of the harmonic oscillator, x and p, also describe the quadrature components of the electromagnetic field, x and p, respectively.

Properties of a Quantum Electromagnetic Mode

An quantum electromagnetic mode of frequency ν is described by a complex wavefunction $\psi(x)$ that governs the uncertainties of the quadrature components x and p of the electromagnetic field, as well as the statistics of the number of photons in the mode.

- The probability $p(n)$ that the mode contains n photons is given by $|c_n|^2$, where the c_n are coefficients of the expansion of $\psi(x)$ in terms of the eigenfunctions $\psi_n(x)$, i.e., $\psi(x) = \sum_n c_n \psi_n(x)$.
- The probability density functions of the quadrature components x and p are given by $|\psi(x)|^2$ and $|\phi(p)|^2$, respectively, where $\psi(x)$ and $\phi(p)$ are related by

$$\phi(p) = \frac{1}{\sqrt{\pi\omega}} \int_{-\infty}^{\infty} \psi(x) \exp(j2px)\, dx. \tag{13.3-9}$$

  This equation is derived from (13.3-6) by use of the transformation (13.3-8) and by noting that $|\psi(x)|^2$ and $|\phi(p)|^2$ integrate to unity.
- The uncertainty relation between the power-RMS widths of the quadrature components is given by

$$\boxed{\sigma_x \sigma_p \geq \tfrac{1}{4},} \tag{13.3-10}$$

Quadrature Uncertainty

  so that these components cannot be simultaneously determined with arbitrary precision.

A. Coherent States

The quadrature uncertainty product $\sigma_x\sigma_p$ attains its minimum value of $1/4$ when the function $\psi(x)$ is Gaussian (Sec. A.2 of Appendix A). In that case

$$\psi(x) \propto \exp\left[-(x-\alpha_x)^2\right], \tag{13.3-11}$$

so that its Fourier transform is also Gaussian,

$$\phi(p) \propto \exp\left[-(p-\alpha_p)^2\right]. \tag{13.3-12}$$

Here, α_x and α_p are arbitrary values that represent the means of x and p, respectively. The quadrature uncertainties, determined from $|\psi(x)|^2$ and $|\phi(p)|^2$, are then given by

$$\sigma_x = \sigma_p = 1/2. \tag{13.3-13}$$

Under these conditions the electromagnetic field is said to be in a **coherent state**. The one-standard-deviation range of uncertainty in the quadrature components x and p, as well as in the complex amplitude a and in the electric field $\mathcal{E}(t)$, are illustrated in Fig. 13.3-2 for coherent-state light. The uncertainties are most pronounced when α_x and α_p are small. The squared-magnitude coefficients $|c_n|^2$ for the expansion of $\psi(x)$ in the Hermite–Gaussian basis are $\overline{n}^n \exp(-\overline{n})/n!$, where $\overline{n} = \alpha_x^2 + \alpha_p^2$, so the photon-number distribution $p(n)$ is Poisson, as suggested in the discussion surrounding (13.2-12). Unlike its status in electromagnetic optics, in the context of quantum optics coherent-state light is *not* deterministic. The coherent state is generated by an ideal, amplitude-stabilized, single-mode laser operated well above its threshold of oscillation.

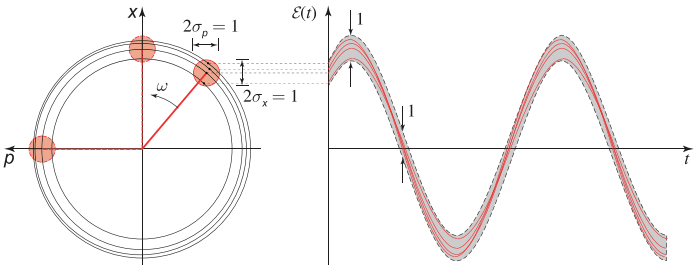


Figure 13.3-2 Quadrature and electric-field uncertainties for the coherent state. Representative values of $\mathcal{E}(t) \propto a\exp(j2\pi\nu t)$ are traced for several arbitrary points within the uncertainty circle; the coefficient of proportionality is chosen to be unity for convenience.

Figure 13.3-3 displays the quadrature uncertainties and time behavior of the electric field for the coherent state when $\alpha_x = \alpha_p = 0$, which is called the **vacuum state**.

B. Quadrature-Squeezed States

Though the uncertainty product $\sigma_x\sigma_p$ cannot be reduced below its minimum value of $1/4$, the uncertainty of one of the quadrature components can be reduced ("squeezed") below $1/2$ but this entails increased uncertainty in the other component. This form of light, which is distinctly nonclassical, is said to be **quadrature-squeezed**. For example, a state for which $\psi(x)$ is a Gaussian function with a squeezed width $\sigma_x = 1/2s$, where $s > 1$, corresponds to a Gaussian function $\phi(p)$ with a stretched width $\sigma_p = s/2$. The product $\sigma_x\sigma_p$ maintains its minimum value of $1/4$, but the uncertainty circle of the phasor a for the coherent state (Fig. 13.3-2) is squeezed into an ellipse, as displayed in Fig. 13.3-4.

The particular quadrature-squeezed state illustrated in Fig. 13.3-4 is known as an amplitude-squeezed state since the phasor amplitude uncertainty is reduced at the

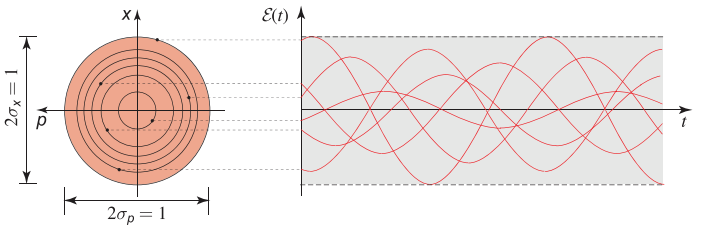


Figure 13.3-3 Quadrature and electric-field uncertainties for the vacuum state. This state is a limiting case for both the coherent state ($\alpha_x = \alpha_p = 0$) and the number state ($n = 0$). The mode carries zero photons and has only the residual zero-point energy $\frac{1}{2}h\nu$. The circle of uncertainty is squeezed into an ellipse for the squeezed vacuum state; however, squeezing the vacuum endows it with a finite mean photon number.

expense of its phase uncertainty. The asymmetry in the uncertainties of the two quadratures is manifested in the time course of the electric field by periodic occurrences of decreased uncertainty, followed each quarter cycle later by occurrences of increased uncertainty. If the field is measured only at those times when its uncertainty is minimal, its noise will be reduced below that of the coherent state. The selection of those times may be achieved by heterodyning the squeezed field with a coherent optical field of appropriate phase (heterodyne receivers are discussed in Sec. 25.4).

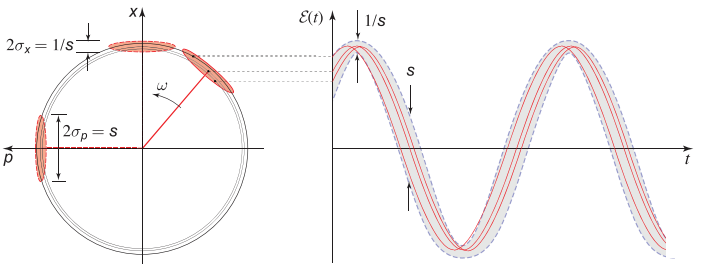


Figure 13.3-4 Quadrature and electric-field uncertainties for a quadrature-squeezed state (specifically, an *amplitude-squeezed state*). The uncertainty circle associated with the coherent state (Fig. 13.3-2) is squeezed into an ellipse of the same area. If the uncertainty ellipse were to be rotated by 90°, so that its long axis lay along the phasor rather than perpendicular to it, the result would be a *phase-squeezed state* since the phase uncertainty would then be reduced at the expense of the amplitude uncertainty. The squeezed vacuum state takes the form of an ellipse at the origin.

Generation and applications of quadrature-squeezed light. Though not robust in the face of optical losses, quadrature-squeezed light has garnered an important niche in precision measurements because of the reduced noisiness of one of its quadratures. The textbook example is the merit of using quadrature-squeezed light in the LIGO gravitational-wave interferometer. In the interferometer configuration considered in Example 2.5-1, for example, all beamsplitter ports are covered by impinging light beams except for the output port, which is exposed to the normal vacuum. Improved interferometer sensitivity is obtained by injecting squeezed vacuum (of appropriate phase) into the output port.[†] Quadrature-squeezed light may be generated in a number of ways, including via optical parametric downconversion in a cavity (Sec. 22.2C).

[†] See, e.g., M. C. Teich and B. E. A. Saleh, Squeezed and Antibunched Light, *Physics Today*, vol. 43, no. 6, pp. 26–34, 1990.

C. Photon-Number-Squeezed States

Quadrature-squeezed light exhibits reduced uncertainty in one of its quadrature field components relative to that of the coherent state. Another form of nonclassical light is **photon-number-squeezed** light, also referred to as **intensity-squeezed** or **sub-Poisson** light. This form of light has a photon-number variance that is squeezed below the coherent-state (Poisson) value, i.e., $\sigma_n^2 < \overline{n}$. Photon-number fluctuations obeying this relation are nonclassical since (13.2-28) cannot be satisfied.

An electromagnetic mode described by the harmonic oscillator eigenstate $\psi(x) = \psi_{n0}(x)$ provides an example of photon-number-squeezed light. This state is referred to as the **number state** because $p(n) = |c_n|^2 = 1$ for a fixed number of photons $n = n_0$; the number state is also called the **Fock state**. Since the number of photons carried by the mode is deterministic, we have $\overline{n} = n_0$ and $\sigma_n^2 = 0$. The case $n_0 = 1$ corresponds to a single photon in the mode. The uncertainties associated with number-state light are illustrated in Fig. 13.3-5. Though the quadrature components as well as the phasor magnitude and phase are all uncertain, the photon number is fixed. For $n_0 = 0$, the number state reduces to the vacuum state displayed in Fig. 13.3-3. Photon-number-squeezing can also be generated using other states, e.g., the binomial state (Prob. 13.3-1).

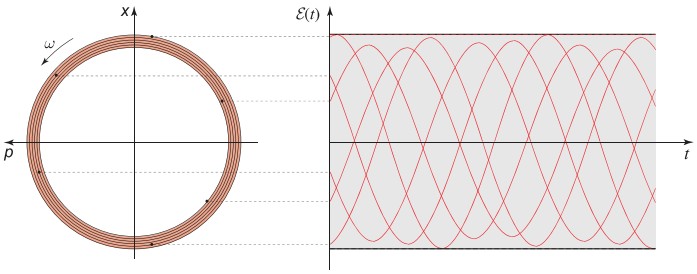


Figure 13.3-5 Representative uncertainties for the number state. The mode contains a fixed number of photons, $n = n_0$. This state is photon-number-squeezed but not quadrature-squeezed.

Generation and applications of photon-number-squeezed light. Just as with quadrature-squeezed light, losses adversely affect photon-number-squeezed light (Sec. 13.2D). Nevertheless, this form of nonclassical light enjoys a range of applications in quantum information processing, communications, computing, and cryptography. Photon-number-squeezed light may be generated via mechanisms that include: 1) the use of feedback to impart anticorrelations to an otherwise Poisson sequence of photons; 2) sub-Poisson excitation (e.g., via a stream of electrons subject to space charge) of an atom or other entity that emits a single photon in response to the excitation; 3) conversion from two-photon light whereby the arrival of one photon of a pair serves to herald the presence of its simultaneously arriving companion, as described in Sec. 13.3D; and 4) emission from a quantum dot, defect, or similar entity embedded in a photonic structure, as considered in Sec. 17.2E.

D. Two-Photon Light

Two-photon light, yet another form of nonclassical light, is described by a joint wavefunction that represents its joint probability amplitude. For example, its spatial properties are described by the wavefunction $U(x_1, x_2)$, where $|U(x_1, x_2)|^2$ is the joint probability density of finding the photons at positions x_1 and x_2 in the transverse plane. The corresponding directional (or spatial-frequency) properties are described by the two-dimensional Fourier transform of $U(x_1, x_2)$, which is denoted $V(\nu_{x1}, \nu_{x2})$. The joint probability density of finding photons with wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ (momenta

$\hbar\mathbf{k}_1$ and $\hbar\mathbf{k}_2$) is given by $|V(\nu_{x1}, \nu_{x2})|^2$, where $k_{x1} = 2\pi\nu_{x1}$ and $k_{x2} = 2\pi\nu_{x2}$ are the transverse components of the wavevectors. The temporal and spectral properties are similarly described by the joint wavefunctions $U(t_1, t_2)$ and its two-dimensional Fourier transform $V(\nu_1, \nu_2)$.

While the polarization state of classical or single-photon light is described by the Jones vector $\mathbf{J} = \begin{bmatrix} A_x \\ A_y \end{bmatrix}$, that of two-photon light is described by a 2×2 joint matrix $\mathbf{J} = \begin{bmatrix} A_{xx} & A_{yx} \\ A_{xy} & A_{yy} \end{bmatrix}$, where the first and second subscripts denote the polarizations of the first and second photons, respectively. For example, $|A_{yx}|^2$ represents the probability of detecting the first and second photons in the y (vertical or V) and x (horizontal or H) polarizations, respectively.

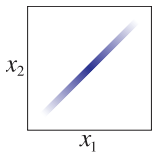
Entangled Photons

If the two photons are independent, their joint wavefunction factors into a product of the individual wavefunctions. In that case

$$U(x_1, x_2) = U_1(x_1)U_2(x_2), \tag{13.3-14}$$

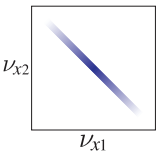
so that the probability of detecting the two photons also factors, indicating that the positions of the two photons are statistically independent and uncorrelated.

If the joint wavefunction is not factorable into a product, however, the photons are said to be **entangled**. An extreme case is the wavefunction



$$U(x_1, x_2) = U_s(x_1)\delta(x_1 - x_2), \tag{13.3-15}$$

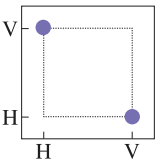
which describes two photons that are always detected at the same position ($x_1 = x_2$), although that position is random with probability density $|U_s(x_1)|^2$. In that case, the photons are said to be **maximally entangled**. The Fourier transform of (13.3-15) is



$$V(\nu_{x1}, \nu_{x2}) = V_s(\nu_{x1})\delta(\nu_{x1} + \nu_{x2}), \tag{13.3-16}$$

where $V_s(\nu_x)$ is the one-dimensional Fourier transform of $U_s(x)$. The transverse components of the wavevector are then anticorrelated ($k_{x1} = -k_{x2}$), indicating that the photon directions are also anticorrelated.

An example of two photons that are maximally entangled in polarization is provided by the polarization state



$$\mathbf{J} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{13.3-17}$$

There are then two polarization possibilities and they have equal probabilities: photon 1 is y polarized and photon 2 is x polarized, or *vice-versa*. Each photon can be either vertically or horizontally polarized, but if one is vertically polarized, the other *must* be horizontally polarized. In the quantum-optics literature, the photon pair is said to be in

a VH+HV superposition state, where V and H represent the vertical (y) and horizontal (x) polarizations, respectively. An entangled photon pair is also called a **biphoton** and two-photon light is sometimes called **twin-beam light**.

Generation and applications of two-photon light. Entangled two-photon light may be generated by means of **spontaneous parametric downconversion (SPDC)**, a nonlinear optical process whereby some fraction of a beam of photons incident on a nonlinear optical crystal are split into pairs of photons, while conserving energy and momentum (Sec. 22.2C). Given that the energy of a photon is $E = \hbar\omega$, conservation of energy dictates that the frequency of the incident photon (called the *pump*) must equal the sum of the energies of the two downconverted photons (called the *signal* and *idler*), i.e., $\omega_p = \omega_1 + \omega_2$ or $(\omega_1 - \frac{1}{2}\omega_p) = -(\omega_2 - \frac{1}{2}\omega_p)$. If the pump is monochromatic, ω_p is fixed and ω_1 and ω_2 are anticorrelated variables. By virtue of the Fourier-transform relation between the spectral and temporal wavefunctions, this implies that the emission times are fully correlated, i.e., $t_1 = t_2$.

Similarly, given that the momentum of a photon is $\mathbf{p} = \hbar\mathbf{k}$, conservation of momentum dictates that the pump wavevector must equal the sum of the wavevectors of the two downconverted photons, i.e., $\mathbf{k}_p = \mathbf{k}_1 + \mathbf{k}_2$. If the pump is a plane-wave traveling along the axial (z) direction, the transverse components k_{1x} and k_{2x} of the wavevectors of the generated photons must then sum to zero. Thus, $0 = k_{1x} + k_{2x}$ or $k_{2x} = -k_{1x}$, revealing that these components are anticorrelated; the positions from which the two photons are emitted are thus correlated, i.e., $x_1 = x_2$. This two-photon wavefunction thus assumes the form specified in (13.3-15), where $U_s(x)$ is determined by the dimensions and optical properties of the nonlinear crystal in which the entangled photons are generated. Downconverted photons can also exhibit polarization entanglement if the nonlinear process requires that they have orthogonal polarizations. SPDC is useful for generating entangled two-photon light in crystal-based bulk optics (Example 22.2-3) as well as in monolithic semiconductor chips. Hollow-core photonic-crystal fibers (Sec. 10.4) filled with noble gases such as argon can also be configured to generate two-photon light via a modulation instability.

Two-photon light can be used to generate single-photon light by using the arrival of one member of the photon pair to herald the presence of the other. Entangled photons find use in applications such as secure quantum communications, cryptography, sensing, and imaging. Entanglement can be distributed over long distances — satellite-based entanglement distribution has been achieved for locations on earth separated by more than 1200 km.

Two-Photon Optics

The transmission of two-photon light through a linear optical system obeys equations based on classical optics, as we proceed to describe.

Polarization optics. When a classical plane wave whose polarization described by the Jones vector $\mathbf{J}_i$ is transmitted through a polarization element with Jones matrix $\mathbf{T}$, the Jones vector of the outgoing wave is $\mathbf{J}_o = \mathbf{T}\mathbf{J}_i$ (Sec. 6.1B). This relation, which is also applicable for single-photon light, is readily generalized to two-photon light by applying the matrix $\mathbf{T}$ twice, once for each photon, which gives rise to

$$\mathbf{J}_o = \mathbf{T}\mathbf{J}_i\mathbf{T}^{\mathsf{T}}, \tag{13.3-18}$$

where the superscript ^{T} signifies the matrix transpose.

EXAMPLE 13.3-1. ***Polarization Rotator.*** The Jones matrix of a 45° polarization rotator is $\mathbf{T} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ [see (6.1-20)]. Upon transmission through such a device, polarization-entangled photons described by the VH+HV Jones vector $\mathbf{J}_i = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ [see (13.3-17)] are characterized by $\mathbf{J}_o = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$. This Jones vector represents entangled photons in the VV−HH state, for which the outgoing photons must be of the same polarization; the probability that they are of orthogonal polarization is zero. It can be demonstrated that maximally entangled light remains maximally entangled when subjected to polarization rotation for an arbitrary angle.

Spatial optics. When a classical wave with wavefunction $U_i(x, y)$ is transmitted through an optical system with impulse response function $h(x, y; x', y')$, the outgoing wave $U_o(x, y)$ is described by the integral (13.1-9), which also applies for single-photon light. This result may be generalized to two-photon light by applying the impulse response function twice, once for each photon. Ignoring the y dependence for simplicity, we then have

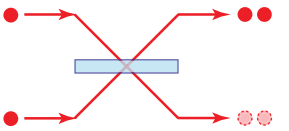
$$U_o(x_1, x_2) = \int\int_{-\infty}^{\infty} h(x_1; x_1')\, h(x_2; x_2')\, U_i(x_1', x_2')\, dx_1'\, dx_2', \tag{13.3-19}$$

where the subscripts 1 and 2 denote photons 1 and 2, respectively. Note that (13.3-19) is analogous to (12.3-5) in Sec. 12.3A, which describes the propagation of partially coherent light with mutual intensity $G(x_1, x_2)$ through a linear optical system. But note also that the conjugation in (12.3-5) is absent in (13.3-19).

EXAMPLE 13.3-2. ***Fourier-Transform System.*** For a 2-f optical system that implements a Fourier-transform operation (Sec. 4.2B), i.e., $h(x; x') \propto \exp(-jxx'/\lambda f)$, where λ is the wavelength and f is the focal length of the lens, we have $U_o(x_1, x_2) \propto V_i(x_1/\lambda f, x_2/\lambda f)$, where $V_i(\nu_{x1}, \nu_{x2})$ is the two-dimensional Fourier transform of $U_i(x_1, x_2)$. The 2-f system manifests the directional, or momentum, characteristics of two-photon light. Correlated photons are converted by the Fourier-transforming lens into anticorrelated photons.

Two-beam optics. Two-photon light in the form of two beams, labeled a and b, is described by a 2×2 joint matrix $\mathbf{J}_i = \frac{1}{\sqrt{2}}\begin{bmatrix} A_{\mathrm{aa}} & A_{\mathrm{ab}} \\ A_{\mathrm{ba}} & A_{\mathrm{bb}} \end{bmatrix}$, where, for example, $|A_{\mathrm{ba}}|^2$ is the probability that the first and second photons are found in beams b and a, respectively. If the two beams are mixed at a lossless device characterized by a scattering matrix $\mathbf{S}$ (see Sec. 7.1A), the outgoing pair of beams of two-photon light is described by the matrix $\mathbf{J}_o = \mathbf{S}\mathbf{J}_i\mathbf{S}^{\mathsf{T}}$, a relation that assumes the same form as (13.3-18).

EXAMPLE 13.3-3. ***Hong–Ou–Mandel (HOM) Interferometer.*** A lossless symmetric beamsplitter is characterized by the scattering matrix $\mathbf{S} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & j \\ j & 1 \end{bmatrix}$, as provided in (7.1-18). For incoming light described by $\mathbf{J}_i = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, indicating that there is a single photon in each beam, we obtain $\mathbf{J}_o = \mathbf{S}\mathbf{J}_i\mathbf{S}^{\mathsf{T}} = \frac{j}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. This reveals that the outgoing photons always emerge together as a pair, out of one or the other port, and there is zero probability that each photon emerges from a different port.[†] Hence, the two



[†] See C. K. Hong, Z. Y. Ou, and L. Mandel, Measurement of Subpicosecond Time Intervals Between Two Photons by Interference, *Physical Review Letters*, vol. 59, pp. 2044–2046, 1987.

indistinguishable photons that were initially separate, each entering a different input port of the beamsplitter, exit the beamsplitter "stuck together" from one of the (randomly chosen) output ports. This outcome may be understood by observing that there are two paths the input photons could follow at the beamsplitter for them to emerge from *different* output ports — both could undergo reflection or both could undergo transmission. However, the probability amplitudes for these two possibilities are equal since the photons are indistinguishable, and they cancel because of the phase shift associated with reflection at the beamsplitter. Hence, the only remaining possibilities are that the photons emerge together from the same output port. The beamsplitter thus serves to convert two indistinguishable photons at its input ports into two entangled photons at its output ports, since the two photons always emerge from the same output port, although which port is random. This form of **two-photon interference**, known as **Hong–Ou–Mandel interference**, is widely used in quantum optics, quantum information, and quantum imaging.

Applications of two-photon optics. Optical systems employing two-photon light find application in a number of areas. An example in the domain of imaging is provided by quantum optical coherence tomography[†] (QOCT), a two-photon interferometric technique that allows a multilayered medium to be axially sectioned. This imaging modality makes use of two indistinguishable photons; a Hong–Ou–Mandel interferometer in which one of the photons is reflected from a movable mirror while the other is reflected from the sample before being presented to the two input ports of the beamsplitter; and a pair of photodetectors connected to register the photon-coincidence rate at the output ports of the beamsplitter. The depths of the reflective layers in the sample are revealed by determining the path delays of the movable mirror that lead to dips in the coincidence rate, thereby indicating the presence of HOM interference. QOCT is the two-photon analog of optical coherence tomography (OCT), discussed in Sec. 12.2B, which makes use of partially coherent light of short coherence length; a Michelson interferometer in which one the mirrors is replaced by the sample; and a photodetector responsive to intensity (Fig. 12.2-3). For OCT, the depths of the reflective layers in the sample are revealed by determining the path delays of the movable mirror that lead to interference fringes in the intensity. Though QOCT is more complex to implement than OCT, it has the merit that it is immune to even-order group velocity dispersion (GVD) in the sample, which serves to increase the attainable resolution and sectioning depth; the technique simultaneously permits the GVD coefficients of the media that comprise the sample to be determined.

The implementation of two-photon optical systems has been substantially advanced by the development of **integrated quantum photonics**, a platform in which **on-chip quantum circuits** generate and manipulate quantum states of light such as single photons, entangled photons, and complex states in which entanglement is shared among multiple modes.

READING LIST

Quantum Optics

Z.-Y. J. Ou, *Quantum Optics for Experimentalists*, World Scientific, 2017.

M. Orszag, *Quantum Optics: Including Noise Reduction, Trapped Ions, Quantum Trajectories, and Decoherence*, Springer-Verlag, 3rd ed. 2016.

C. W. Gardiner and P. Zoller, *The Quantum World of Ultra-Cold Atoms and Light. Book I: Foundations of Quantum Optics*, Imperial College Press, 2014.

[†] See M. B. Nasr, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich, Dispersion-Cancelled and Dispersion-Sensitive Quantum Optical Coherence Tomography, *Optics Express*, vol. 12, pp. 1353–1362, 2004.

G. S. Agarwal, *Quantum Optics*, Cambridge University Press, 2012.

G. Grynberg, A. Aspect, and C. Fabre, *Introduction to Quantum Optics: From the Semi-Classical Approach to Quantized Light*, Cambridge University Press, 2010.

U. Leonhardt, *Essential Quantum Optics: From Quantum Measurements to Black Holes*, Cambridge University Press, 2010.

D. F. Walls and G. J. Milburn, *Quantum Optics*, Springer-Verlag, 2nd ed. 2008.

J. C. Garrison and R. Y. Chiao, *Quantum Optics*, Oxford University Press, 2008.

P. Meystre and M. Sargent III, *Elements of Quantum Optics*, Springer-Verlag, 4th ed. 2007.

M. Fox, *Quantum Optics: An Introduction*, Oxford University Press, 2006.

W. Vogel and D.-G. Welsch, *Quantum Optics*, Wiley–VCH, 3rd ed. 2006.

C. C. Gerry and P. L. Knight, *Introductory Quantum Optics*, Cambridge University Press, 2005.

H.-A. Bachor and T. C. Ralph, *A Guide to Experiments in Quantum Optics*, Wiley–VCH, 2nd ed. 2004.

W. P. Schleich, *Quantum Optics in Phase Space*, Wiley–VCH, 2001.

R. Loudon, *The Quantum Theory of Light*, Oxford University Press, 3rd ed. 2000.

V. Peřinová, A. Lukš, and J. Peřina, *Phase in Optics*, World Scientific, 1998.

M. O. Scully and M. S. Zubairy, *Quantum Optics*, Cambridge University Press, paperback ed. 1997.

G. S. Agarwal, ed., *Selected Papers on Fundamentals of Quantum Optics*, SPIE Optical Engineering Press (Milestone Series Volume 103), 1995.

R. P. Feynman (with an introduction by A. Zee), *QED: The Strange Theory of Light and Matter*, Princeton University Press, 1985, reissued 2014.

S. Weinberg, Light as a Fundamental Particle, *Physics Today*, vol. 28, no. 6, pp. 32–37, 1975.

J. R. Klauder and E. C. G. Sudarshan, *Fundamentals of Quantum Optics*, Benjamin, 1968; Dover, reissued 2006.

Coherence, Quantum Fluctuations, Antibunching, and Photon Statistics

B. R. Masters, Satyendra Nath Bose and Bose–Einstein Statistics, *Optics & Photonics News*, vol. 24, no. 4, pp. 40–47, 2013.

H. J. Carmichael, *Statistical Methods in Quantum Optics 2: Non-Classical Fields*, Springer-Verlag, 2010.

R. J. Glauber, *Quantum Theory of Optical Coherence: Selected Papers and Lectures*, Wiley–VCH, 2007.

R. J. Glauber, Nobel Lecture: One Hundred Years of Light Quanta, *Reviews of Modern Physics*, vol. 78, pp. 1267–1278, 2006.

Z. Ficek and S. Swain, *Quantum Interference and Coherence: Theory and Experiments*, Springer-Verlag, 2005.

J. Peřina, ed., *Coherence and Statistics of Photons and Atoms*, Wiley, 2001.

H. J. Carmichael, *Statistical Methods in Quantum Optics 1: Master Equations and Fokker–Planck Equations*, Springer-Verlag, 1999.

L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge University Press, 1995.

L. Mandel and E. Wolf, eds., *Selected Papers on Coherence and Fluctuations of Light (1850–1966)*, SPIE Optical Engineering Press (Milestone Series Volume 19), 1990.

R. A. Campos, B. E. A. Saleh, and M. C. Teich, Quantum-Mechanical Lossless Beam Splitter: SU(2) Symmetry and Photon Statistics, *Physical Review A*, vol. 40, pp. 1371–1384, 1989.

E. R. Pike and H. Walther, eds., *Photons and Quantum Fluctuations*, CRC Press, 1988.

M. C. Teich and B. E. A. Saleh, Photon Bunching and Antibunching, in E. Wolf, ed., *Progress in Optics*, North-Holland, 1988, vol. 26, pp. 1–104.

J. Peřina, *Coherence of Light*, Reidel, 2nd ed. 1985.

E. Wolf, Einstein's Researches on the Nature of Light, *Optics News*, vol. 5, no. 1, pp. 24–39, 1979.

B. E. A. Saleh, *Photoelectron Statistics*, Springer-Verlag, 1978.

H. J. Kimble, M. Dagenais, and L. Mandel, Photon Antibunching in Resonance Fluorescence, *Physical Review Letters*, vol. 39, pp. 691–695, 1977.

W. H. Louisell, *Quantum Statistical Properties of Radiation*, Wiley, 1973; Wiley–VCH, reprinted 1990.

L. Mandel and E. Wolf, eds., *Selected Papers on Coherence and Fluctuations of Light*, Volumes 1 and 2, Dover, 1970.

L. Mandel, Fluctuations of Light Beams, in E. Wolf, ed., *Progress in Optics*, North-Holland, 1963, vol. 2, pp. 181–248.

S. N. Bose, Plancks Gesetz und Lichtquantenhypothese (Planck's Law and the Light-Quantum Hypothesis), *Zeitschrift für Physik*, vol. 26, pp. 178-181, 1924.

Quadrature-Squeezed and Photon-Number-Squeezed Light

X. He, N. F. Hartmann, X. Ma, Y. Kim, R. Ihly, J. L. Blackburn, W. Gao, J. Kono, Y. Yomogida, A. Hirano, T. Tanaka, H. Kataura, H. Htoon, and S. K. Doorn, Tunable Room-Temperature Single-Photon Emission at Telecom Wavelengths from sp^3 Defects in Carbon Nanotubes, *Nature Photonics*, vol. 11, pp. 577–582, 2017.

H. Vahlbruch, M. Mehmet, K. Danzmann, and R. Schnabel, Detection of 15 dB Squeezed States of Light and their Application for the Absolute Calibration of Photoelectric Quantum Efficiency, *Physical Review Letters*, vol. 117, 110801, 2016.

J. Aasi *et al.* (LIGO Scientific Collaboration), Enhanced Sensitivity of the LIGO Gravitational Wave Detector by Using Squeezed States of Light, *Nature Photonics*, vol. 7, pp. 613–619, 2013.

H. J. Carmichael, *Statistical Methods in Quantum Optics 2: Non-Classical Fields*, Springer-Verlag, 2010.

M. C. Teich and B. E. A. Saleh, Squeezed and Antibunched Light, *Physics Today*, vol. 43, no. 6, pp. 26–34, 1990 (Erratum: vol. 43, no. 11, pp. 123–124, 1990).

M. C. Teich and B. E. A. Saleh, Squeezed States of Light, *Quantum Optics: Journal of the European Physical Society B*, vol. 1, pp. 153–191, 1989.

R. E. Slusher, L. W. Hollberg, B. Yurke, J. C. Mertz, and J. F. Valley, Observation of Squeezed States Generated by Four-Wave Mixing in an Optical Cavity, *Physical Review Letters*, vol. 55, pp. 2409–2412, 1985.

M. C. Teich and B. E. A. Saleh, Observation of Sub-Poisson Franck–Hertz Light at 253.7 nm, *Journal of the Optical Society of America B*, vol. 2, pp. 275–282, 1985.

D. Stoler, B. E. A. Saleh, and M. C. Teich, Binomial States of the Quantized Radiation Field, *Optica Acta* (*Journal of Modern Optics*), vol. 32, pp. 345–355, 1985.

R. Short and L. Mandel, Observation of Sub-Poissonian Photon Statistics, *Physical Review Letters*, vol. 51, pp. 384–387, 1983.

C. M. Caves, Quantum-Mechanical Noise in an Interferometer, *Physical Review D*, vol. 23, pp. 1693–1708, 1981.

Entangled-Photon and Two-Photon Light

J. Wang, S. Paesani, Y. Ding, R. Santagati, P. Skrzypczyk, A. Salavrakos, J. Tura, R. Augusiak, L. Mančinska, D. Bacco, D. Bonneau, J. W. Silverstone, Q. Gong, A. Acín, K. Rottwitt, L. K. Oxenløwe, J. L. O'Brien, A. Laing, and M. G. Thompson, Multidimensional Quantum Entanglement With Large-Scale Integrated Optics, *Science*, vol. 360, pp. 285–291, 2018.

M. A. Finger, T. Sh. Iskhakov, N. Y. Joly, M. V. Chekhova, and P. St. J. Russell, Raman-Free, Noble-Gas-Filled Photonic-Crystal Fiber Source for Ultrafast, Very Bright Twin-Beam Squeezed Vacuum, *Physical Review Letters*, vol. 115, 143602, 2015.

D. Kang, M. Kim, H. He, and A. S. Helmy, Two Polarization-Entangled Sources from the Same Semiconductor Chip, *Physical Review A*, vol. 92, 013821, 2015.

F. Boitier, A. Orieux, C. Autebert, A. Lemaître, E. Galopin, C. Manquest, C. Sirtori, I. Favero, G. Leo, and S. Ducci, Electrically Injected Photon-Pair Source at Room Temperature, *Physical Review Letters*, vol. 112, 183901, 2014.

M. F. Saleh, G. Di Giuseppe, B. E. A. Saleh, and M. C. Teich, Photonic Circuits for Generating Modal, Spectral, and Polarization Entanglement, *IEEE Photonics Journal*, vol. 2, pp. 736–752, 2010.

M. B. Nasr, S. Carrasco, B. E. A. Saleh, A. V. Sergienko, M. C. Teich, J. P. Torres, L. Torner, D. S. Hum, and M. M. Fejer, Ultrabroadband Biphotons Generated via Chirped Quasi-Phase-Matched Optical Parametric Down-Conversion, *Physical Review Letters*, vol. 100, 183601, 2008.

J. Audretsch, *Entangled Systems*, Wiley–VCH, 2007.

J. S. Bell (with an introduction by A. Aspect), *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press, 2nd ed. 2004.

B. E. A. Saleh, A. F. Abouraddy, A. V. Sergienko, and M. C. Teich, Duality Between Partial Coherence and Partial Entanglement, *Physical Review A*, vol. 62, 043816, 2000.

D. N. Klyshko, *Photons and Nonlinear Optics*, Nauka (Moscow), 1980 [Translation: Gordon and Breach, New York, 1988).

A. Aspect, P. Grangier, and G. Roger, Experimental Tests of Realistic Local Theories via Bell's Theorem, *Physical Review Letters*, vol. 47, pp. 460-463, 1981.

D. C. Burnham and D. L. Weinberg, Observation of Simultaneity in Parametric Production of Optical Photon Pairs, *Physical Review Letters*, vol. 25, pp. 84–87, 1970.

D. Magde and H. Mahr, Study in Ammonium Dihydrogen Phosphate of Spontaneous Parametric Interaction Tunable from 4400 to 16 000 Å, *Physical Review Letters*, vol. 18, pp. 905–907, 1967.

S. E. Harris, M. K. Oshman, and R. L. Byer, Observation of Tunable Optical Parametric Fluorescence, *Physical Review Letters*, vol. 18, pp. 732–734, 1967.

A. Einstein, B. Podolsky, and N. Rosen, Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?, *Physical Review*, vol. 47, pp. 777–780, 1935.

E. Schrödinger, Die gegenwärtige Situation in der Quantenmechanik, *Die Naturwissenschaften*, vol. 23, pp. 807–812, 823–828, 844–849; 1935 [Translation: The Present Situation in Quantum Mechanics: A Translation of Schrödinger's 'Cat Paradox' Paper, *Proceedings of the American Philosophical Society*, vol. 124, pp. 323–338, 1980].

Quantum Imaging, Metrology, Lithography, and Spectroscopy

D. S. Simon, G. Jaeger, A. V. Sergienko, *Quantum Metrology, Imaging, and Communication*, Springer-Verlag, 2017.

M. A. Taylor, J. Janousek, V. Daria, J. Knittel, B. Hage, H.-A. Bachor, and W. P. Bowen, Subdiffraction-Limited Quantum Imaging within a Living Cell, *Physical Review X*, vol. 4, 011017, 2014.

T. Ono, R. Okamoto, and S. Takeuchi, An Entanglement-Enhanced Microscope, *Nature Communications* **4**, 2426 doi: 10.1038/ncomms3426, 2013.

M. C. Teich, B. E. A. Saleh, F. N. C. Wong, and J. H. Shapiro, Variations on the Theme of Quantum Optical Coherence Tomography: A Review, *Quantum Information Processing*, vol. 11, pp. 903–923, 2012.

M. B. Nasr, D. P. Goode, N. Nguyen, G. Rong, L. Yang, B. M. Reinhard, B. E. A. Saleh, and M. C. Teich, Quantum Optical Coherence Tomography of a Biological Sample, *Optics Communications*, vol. 282, pp. 1154–1159, 2009.

M. B. Nasr, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich, Demonstration of Dispersion-Cancelled Quantum-Optical Coherence Tomography, *Physical Review Letters*, vol. 91, 083601, 2003.

A. F. Abouraddy, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich, Entangled-Photon Fourier Optics, *Journal of the Optical Society of America B*, vol. 19, pp. 1174–1184, 2002.

A. F. Abouraddy, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich, Role of Entanglement in Two-Photon Imaging, *Physical Review Letters*, vol. 87, 123602, 2001.

A. F. Abouraddy, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich, Quantum Holography, *Optics Express*, vol. 9, pp. 498–505, 2001.

A. N. Boto, P. Kok, D. S. Abrams, S. L. Braunstein, C. P. Williams, and J. P. Dowling, Quantum Interferometric Optical Lithography: Exploiting Entanglement to Beat the Diffraction Limit, *Physical Review Letters*, vol. 85, pp. 2733-2736, 2000.

B. E. A. Saleh, B. M. Jost, H.-B. Fei, and M. C. Teich, Entangled-Photon Virtual-State Spectroscopy, *Physical Review Letters*, vol. 80, pp. 3483–3486, 1998.

M. C. Teich and B. E. A. Saleh, Entangled-Photon Microscopy, Spectroscopy, and Display, *U.S. Patent 5,796,477*, Patented August 18, 1998.

Quantum Information, Communications, Computing, and Cryptography

I. Khan, B. Heim, A. Neuzner, and C. Marquardt, Satellite-Based QKD, *Optics & Photonics News*, vol. 29, no. 2, pp. 26–33, 2018.

M. M. Wilde, *Quantum Information Theory*, Cambridge University Press, 2nd ed. 2017.

J. Yin *et al.*, Satellite-Based Entanglement Distribution Over 1200 Kilometers, *Science*, vol. 356, pp. 1140–1144, 2017.

M. Hayashi, S. Ishizaka, A. Kawachi, G. Kimura, and T. Ogawa, *Introduction to Quantum Information Science*, Springer-Verlag, 2015.

J. Carolan, C. Harrold, C. Sparrow, E. Martín-López, N. J. Russell, J. W. Silverstone, P. J. Shadbolt, N. Matsuda, M. Oguma, M. Itoh, G. D. Marshall, M. G. Thompson, J. C. F. Matthews, T. Hashimoto, J. L. O'Brien, and A. Laing, Universal Linear Optics, *Science*, vol. 349, pp. 711–716, 2015.

I. A. Walmsley, Quantum Optics: Science and Technology in a New Light, *Science*, vol. 348, pp. 525–530, 2015.

J. A. Bergou and M. Hillery, *Introduction to the Theory of Quantum Information Processing*, Springer-Verlag, 2013.

J. A. Jones and D. Jaksch, *Quantum Information, Computation and Communication*, Cambridge University Press, 2012.

R. T. Perry, *Quantum Computing from the Ground Up*, World Scientific, 2012.

A. Zeilinger, *Dance of the Photons: From Einstein to Quantum Teleportation*, Farrar, Straus and Giroux, 2010.

S. M. Barnett, *Quantum Information*, Oxford University Press, 2009.

G. Jaeger, *Entanglement, Information, and the Interpretation of Quantum Mechanics*, Springer-Verlag, 2009.

J. L. O'Brien, A. Furusawa, and J. Vučković, Photonic Quantum Technologies, *Nature Photonics*, vol. 3, pp. 687–695, 2009.

H. Paul, *Introduction to Quantum Optics: From Light Quanta to Quantum Teleportation*, Cambridge University Press, 2004.

M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, 2002.

PROBLEMS

13.1-5 **Photon Energy.**

(a) What voltage should be applied to an electrode to accelerate an electron from zero velocity such that it acquires the same energy as a photon of wavelength $\lambda_o = 0.87\ \mu\text{m}$?

(b) A photon of wavelength $1.06\ \mu\text{m}$ is combined with a photon of wavelength $10.6\ \mu\text{m}$ to create a photon whose energy is the sum of the energies of the two photons. What is the wavelength of the resultant photon? This process, known as sum-frequency generation, is illustrated in Fig. 22.2-6.

13.1-6 **Position of a Single Photon at a Screen.** Consider monochromatic light of wavelength λ_o incident on an infinite screen in the plane $z = 0$, with an intensity $I(\rho) = I_0 \exp(-\rho/\rho_0)$, where $\rho = \sqrt{x^2 + y^2}$. Assume that the intensity of the source is reduced to a level at which only a single photon strikes the screen.

(a) Find the probability that the photon strikes the screen within a radius ρ_0 of the origin.

(b) If the incident light contains exactly 10^6 photons, how many photons strike within a circle of radius ρ_0 on average?

13.1-7 **Photon Momentum.** Compare the magnitude of the total momentum in all of the photons in a 10-J laser pulse with that of:

(a) a 1-g mass moving at a velocity of 1 cm/s.

(b) an electron moving at a velocity $c_o/10$.

*13.1-8 **Momentum of a Photon in a Gaussian Beam.**

(a) What is the probability that the momentum vector of a photon associated with a Gaussian beam of waist radius W_0 lies within the beam divergence angle θ_0? Refer to Sec. 3.1 for definitions.

(b) Does the relation $p = E/c_o$ hold in this case? Explain.

13.1-9 **Levitation by Light Pressure.** An isolated hydrogen atom has mass 1.67×10^{-27} kg.

(a) Find the gravitational force on this hydrogen atom near the surface of the earth (assume that at sea level the gravitational acceleration constant $g = 9.8\ \mathrm{m/s^2}$).

(b) Let an upwardly directed laser beam emitting 1-eV photons be focused in such a way that the full momentum of each of its photons is transferred to the atom. Find the average upward force on the atom provided by one photon striking each second.

(c) Find the number of photons that must strike the atom per second, and the corresponding optical power, for it not to fall under the effect of gravity, assuming ideal conditions in vacuum.

(d) How many photons per second would be required to keep the atom from falling if it were perfectly reflecting?

*13.1-10 **Single Photon in a Fabry–Perot Resonator.** Consider a Fabry–Perot resonator of length $d = 1$ cm that contains nonabsorbing material of refractive index $n = 1.5$ and perfectly reflecting mirrors. Assume that there is exactly one photon in the mode described by the standing wave $\sin(10^5 \pi x/d)$.

(a) Determine the photon wavelength and energy (in eV).

(b) Estimate the uncertainty in the photon's position and momentum (magnitude and direction). Compare with the value obtained from the relation $\sigma_x \sigma_p = \hbar/2$.

13.1-11 **Single-Photon Beating (Time Interference).** Consider a detector illuminated by a polychromatic plane wave consisting of two superposed monochromatic waves traveling in the same direction. The constituent waves have complex wavefunctions given by

$$U_1(t) = \sqrt{I_1}\exp(j2\pi\nu_1 t) \qquad \text{and} \qquad U_2(t) = \sqrt{I_2}\exp(j2\pi\nu_2 t),$$

with frequencies ν_1 and ν_2 and intensities I_1 and I_2, respectively. In accordance with Sec. 2.6B, the intensity of this wave is given by $I(t) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos[2\pi(\nu_2 - \nu_1)t]$. Assume that the two constituent plane waves have equal intensities ($I_1 = I_2$) and that the wave is sufficiently weak so that only a single polychromatic photon reaches the detector during the time interval $T = 1/|\nu_2 - \nu_1|$.

(a) Sketch the probability density $p(t)$ for detecting the photon in the interval $0 \le t \le 1/|\nu_2 - \nu_1|$. At what time instant during T is the probability density zero that the photon will be detected?

(b) An attempt to discover from which of the two constituent waves the photon arrives entails an energy measurement with a precision better than

$$\sigma_E < h|\nu_2 - \nu_1|.$$

Use the time–energy uncertainty relation to show that the time required for such a measurement is of the order of the beat-frequency period. This signifies that the very process of measurement washes out the interference and thereby precludes the measurement from being made.

13.1-12 **Photon Momentum Exchange at a Beamsplitter.** Consider a single photon, in a mode described by a plane wave, impinging on a lossless beamsplitter. What is the momentum vector of the photon before it impinges on the mirror? What are the possible values of the momentum vector of the photon, and the probabilities of observing these values, after passage through the beamsplitter?

13.2-2 **Photon Flux.** Demonstrate that the power of a monochromatic optical beam that carries an average of one photon per optical cycle is inversely proportional to the square of the wavelength.

13.2-3 **The Poisson Distribution.** Verify that the Poisson probability distribution provided in (13.2-12) is normalized to unity and has mean $\bar{n}$ and variance $\sigma_n^2 = \bar{n}$.

13.2-4 **Photon Statistics of a Coherent Gaussian Beam.** Assume that a 100-pW single-mode He–Ne laser emits light at 633 nm in a TEM_{00} Gaussian beam (Sec. 3.1).

(a) What is the mean number of photons crossing a circle of radius equal to the waist radius of the beam W_0 in a time $T = 100$ ns?

(b) What is the root-mean-square value of the number of photons in (a)?

(c) What is the probability that the number of photons is zero in (a)?

13.2-5 **The Bose–Einstein Distribution.**

(a) Verify that the Bose–Einstein probability distribution provided in (13.2-20) is normalized and has mean $\overline{n}$ and variance $\sigma_n^2 = \overline{n} + \overline{n}^2$.

(b) If a stream of photons obeying Bose–Einstein statistics contains an average of $\Phi = 1$ photon per nanosecond, what is the probability that zero photons will be detected in a 20-ns time interval?

*13.2-6 **The Negative–Binomial Distribution.** It is well known in the literature of probability theory that the sum of $\mathcal{M}$ identically distributed random variables, each with a geometric (Bose–Einstein) distribution, obeys the negative binomial distribution with overall mean $\overline{n}$,

$$p(n) = \binom{n + \mathcal{M} - 1}{n} \frac{(\overline{n}/\mathcal{M})^n}{(1 + \overline{n}/\mathcal{M})^{n+\mathcal{M}}} .$$

Verify that the negative-binomial distribution reduces to the Bose–Einstein distribution for $\mathcal{M} = 1$ and to the Poisson distribution as $\mathcal{M} \to \infty$.

*13.2-7 **Photon-Number Statistics for Multimode Thermal Light in a Cavity.** Consider $\mathcal{M}$ modes of thermal radiation sufficiently close to each other in frequency that each can be considered to be occupied in accordance with a Bose–Einstein distribution of the same mean photon number $1/[\exp(h\nu/kT) - 1]$. Show that the variance of the *total* number of photons n is related to its mean by

$$\sigma_n^2 = \overline{n} + \frac{\overline{n}^2}{\mathcal{M}} ,$$

which indicates that multimode thermal light has less variance than does single-mode thermal light. The presence of the multiple modes provides averaging, thereby reducing the noisiness of the light.

*13.2-8 **Photon-Number Statistics for a Beam of Multimode Thermal Light.** A multimode thermal light source that carries $\mathcal{M}$ identical modes, each with an exponentially distributed (random) integrated rate, has an overall probability density $p(w)$ describable by the gamma density

$$p(w) = \frac{1}{(\mathcal{M} - 1)!} \left(\frac{\langle w \rangle}{\mathcal{M}} \right)^{-\mathcal{M}} w^{\mathcal{M}-1} \exp\left(-\frac{w}{\langle w \rangle / \mathcal{M}} \right) , \qquad w \geq 0.$$

Use Mandel's formula (13.2-26) to show that the resulting photon-number distribution assumes the form of the negative-binomial distribution defined in Prob. 13.2-6.

*13.2-9 **Mean and Variance of the Doubly Stochastic Poisson Distribution.** Prove (13.2-27) and (13.2-28).

*13.2-10 **Random Partitioning of Coherent Light.**

(a) Use (13.2-33) to show that the photon-number distribution of randomly partitioned coherent light retains its Poisson form.

(b) Show explicitly that the mean photon number for light reflected from a lossless beamsplitter is $(1 - \mathcal{T})\,\overline{n}$.

(c) Prove (13.2-34) for coherent light.

13.2-11 **Random Partitioning of Single-Mode Thermal Light.**

(a) Use (13.2-33) to show that the photon-number distribution of randomly partitioned single-mode thermal light retains its Bose–Einstein form.

(b) Show explicitly that the mean photon number for light reflected from a lossless beamsplitter is $(1 - \mathcal{T})\,\overline{n}$.

(c) Prove (13.2-35) for single-mode thermal light.

*13.2-12 **Exponential Decay of Mean Photon Number in an Absorber.**

(a) Consider an absorptive material of thickness d and absorption coefficient α (cm^{-1}). If the average number of photons entering the material is $\overline{n}_0$, write a differential equation to find the average number of photons $\overline{n}(x)$ at position x, where x is the depth into the material ($0 \leq x \leq d$).

(b) Solve the differential equation. State why your result turns out to be the exponential decay law obtained in electromagnetic optics (Sec. 5.5A).

(c) If the incident light is coherent, write an expression for the photon-number distribution $p(n)$ at an arbitrary position x into the absorber.

(d) What is the probability that a single photon incident on the absorber survives passage through it?

*13.3-1 **Statistics of the Binomial Photon-Number Distribution.** The binomial probability distribution, which is written as

$$p(n) = \frac{M!}{(M-n)!\,n!}\,p^n\,(1-p)^{M-n},$$

describes the photon-number statistics for certain sources of photon-number-squeezed light.[†]

(a) Indicate a plausible mechanism whereby number-state light is converted into light described by binomial photon statistics.

(b) Prove that the binomial probability distribution is normalized to unity.

(c) Find the photon-number mean $\overline{n}$ and the photon-number variance σ_n^2 of the binomial probability distribution in terms of its two parameters, p and M.

(d) Find an expression for the SNR in terms of $\overline{n}$ and p. Evaluate the SNR for the limiting cases $p \to 0$ and $p \to 1$. What is the nature of the light corresponding to those two limits?

*13.3-2 **Noisiness of the Uniform Photon-Number Distribution.** Consider a light source that generates a photon stream with the discrete-uniform photon-number distribution

$$p(n) = \begin{cases} \dfrac{1}{2\,\overline{n}+1}, & 0 \le n \le 2\,\overline{n} \\ 0, & \text{otherwise.} \end{cases}$$

(a) Verify that the distribution is normalized to unity and has mean $\overline{n}$. Calculate the photon-number variance σ_n^2 and the signal-to-noise ratio (SNR) and compare these quantities with those for the Bose–Einstein and Poisson distributions of the same mean.

(b) In terms of SNR, would this source be quieter or noisier than an ideal single-mode laser when $\overline{n} < 2$? When $\overline{n} = 2$? When $\overline{n} > 2$?

(c) By what factor is the SNR for this light larger than that for single-mode thermal light?

(d) Suggest a mechanism for generating light with this photon-number distribution.

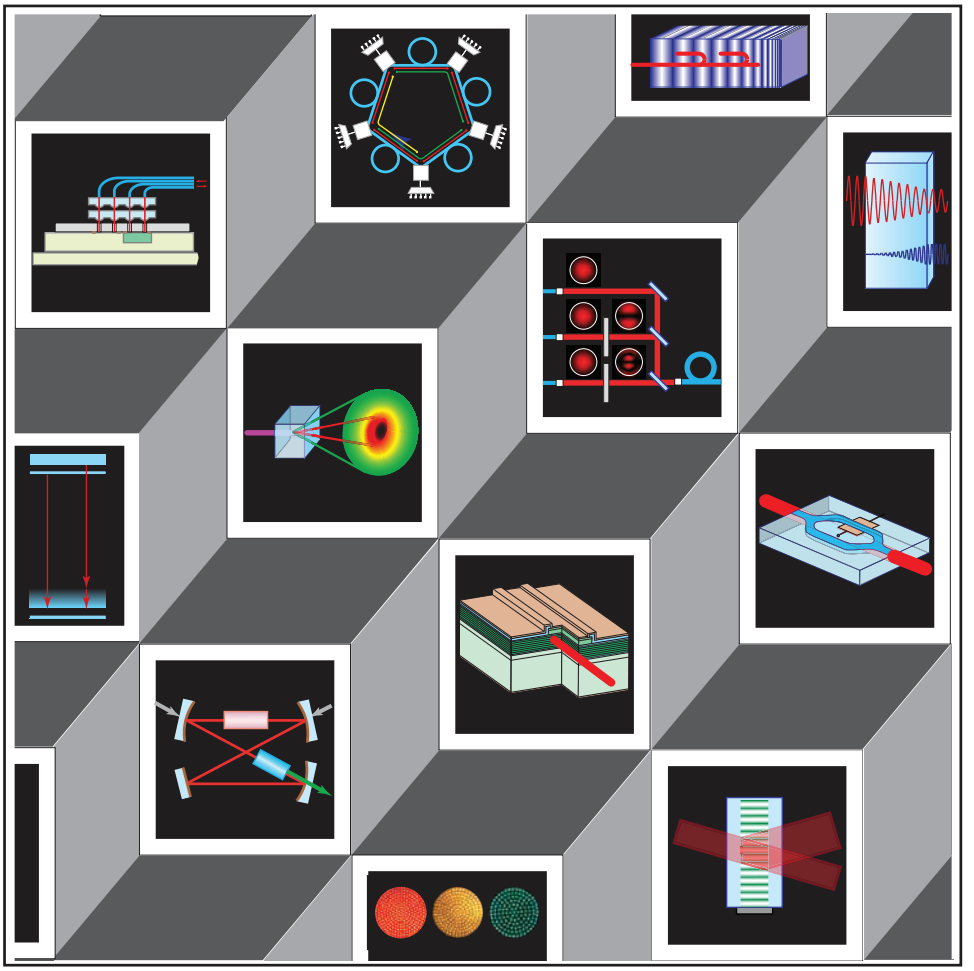
Useful formulas:

$$1 + 2 + 3 + \cdots + j = \frac{j(j+1)}{2}, \quad 1^2 + 2^2 + 3^2 + \cdots + j^2 = \frac{j(j+1)(2j+1)}{6}.$$

[†] See D. Stoler, B. E. A. Saleh, and M. C. Teich, Binomial States of the Quantized Radiation Field, *Optica Acta* (*Journal of Modern Optics*), vol. 32, pp. 345–355, 1985.

FUNDAMENTALS OF PHOTONICS

Part II: Photonics
(Chapters 14–25)



CHAPTER
14

LIGHT AND MATTER

14.1 ENERGY LEVELS 562
- A. Atoms
- B. Ions and Doped Dielectric Media
- C. Molecules
- D. Solids

14.2 OCCUPATION OF ENERGY LEVELS 581
- A. Boltzmann Distribution
- B. Fermi–Dirac Distribution

14.3 INTERACTIONS OF PHOTONS WITH ATOMS 583
- A. Interaction of Single-Mode Light with an Atom
- B. Spontaneous Emission
- C. Stimulated Emission and Absorption
- D. Line Broadening
- *E. Enhanced Spontaneous Emission
- *F. Laser Cooling, Laser Trapping, and Atom Optics

14.4 THERMAL LIGHT 602
- A. Thermal Equilibrium Between Photons and Atoms
- B. Blackbody Radiation Spectrum

14.5 LUMINESCENCE AND SCATTERING 607
- A. Forms of Luminescence
- B. Photoluminescence
- C. Scattering

Niels Bohr
(1885–1962)

Albert Einstein
(1879–1955)

Bohr and Einstein laid the theoretical foundations for describing the interaction of light with matter.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Light interacts with matter because matter contains electric charges. The time-varying electric field of light interacts with the electric charges and dipoles of atoms, molecules, and solids. A photon may interact with an atom (or ion) if its energy matches the difference between two atomic energy levels. The allowed energy levels and energy bands of matter are determined by the rules of quantum mechanics. If the atom is initially in the lower energy level, the photon may impart its energy to the atom and thereby raise it to the higher level; the photon is then said to be absorbed (or annihilated). Alternatively, if the atom is in the higher energy level, the photon may stimulate the atom to undergo a transition to the lower level, resulting in the emission (or creation) of a second photon whose energy is equal to the difference between the atomic energy levels. Under appropriate circumstances, stimulated emission can lead to laser action.

Thermal excitations cause the atoms of matter to constantly undergo upward and downward transitions among their allowed energy levels via the absorption and emission of photons. For blackbodies in thermal equilibrium, under steady-state conditions the resulting collection of photons and atoms produces thermal light. All blackbodies whose temperatures lie above absolute zero radiate thermal light, which has a distribution of frequencies known as the blackbody radiation spectrum. As the temperature of the object increases, the higher atomic energy levels become increasingly populated, causing the peak of the blackbody radiation spectrum to shift toward higher frequencies (shorter wavelengths).

Photon emission may also be instigated by external sources of energy other than thermal excitations. Exposure to sound waves, electric currents, ultraviolet radiation, and chemical reactions can cause materials to emit light called luminescence. Light can also interact with atoms via scattering and dispersive (gradient or dipole) forces. A photon incident on a material that has its direction and energy altered via scattering often serves to elucidate the internal energy levels of the material, such as those associated with molecular vibrations. Yet other processes can also result in the emission of light; examples include charged particles traveling faster than the velocity of light in a medium (Čerenkov radiation) and the deceleration of charged particles as they penetrate matter (Bremsstrahlung).

This Chapter

The purpose of this chapter is to introduce the laws responsible for the generation of laser, thermal, and luminescence light. The chapter begins with a brief review of the generic energy levels associated with different types of matter (Sec. 14.1) and the occupation of these energy levels (Sec. 14.2). In Sec. 14.3 we discuss the absorption and emission of photons by an atom; those results form the basis of the operation of laser amplifiers and oscillators, as set forth in Chapters 15 and 16, respectively. The interaction of many photons with many atoms, under conditions of thermal equilibrium and steady state, is considered in Sec. 14.4. Finally, an overview of luminescence and light scattering is provided in Sec. 14.5.

14.1 ENERGY LEVELS

The atoms of matter may exist in relative isolation, as in the case of a dilute atomic gas, or they may interact with neighboring atoms to form molecules, liquids, and solids. The constituents of matter obey the laws of quantum mechanics.

The behavior of a single nonrelativistic particle of mass m (an electron, for example) in a potential is governed by a complex **wavefunction** $\Psi(\mathbf{r}, t)$ that satisfies the

Schrödinger equation[†]

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(\mathbf{r},t) + V(\mathbf{r},t)\Psi(\mathbf{r},t) = -j\hbar\frac{\partial\Psi(\mathbf{r},t)}{\partial t}\,. \tag{14.1-1}$$

The potential energy $V(\mathbf{r},t)$ characterizes the environment of the particle, including contributions from externally applied fields. The partial differential equation displayed in (14.1-1) thus has an enormous variety of solutions, depending on the form of $V(\mathbf{r},t)$. Systems that comprise multiple particles, such as atoms, ions, molecules, liquids, and solids, obey a more complex version of this equation in which the potential energy contains terms that accommodate interactions among the particles. Equation (14.1-1) is mathematically similar to the paraxial Helmholtz equation of wave optics (2.2-23) and to the paraxial slowly varying envelope equation of ultrafast optics (23.1-24).

The **Born postulate** of quantum mechanics specifies that the probability of finding the particle within an incremental volume dV surrounding the position $\mathbf{r}$, in the time interval between t and $t + dt$, is

$$p(\mathbf{r},t)\,dV\,dt = |\Psi(\mathbf{r},t)|^2\,dV\,dt. \tag{14.1-2}$$

Equation (14.1-2) resembles (13.1-15) for the probability of finding a photon within an incremental area and time.

In the absence of a time-varying potential, the allowed energy levels E of the particle are determined by using the technique of separation of variables. This leads to a solution of (14.1-1) of the form $\Psi(\mathbf{r},t) = \psi(\mathbf{r})\exp[j(E/\hbar)t]$, where $\psi(\mathbf{r})$ satisfies the **time-independent Schrödinger equation**

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + V(\mathbf{r})\psi(\mathbf{r}) = E\psi(\mathbf{r})\,. \tag{14.1-3}$$

Equation (14.1-3), which is similar to the Helmholtz equation (2.2-7), may be regarded as an eigenvalue problem for which the allowed values of the energy E are the eigenvalues, while the solutions $\psi(\mathbf{r})$ are the eigenfunctions (Appendix C).

Systems of multiple particles obey a generalized form of (14.1-3). The solutions provide the allowed values of the energy E of the system. These values can be discrete (as for an atom), or continuous (as for a free particle), or can comprise sets of densely packed discrete levels called bands (as for a semiconductor). The presence of thermal excitation, or of an external field such as light illuminating the material, can induce the system to move from one of its energy levels to another. This provides a means by which the system can exchange energy with the outside world.

In the following sections we schematically illustrate typical energy-level structures for selected atoms, ions, molecules, and solids.

A. Atoms

Atomic energy levels are established by the potential energies of the electrons in the presence of the atomic nucleus and the other electrons, as well as by interactions involving the orbital and spin angular momenta, which are usually much weaker than those involving charges. Isolated atoms such as Ne, Cu, and Cd, among others, serve as active laser media.

[†] The Schrödinger equation is cast in a form commonly used in electrical engineering. As explained in the footnote on page 47, correspondence with the form usually used in the physics literature is attained by simply replacing $-j$ with i, where $j = i = \sqrt{-1}$. Of course, this choice has no bearing on the final results.

Hydrogen

Energy levels. The energy levels of a hydrogen-like atom comprising a nucleus of charge $+Ze$, where Z is the **atomic number** ($Z = 1$ for H), and a single electron of charge $-e$ and mass m_0, are determined by inserting the Coulomb potential energy, $V(r) = -Ze^2/4\pi\epsilon_o r$, in the time-independent Schrödinger equation (14.1-3). Since $V(r)$ is a function of the radial coordinate alone, the Laplacian may be written in spherical coordinates, whereupon the partial differential equation splits into three ordinary differential equations via separation of variables. This leads to a ready solution for the eigenvalue problem, in which the eigenvalues comprise an infinite number of discrete energy levels with values

$$E_{\mathrm{n}} = -\frac{M_r Z^2 e^4}{(4\pi\epsilon_o)^2\, 2\hbar^2}\,\frac{1}{\mathrm{n}^2}, \qquad \mathrm{n} = 1, 2, 3, \ldots, \tag{14.1-4}$$

where the reduced mass of the atom M_r replaces the electron mass m_0 to accommodate the finite mass of the nucleus. The energy levels in (14.1-4), which are characterized by a single quantum number n called the principal quantum number, are displayed in Fig. 14.1-1 for H and C^{5+}.

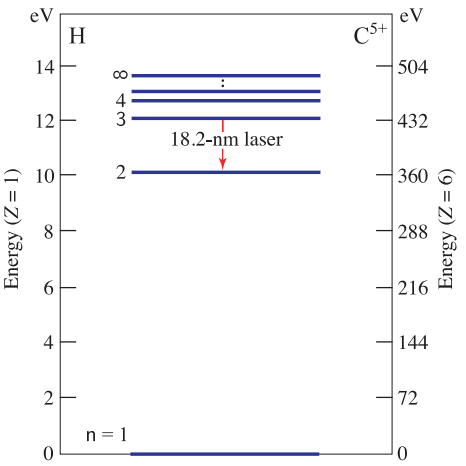


Figure 14.1-1 Energy levels of a hydrogen atom ($Z = 1$; left ordinate) and a C^{5+} ion (a hydrogen-like atom with $Z = 6$; right ordinate). In this depiction, the arbitrary zero of energy is placed at the ground-state level ($\mathrm{n} = 1$), which lies $\approx$ 13.6 eV below the ionization energy level ($\mathrm{n} = \infty$) for hydrogen and $\approx$ 489.8 eV below the ionization energy level for C^{5+}. The $\mathrm{n} = 3$ to $\mathrm{n} = 2$ transition for C^{5+}, indicated by the red vertical arrow, corresponds to the extreme-ultraviolet laser transition at 18.2 nm, as discussed in Sec. 16.3F.

Bohr theory. In the context of the **Bohr atom** associated with the *old quantum theory*, the energy levels provided in (14.1-4) can be obtained by equating the Coulomb force of attraction to the centrifugal force required to keep the electron in a circular orbit, while assuming that the electron orbital angular momentum is quantized to integer multiples of $\hbar$. The radii of these Bohr orbits turn out to be $r_{\mathrm{n}} = (4\pi\epsilon_o)\,\mathrm{n}^2\hbar^2/m_0 Ze^2$, $\mathrm{n} = 1, 2, 3, \ldots$. The radius of the lowest angular-momentum Bohr orbit for hydrogen, $a_0 \equiv r_{\{\mathrm{n}=1, Z=1\}} \approx 0.053$ nm, is known as the **Bohr radius**. The associated **Bohr period** is given by $T_0 = 2\pi a_0^2 m_0/\hbar \approx 150$ as.

Quantum numbers. The eigenfunctions of the Schrödinger equation take the form of a product of three functions, $\psi_{\mathrm{n}\ell\mathrm{m}}(r, \theta, \phi) = \mathbb{R}_{\mathrm{n}\ell}(r)\,\Theta_{\ell\mathrm{m}}(\theta)\,\Phi_{\mathrm{m}}(\phi)$, where $\mathrm{n} = 1, 2, 3, \ldots$ is the **principal quantum number**; $\ell = 0, 1, 2, \ldots, \mathrm{n} - 1$ is the **azimuthal quantum number**; and $\mathrm{m} = 0, \pm 1, \pm 2, \ldots, \pm\ell$ is the **magnetic quantum number**. Here, the $\mathbb{R}_{\mathrm{n}\ell}(r)$ represent associated Laguerre functions (which are related to the generalized Laguerre polynomials discussed in the footnote on page 103); the $\Theta_{\ell\mathrm{m}}(\theta)$ are associated Legendre functions; and the $\Phi_{\mathrm{m}}(\phi)$ are phase functions. These solutions bear some similarity to those for the microsphere microresonator discussed in Sec. 11.4C. Incorporating the intrinsic spin of the electron requires a fourth quantum number known as the **spin quantum number**: $\mathrm{s} = \pm 1/2$.

Fine structure, hyperfine structure, and relativistic effects. The electromagnetic interaction between the magnetic dipole moment associated with the electron spin and the magnetic field generated by the orbital angular momentum of the electron rotating about the nucleus is referred to as **spin–orbit coupling**. It serves to split the otherwise degenerate energy levels of the hydrogen atom into closely spaced, but distinct, components called **fine structure**. The electron spin and its orbital angular momentum also interact with the nuclear magnetic dipole moment to produce yet finer splittings, called **hyperfine structure**. Other interactions (e.g., spin–spin coupling) are also present, but are negligible. All of these effects cause the energy levels of hydrogen to differ slightly from those specified in (14.1-4). **Relativistic shifts** of the energy levels, which are small but measurable, can be accommodated by using the relativistically invariant **Dirac equation** in place of the nonrelativistic Schrödinger equation. The Dirac equation also intrinsically gives rise to the existence of electron spin and antiparticles such as the positron.

Multielectron Atoms

Shells and subshells. Multielectron atoms comprise a nucleus of charge $+Ze$ surrounded by Z electrons, each of charge $-e$. As the atomic number Z increases, the occupation of successive single-electron states proceeds by minimizing the total energy while satisfying the **Pauli exclusion principle**, which provides that, as fermions, no two electrons may have the same set of four quantum numbers. The electron states thus fill in the form of **shells**, designated by the principal quantum number n, each of which has the capacity to hold a specific number of electrons. Within each shell are constituent **subshells**, designated by the azimuthal quantum number ℓ, as specified in spectroscopic notation (s, p, d, f, g, h, i, ... correspond to $\ell = 0, 1, 2, 3, 4, 5, 6, \ldots$, respectively). The **electron configuration** $\mathrm{n}\ell^{\mathrm{u}}$ represents the arrangement of electrons in the subshells; the superscript u indicates the number of electrons present in each. For example, the configuration for the ground state of He ($Z = 2$) is $1s^2$, its two electrons just filling the $\mathrm{n} = 1$ shell, for which $\ell = 0$. Low-lying excited-state configurations of He in which one of the electrons has been excited to the $\mathrm{n} = 2$ shell include $1s2s$ and $1s2p$. For Ne ($Z = 10$), the ground-state configuration is $1s^2 2s^2 2p^6$; its 10 electrons just fill the $\mathrm{n} = 1$ and $\mathrm{n} = 2$ shells, which accommodate 2 and 8 electrons, respectively.

Energy levels. The energy levels of multielectron atoms can be approximately determined by using the Schrödinger theory, which is suitable for atoms in which relativistic effects are insignificant. Because of the myriad Coulomb interactions inherent in a collection of electrons, the Schrödinger equation is typically solved via an approximate, self-consistent approach called the Hartree method. Each electron is considered to move independently in a spherically symmetric net potential, which is taken to be the sum of the spherically symmetric attractive Coulomb potential arising from the nucleus and a spherically symmetric repulsive potential representing the average effect of the Coulomb forces from all other electrons. The Z-electron Schrödinger equation then splits into Z single-electron Schrödinger equations with an overall eigenfunction that is a product of the individual-electron eigenfunctions, and a total energy that is the sum of the energies of the individual electrons. Finally, perturbation theory is used to accommodate the deviations from spherical symmetry of the repulsive potential and for interactions involving electron spin. The resultant single-electron eigenfunctions are closely related to those for the hydrogen atom and are written in the same form. However, energy formulas similar to (14.1-4) exist only for optically active electrons such as the valence electrons in alkali atoms.

Electron configuration and term symbol. Electron configurations often have corresponding collections of closely spaced energy sublevels that comprise **manifolds**, as displayed schematically in Fig. 14.1-2 for Ne. These fine-structure splittings are principally a consequence of spin–orbit coupling, which introduces level shifts that are typically 1 part in 10^4 for H and grow larger as the atomic number Z increases (hyperfine shifts are a factor of 10^3 smaller and relativistic effects can be safely ignored). For the lighter multielectron atoms, the total orbital angular momentum L and the total spin angular momentum $\mathcal{S}$ are good quantum numbers and the interaction between them, known as Russell–Saunders or $\mathrm{L}\mathcal{S}$ coupling, is responsible for the spin–orbit coupling. Atomic states are then well described by **term symbols** of the form ${}^{2\mathcal{S}+1}\mathrm{L}_{\mathcal{J}}$, which display the various angular momenta: $\mathcal{S}$ is the *total spin angular-momentum* quantum number and $2\mathcal{S}+1$ is the *spin multiplicity* (e.g., triplet, singlet); L is the *total orbital angular-momentum* quantum number in spectroscopic notation (uppercase letters $S,\ P,\ D,\ F, \ldots$ represent $\mathrm{L} = 0,\ 1,\ 2,\ 3, \ldots$, respectively); and $\mathcal{J}$ is the *total overall angular-momentum* quantum number. The term symbol for an atom is often displayed just after the electron configuration. For example, the lowest-lying excited states of He are denoted $1s2s\ {}^3S_1$ and $1s2s\ {}^1S_0$, with triplet and singlet multiplicities, respectively, as shown in Fig. 14.1-2. When all occupied subshells are filled (as is the case for the ground states of the noble gases as well as for a number of other atoms such as Cd and Yb), the term symbol is ${}^1S_0$.

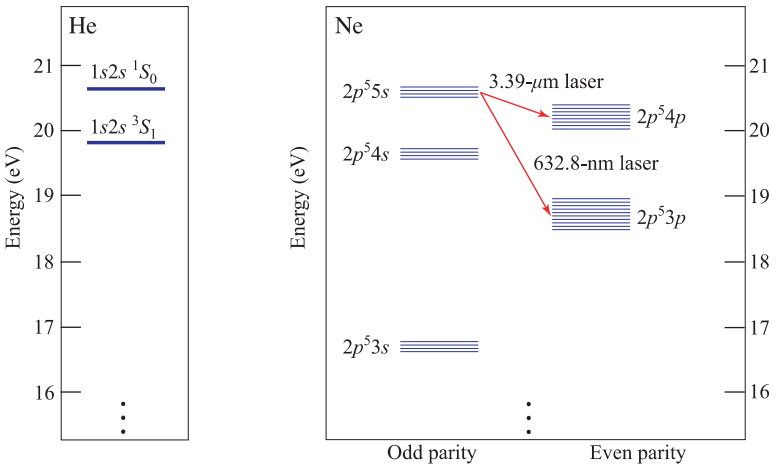


Figure 14.1-2 Selected excited-state energy levels of He and Ne atoms. Electron configurations and term symbols are indicated (by convention, the electron-configuration prefix for Ne, $1s^2 2s^2$, corresponding to filled subshells, is suppressed). The energy spacings between the fine-structure splittings, which are illustrated schematically, are greatly exaggerated. The Ne transitions marked by red arrows correspond to the wavelengths 3.39 μm and 632.8 nm, as indicated. These common Ne laser transitions lie in the mid-infrared and visible, respectively (Secs. 15.3E and 16.3G). The close energy matches between the excited He and Ne energy levels, which are fortuitous, facilitate excitation of the Ne atoms via collisions in a gas-discharge tube — hence the moniker "He–Ne laser."

Periodic table. The larger the value of the principal quantum number n, the less tightly bound are the electrons to the atom because of the moderation of the nuclear potential by the Coulomb screening provided by the inner electrons. As a result, shells typically fill in the order $\mathrm{n} = 1, 2, 3, 4, \ldots$. Similarly, the larger the value of ℓ, the less tightly bound are the electrons because the electron probability density progressively shifts toward the atomic periphery. Hence, subshells typically fill in the order $s, p, d, f, \ldots$. As a consequence of these successive filling processes, many

properties of the elements are periodic functions of Z, as exemplified by the periodic table displayed in Fig. 14.1-3. Successive rows of the table correspond to consecutive values of the principal quantum number n.

Each column of the table contains a group of elements whose physical and chemical properties bear a certain similarity to each other because they contain the same number of electrons in their outermost shells (valence electrons). Column ⑱, for example, comprises the noble gases, including He and Ne, which are monoatomic and chemically inert because they have filled outer shells and a large energy difference between their filled p subshells and the next higher s subshells. Columns ① and ⑰, in contrast, comprise elements that are highly active chemically and easily form molecules. Each alkali-metal atom in column ①, for example, contains a lone outer electron that it will readily share with any nearby halogen atom in column ⑰, which needs just such a lone electron to complete its outer shell.

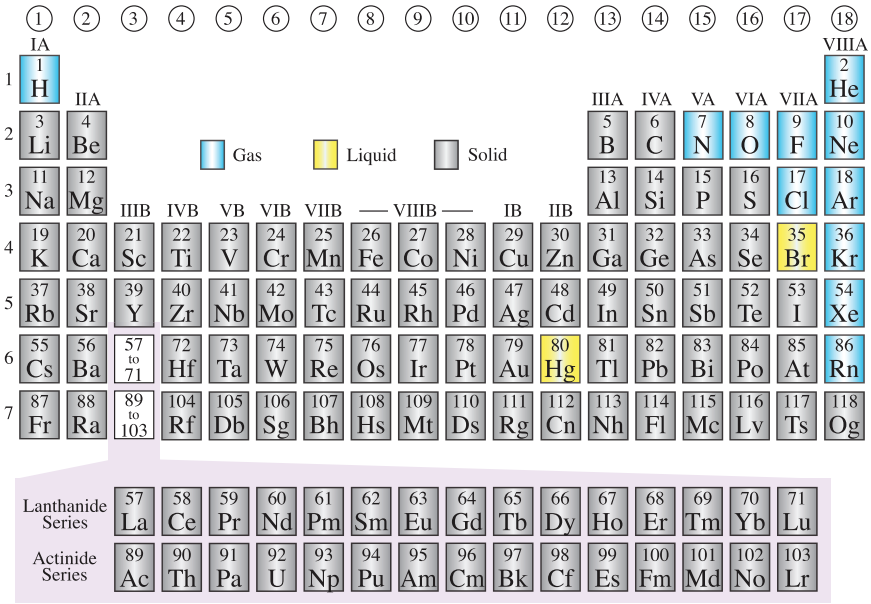


Figure 14.1-3 Periodic table of the elements, with element abbreviations and atomic numbers Z indicated. Each successive row of the table, called a *period*, comprises elements containing electrons that reside in shells designated by the principal quantum number n indicated by the arabic numeral at left. Each column of the table, called a *group*, comprises elements with similar physical and chemical properties, and is designated by an encircled arabic numeral at top. The roman numerals are an older, but still widely used system for designating the groups; it remains prevalent in semiconductor physics. The *lanthanide* and *actinide* series depicted in the lower portion of the figure reside in rows 6 and 7, and comprise the successive additions of electrons to the inner $4f$ and $5f$ subshells, respectively. The *rare-earth elements* comprise the elements in the lanthanide series (except for La itself) plus Sc and Y. Elements that take the form of gases, liquids, and solids at room temperature are indicated in blue, yellow, and silver, respectively.

In general, multielectron atoms exhibit an enormous variety of allowed energy levels. Even though optical transitions typically involve only valence electrons, the sheer abundance of energy levels gives rise to a cornucopia of energy differences, many of which serve as viable laser wavelengths (Secs. 15.3E and 16.3G indicate but a few). The energy differences between the excited atomic levels of Ne displayed in Fig. 14.1-2, for example, extend to several eV, which covers the infrared and optical regions of the spectrum.

Relative atomic mass. The nucleus of an element of fixed atomic number Z contains Z protons and an assembly of neutrons. Different *isotopes* of a given element all contain Z protons, but different numbers of neutrons. Naturally occurring elements often comprise a collection of isotopes; the abundance-weighted *relative atomic mass* of a particular sample of the element is denoted A_r. Mononuclidic elements, such as ${}^{A_r=23}_{Z=11}\mathrm{Na}$, have a single naturally occurring (or strongly dominant) isotope. Just as elementary particles are either bosons or fermions (Sec. 13.1D), so too are composite particles such as atoms. If the total number of constituent fermions (electrons, protons, and neutrons) in the atom is even (odd), it is a composite boson (fermion), and has integer (half-integer) spin. Examples of bosonic atoms are ${}^{4}_{2}\mathrm{He}$ and ${}^{23}_{11}\mathrm{Na}$; examples of fermionic atoms are ${}^{3}_{2}\mathrm{He}$, ${}^{22}_{11}\mathrm{Na}$, and ${}^{87}_{38}\mathrm{Sr}$. Bosons and fermions obey Bose–Einstein and Fermi–Dirac statistics, respectively.

External fields. The application of an external magnetic or electric field to an atom serves to split and modify otherwise degenerate energy levels via the **Zeeman effect** and **Stark effect**, respectively. The Zeeman effect results from the interaction of an external magnetic field with the overall magnetic dipole moment of an atom (including its orbital and spin components). The Stark effect, which is the electric-field analog of the Zeeman effect, results from the interaction of an external electric field with the induced electric dipole moment of the atom. Splittings resulting from an applied AC electric field are said to arise from the **AC Stark effect**. The magnitudes of both the Zeeman and Stark energy splittings increase with increasing field strength. Both effects play important roles in laser cooling and trapping, as well as in other applications.

Ionization energies. The ionization energy of a neutral atom is the energy necessary to remove its most loosely bound valence electron to the vacuum, leaving behind a cation. In each row of the periodic table presented in Fig. 14.1-3, the ionization energy generally increases as one moves to the right, exhibiting a minimum for the alkali metal in column ① (which has only a single electron outside a closed shell), and a maximum for the noble gas in column ⑱ (which has a closed shell). The energy of the electron in the ground state of hydrogen, for example, is -13.6 eV with respect to the vacuum level (Fig. 14.1-1), so the ionization energy of hydrogen is $\mathbb{W} = 13.6$ eV. It is also possible to ionize a multielectron atom by removing an inner-shell electron, which has a far higher ionization energy; and to create highly ionized multielectron atoms, in which multiple electrons are removed (Sec. 16.3F).

B. Ions and Doped Dielectric Media

The removal of one or more electrons from an atom leaves behind a residual cation with its own electron configuration and term symbol. Though the energy levels associated with the core electrons remain the same as those of the parent atom, the subshells in which the optically active electrons reside are unique to the ion, as are their energy levels. Much as with multielectron atoms, ions offer an enormous variety of energy levels and potential laser wavelengths. Among the earliest lasers to be developed were ionic gas lasers that relied on the energy levels of noble-gas ions such as Ar^{+} and Kr^{+}, created from their respective neutral atoms in a gas-discharge tube (Sec. 16.3E). Solid-state lasers that relied on the energy levels of lanthanide-metal ions, such as Nd^{3+} substituting for a small fraction of the Y^{3+} ions in a $\mathrm{Y_3Al_5O_{12}}$ crystal, were developed in the very same year (1964). However, by virtue of their superior performance and robustness, solid-state lasers have nearly totally eclipsed ionic gas lasers, except in the most specialized of applications.

Much as with atoms, the energy levels of the laser ions are established via quantum-mechanical calculations or, as is more often the case in practice, empirically. Transition-metal and lanthanide-metal ions are generally used as dopants for solid-state laser amplifiers and lasers (Secs. 15.3 and 16.3A). The host media are usually insulating ionic or covalent dielectrics that are transparent in a particular region of the spectrum, with suitable optical, thermal, and mechanical properties. The optical properties of dielectric materials were considered in Sec. 5.5C in the context of the Lorentz oscillator model, which is suitable for characterizing transparent host materials.

The extent to which the energy levels of an active laser ion remain unaffected by the host medium is determined principally by how well the ion's optically active electrons are shielded from the host's neighboring lattice atoms. It will become clear that the energy levels of transition-metal ions are substantially modified by the local field effects of the host whereas those of lanthanide-metal (rare-earth) and actinide-metal ions are scarcely affected. By way of example, we will consider the energy levels of four well-known laser systems: 1) Cr^{3+}:Al_2O_3 (ruby); 2) Cr^{3+}:$BeAl_2O_4$ (alexandrite); 3) Nd^{3+}:$Y_3Al_5O_{12}$ (Nd^{3+}:YAG); and 4) Nd^{3+}:glass.

Transition-Metal Ions

The most commonly encountered transition-metal ion dopants for lasers are the trivalent ions Ti^{3+} and Cr^{3+}, although some lasers make use of Cr^{2+}, Cr^{4+}, Ni^{2+}, Co^{2+}, as well as other ions. The ground-state electron configurations and term symbols for some of these ions, and for their respective elements, are presented in Table 14.1-1. It is evident from the table that the optically active electrons for the transition-metal ions reside in the $3d$ subshell.

Table 14.1-1 Selected transition-metal, lanthanide-metal (rare-earth), and actinide-metal ions used in solid-state lasers. Ground-state electron configurations and term symbols for isolated atoms and ions are provided.

	Atom			Ion		
Atomic Number Z	Element	Configuration[a]	Term	Ion	Configuration[a]	Term
Transition Metals						
22	Ti	$3d^2\,4s^2$	$^3F_2$	Ti^{3+}	$3d^1$	$^2D_{3/2}$
				Cr^{2+}	$3d^4$	$^5D_0$
24	Cr	$3d^5\,4s^1$	$^7S_3$	Cr^{3+}	$3d^3$	$^4F_{3/2}$
				Cr^{4+}	$3d^2$	$^3F_2$
26	Fe	$3d^6\,4s^2$	$^5D_4$	Fe^{2+}	$3d^6$	$^5D_4$
Lanthanide Metals						
60	Nd	$4f^4\,6s^2$	$^5I_4$	Nd^{3+}	$4f^3$	$^4I_{9/2}$
68	Er	$4f^{12}\,6s^2$	$^3H_6$	Er^{3+}	$4f^{11}$	$^4I_{15/2}$
69	Tm	$4f^{13}\,6s^2$	$^2F_{7/2}$	Tm^{3+}	$4f^{12}$	$^3H_6$
70	Yb	$4f^{14}\,6s^2$	$^1S_0$	Yb^{3+}	$4f^{13}$	$^2F_{7/2}$
Actinide Metals						
92	U	$5f^3\,6d^1\,7s^2$	$^5L_6$	U^{3+}	$5f^3$	$^4I_{9/2}$

[a] By convention, the electron configurations for filled subshells are omitted; this includes those for the $5s^2\,5p^6$ filled subshells in the $n = 5$ shell of the lanthanides and those for the $6s^2\,6p^6$ filled subshells in the $n = 6$ shell of the actinides.

Ruby and alexandrite. We now proceed to examine the energy levels of two dielectric media doped with Cr^{3+}, namely ruby and alexandrite (Fig. 14.1-4). Ruby is celebrated because it was the material from which the first laser was made (page 657), whereas alexandrite received considerable early attention because its output is tunable

over a range of wavelengths. The energy levels of Ti:sapphire, perhaps the most important transition-ion laser material, will be considered in Sec. 16.3A.

Ruby (Cr^{3+}:Al_2O_3) is chromium aluminum oxide. It is a dielectric medium with refractive index $n \approx 1.76$ that is composed principally of sapphire (Al_2O_3, also called aluminum oxide, alumina, and corundum), in which a small fraction of the Al^{3+} ions ($\approx 0.05\%$) are replaced by Cr^{3+} ions. **Alexandrite** (Cr^{3+}:$BeAl_2O_4$) is formed by doping a small amount of chromium oxide ($\approx 0.1\%$) into a chrysoberyl ($BeAl_2O_4$) host. This material has a refractive index that is close to that of ruby, $n \approx 1.75$; however chrysoberyl is biaxial whereas sapphire is uniaxial.

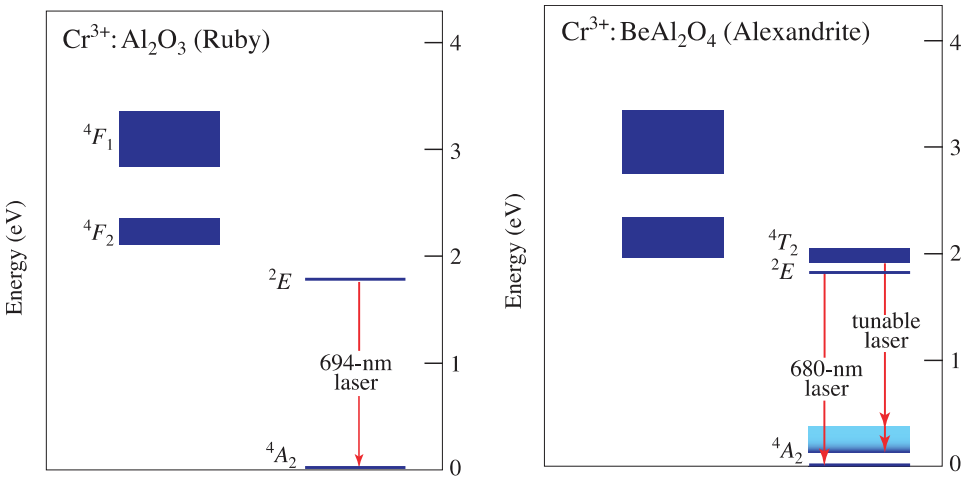


Figure 14.1-4 Selected energy levels and energy bands for Cr^{3+}:Al_2O_3 (ruby) and Cr^{3+}:$BeAl_2O_4$ (alexandrite). The red arrows represent laser transitions. Each laser emits light at a characteristic fixed wavelength (694 nm for ruby and 680 nm for alexandrite). However alexandrite also lases over a range of additional wavelengths. The dark-to-light shading of the lower laser band in alexandrite indicates a decrease in relative occupancy. Because of the important role of the crystal field in determining the energy levels of transition-metal ions in dielectric hosts, group-theoretical symbols, rather than term symbols, are generally used to designate them.

Since the $3d$ electrons of the Cr^{3+} ions in both materials are exposed to neighboring ions, the energy levels of these materials are determined in large part by the surrounding crystal fields and therefore depend substantially on the host material. In particular, each chromium ion is surrounded by oxygen atoms in a configuration that subjects it to a significant spatially varying potential. Best represented in the context of **crystal-field theory** (or **ligand-field theory**), this potential, along with that of the Cr^{3+} nucleus, determine the energy levels of ruby and alexandrite via the Schrödinger equation. As a consequence, the energy levels of transition-metal ions in a dielectric host are generally designated by group-theoretical symbols (e.g., ^2E and $^4A_2$) rather than by term symbols.

The resultant energies are a mixture of discrete levels and energy bands, some of which are shown in Fig. 14.1-4. The energy levels of the two materials are quite distinct even though they share the same dopant ion. In particular, the $^4A_2$ energy band in alexandrite comprises a collection of **vibronic states** (shaded-blue region) that result from coupling between the electronic energy levels and the lattice vibrations of the crystal. Consequently, alexandrite is tunable over a (limited) range of wavelengths whereas ruby is not (Sec. 15.3E). Nevertheless, alexandrite also lases at a characteristic wavelength (680 nm) that is not too far from that of ruby (694 nm).

Lanthanide-Metal Ions

The **lanthanide** elements comprise the series from $_{58}$Ce to $_{71}$Lu that reside in row 6 of the periodic table (Fig. 14.1-3). These elements, plus $_{21}$Sc and $_{39}$Y, are often called **rare earths** because they were long ago thought to be rare (they are in fact rarely rare). The lanthanide elements are constructed by successively adding electrons to the $4f$ subshell, which lies within the filled $5s^2 5p^6$ and $6s^2$ subshells, as shown in Table 14.1-1. The lanthanides usually exist as trivalent cations, in which case the configuration of their valence electrons takes the form $4f^{\mathrm{u}}$, with u varying from 1 (Ce^{3+}) to 14 (Lu^{3+}).

The lanthanide ions Nd^{3+}, Er^{3+}, Tm^{3+}, and Yb^{3+} are particularly important for laser amplifiers and oscillators. The electron configurations and term symbols for these trivalent ions, and their respective elements, are provided in Table 14.1-1. Nd^{3+}:glass and Er^{3+}:silica-glass fibers are widely used as laser amplifiers, as will be highlighted in Secs. 15.3B and 15.3C, respectively. Nd^{3+}:YVO_4, Nd^{3+}:YAG, and Yb^{3+}:YAG often serve as laser oscillators, as discussed in Sec. 16.3A. Among the other lanthanides, Pr^{3+} and Ho^{3+} also find use as active laser ions. Two or more lanthanide ions are frequently used to co-dope laser media in order to improve performance.

Nd^{3+}:YAG and Nd^{3+}:glass. The energy levels of trivalent lanthanide ions in a dielectric host and in isolation are quite similar. This results from the fact that the optically active $4f$ electrons are well shielded from the external effects of the lattice by the filled $5s$ and $5p$ subshells (Table 14.1-1). This is in sharp contrast to the behavior of transition-metal ions; unlike ruby and alexandrite, lanthanide-ion energy levels are rather independent of the host material. This is illustrated in Fig. 14.1-5 for Nd^{3+} in two hosts that are quite different: YAG and phosphate glass. The main near-infrared laser-transition wavelengths in the two materials, corresponding to the energy differences between the $^4F_{3/2}$ and $^4I_{11/2}$ levels, are in close alignment: 1.064 μm for Nd^{3+}:YAG and 1.053 μm for Nd^{3+}:glass.

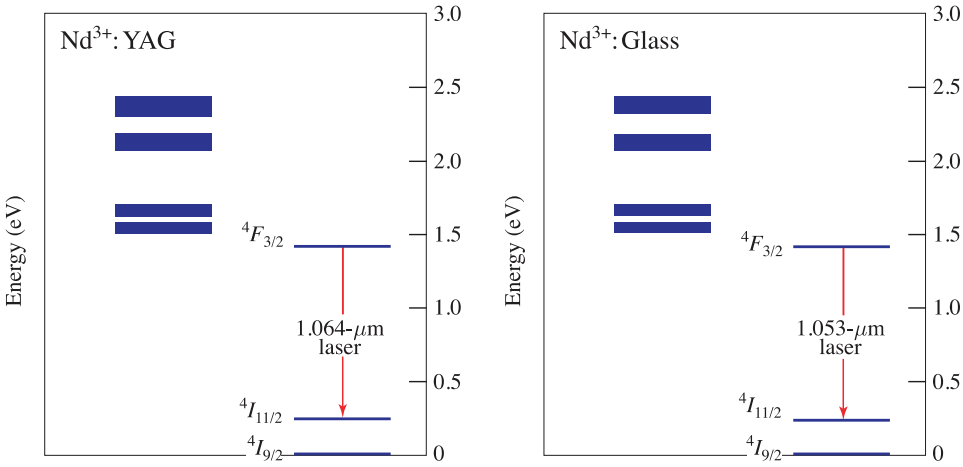


Figure 14.1-5 Selected energy levels of Nd^{3+} in YAG and in phosphate glass. The arrows indicate the principal near-infrared laser transition, which has a wavelength of 1.064 μm in YAG and 1.053 μm in phosphate glass. The energy levels of both materials are rather similar at the scale of this figure. Since the energy levels of the dopant ions are scarcely affected by the host, ionic term symbols are used to designate them (Table 14.1-1). A magnified view of the manifolds for Nd^{3+}:YAG, provided in Fig. 14.1-6, shows the multiple sublevels.

Lanthanide-ion manifolds. However, the energies of the sublevels within the lanthanide-ion manifolds do depend on the host material. Nd^{3+}:YAG and Nd^{3+}:glass exhibit quite different sublevel structures because of differences in their local field environments. Incorporating Nd^{3+} ions in a crystal such as YAG results in homogeneous broadening, whereas embedding them in a less structured material such as phosphate glass results in inhomogeneous broadening (Sec. 14.3D). The sharp Nd^{3+}:YAG sublevels are clearly visible in Fig. 14.1-6, which offers a greatly magnified view of Fig. 14.1-5. For a glass host, these narrow sublevels are smeared into bands.

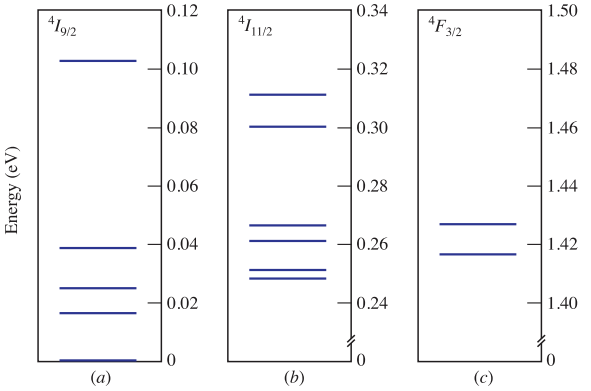


Figure 14.1-6 Sublevels of the three manifolds associated with Nd^{3+}:YAG laser transitions near 1.06 μm: (*a*) ground state $^4I_{9/2}$; (*b*) lower laser level $^4I_{11/2}$; (*c*) upper laser level $^4F_{3/2}$. The specific energies of the sublevels depend on the host material; they are smeared into bands for hosts such as glass, where inhomogeneous broadening prevails.

The number of distinct sublevels within each of the manifolds displayed in Fig. 14.1-6 is determined by the value of $g/2$, where $g = 2\mathcal{J} + 1$. Here g is the **degeneracy parameter** and $\mathcal{J}$ is the total overall angular-momentum quantum number, which is provided in the term symbol $^{2\mathcal{S}+1}\mathrm{L}_{\mathcal{J}}$, as discussed in Sec. 14.1A for atoms. For the three Nd^{3+}:YAG manifolds displayed in Fig. 14.1-6, the numbers of distinct sublevels are $(2\mathcal{J} + 1)/2 = 5$, 6, and 2, respectively. The celebrated 1.06415-μm laser line is associated with a transition between the upper sublevel in the $^4F_{3/2}$ manifold at 1.4269 eV and the third-from-bottom sublevel in the $^4I_{11/2}$ manifold at 0.2616 eV. (When frequency doubled, this transition provides the well-known source of green light at 532 nm.)

Actinide-Metal Ions

The **actinide** elements residing in row 7 of the periodic table (Fig. 14.1-3) are constructed by incrementing the number of electrons in the $5f$ subshell. The optically active $5f$ electrons in the actinide ions are shielded from the host lattice by the filled $6s$ and $6p$ subshells (Table 14.1-1). The chemical behavior of the actinides is thus similar to that of their lanthanide homologs. The U^{3+}:CaF_2 laser, developed just months after the ruby laser was demonstrated in 1960, operates on the $^4I_{11/2} \rightarrow {}^4I_{9/2}$ transition of U^{3+} at $\lambda_o \approx 2.49$ μm. Neodymium lies immediately above uranium in the periodic table; both Nd^{3+} and U^{3+} share the same f^3 ground-state electron configuration and $^4I_{9/2}$ term symbol (Table 14.1-1). While the development of actinide-ion-doped laser materials was impeded by their relative rarity and by the radioactive nature of many of their isotopes, the advancement of lanthanide-ion-doped laser materials such as Nd^{3+}:CaF_2 proceeded apace. Nevertheless, U^{3+}:CaF_2 was the first four-level laser ever operated.

C. Molecules

Molecules can be formed by the combination of two or more atoms. A stable molecule emerges when the sharing of valence electrons by the constituent atoms results in a

reduction of the overall energy. Molecular bonds take several forms and differ widely in their strengths. Two important forms of molecular bonding are **covalent bonding** and **ionic bonding**. In the simplest view of the covalent bond, the negatively charged electrons of the constituent atoms are simultaneously attracted by the positive charges of two or more atoms. For the ionic bond, on the other hand, an electron is essentially transferred from one constituent atom to another, resulting in an electrostatic attraction between them. Weak residual bonding between molecules that arises neither from covalent nor ionic bonding is known as **van der Waals bonding**. Such bonding often arises from the interaction of molecular dipoles and may be attractive or repulsive.

Molecular energy levels are determined in part by the nature of the bonding and the potential energies associated with the interatomic forces that bind the atoms. The energy levels of a molecule are typically associated with three more-or-less distinct features, whose transitions typically fall in different wavelength regions: *rotational transitions* lie in the microwave and far infrared, *vibrational transitions* lie in the infrared, and *electronic transitions* lie in the visible and ultraviolet. Since the time scales of these features differ considerably, to first approximation they may be analyzed separately. Molecules ranging from simple gases to dyes in a solvent serve as active laser media (Sec. 16.3E).

Rotating Diatomic Molecule

The rotation of a diatomic molecule with moment of inertia $\mathcal{I}$ about its center of mass can be considered as the rotation of a rigid rotor about an axis perpendicular to its internuclear axis. The classical rotational energy for such a system is $E_{\mathrm{r}} = L^2/2\mathcal{I}$, where L is the angular momentum of the system about the axis of rotation. In accordance with the laws of quantum mechanics, the square-magnitude of the angular momentum of such a system is quantized according to $L^2 = \mathrm{r}(\mathrm{r}+1)\hbar^2$, where r is the rotational quantum number. The allowed energy levels of the rotating diatomic molecule are thus

$$E_{\mathrm{r}} = \frac{1}{2\mathcal{I}}\,\mathrm{r}(\mathrm{r}+1)\hbar^2, \qquad \mathrm{r} = 0,\ 1,\ 2, \ldots . \tag{14.1-5}$$

The energy separations $\hbar\omega$ of rotational energy levels typically lie in the range between 10^{-4} and 10^{-2} eV, corresponding to photons in the microwave and far-infrared regions of the spectrum. This energy spacing increases with increasing quantum number r, in contrast to the spacing between successive electronic energy levels of hydrogen-like atoms, which decrease with increasing quantum number in accordance with (14.1-4).

Vibrating Diatomic Molecule

The vibrations of a diatomic molecule (e.g., N_2, CO, HCl) are governed by an intramolecular attraction and a restoring force that is approximately proportional to the change in internuclear distance x. The system may therefore be modeled as two masses, M_1 and M_2, joined by a spring, with reduced mass $M_r = M_1M_2/(M_1 + M_2)$. A molecular spring constant κ can be defined such that the potential energy is $V(x) = \frac{1}{2}\kappa x^2$. The molecular vibrations (e.g., for the N_2 molecule portrayed in Fig. 14.1-7) therefore take on the energy levels of a quantum-mechanical harmonic oscillator.

As discussed in Sec. 13.3, the levels are quantized in accordance with

$$E_{\mathrm{v}} = \left(\mathrm{v} + \tfrac{1}{2}\right)\hbar\omega, \qquad \mathrm{v} = 0, 1, 2, \ldots, \tag{14.1-6}$$

where $\omega = \sqrt{\kappa/M_r}$ is the (angular) oscillation frequency, $\frac{1}{2}\hbar\omega$ is the zero-point energy, and v is the vibrational quantum number. Equation (14.1-6) matches the expression for the allowed energies of a mode of the electromagnetic field provided

in (13.1-5). Typical values of $\hbar\omega$ for molecular vibrations lie in the range between 0.05 and 0.5 eV, corresponding to the mid-infrared region of the spectrum (as for the N_2 vibrational energy levels displayed in Fig. 14.1-7).

Unlike the energy levels of the hydrogen atom, and those of the rotating diatomic molecule, the vibrational energy levels of the diatomic molecule are equally spaced. In practice, however, the potential-energy curves for most molecules become anharmonic as the energy increases (Sec. 22.7), which results in a diminution of the energy-level separations as v increases. In the course of undergoing a vibrational transition, the molecule may simultaneously alter its rotational state, whereupon both v and r change in the vibrational–rotational spectrum.

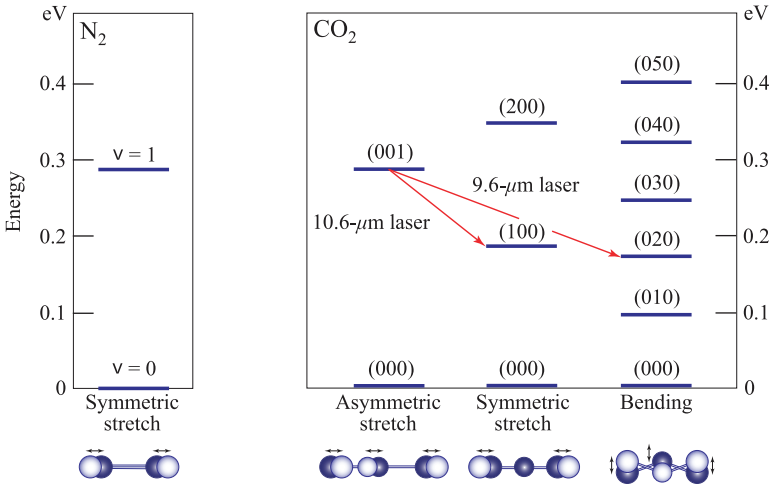


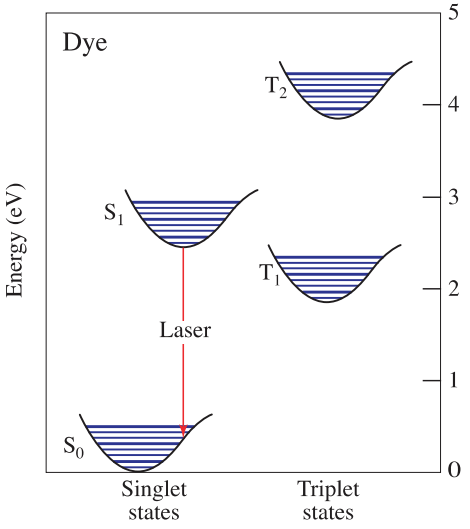
Figure 14.1-7 Vibrational energy levels of the N_2 and CO_2 molecules (the zero of energy is arbitrarily set at $v = 0$). The allowed energy levels for the three modes of vibration of CO_2, antisymmetric stretch (A), symmetric stretch (S), and bending (B), are displayed schematically. A triplet of quantum numbers $(v_S v_B v_A)$ characterize the excitation of the modes. The transitions indicated by red arrows represent laser emission at the iconic CO_2 mid-infrared laser wavelengths $\lambda_o = 10.6\ \mu m$ and $9.6\ \mu m$, as indicated. A manifold of finely spaced rotational energy levels (not shown) is associated with each vibrational level. The close energy match between the excited ($v = 1$) N_2 and (001) CO_2 energy levels, which is fortuitous, facilitates collisional excitation of the CO_2 molecules in a gas-discharge tube.

Vibrating Triatomic Molecule

A triatomic molecule of particular interest in photonics is carbon dioxide, which serves as a useful laser medium for generating high optical powers in the mid-infrared region. Since it comprises three atoms and is linear, the CO_2 molecule may undergo independent vibrations of three kinds, as illustrated in Fig. 14.1-7: asymmetric stretch (A), symmetric stretch (S), and bending (B). These normal modes exhibit the features of quantum-mechanical harmonic oscillators, each with its own spring constant, value of $\hbar\omega$, and equally spaced energy levels. The expression for the energy of the molecule is thus a sum of three terms, each of the form of (14.1-6), and the excitation of the system is characterized by a triplet of vibrational quantum numbers: $(v_S v_B v_A)$. The transitions indicated by the red arrows in Fig. 14.1-7 represent energy exchanges between normal modes that lead to photon emission at the well-known CO_2 laser wavelengths near $\lambda_o = 10.6\ \mu m$ $[(001) \rightarrow (100)]$ and $\lambda_o = 9.6\ \mu m$ $[(001) \rightarrow (020)]$ (Secs. 15.3E and 16.3G). As with diatomic molecules, each vibrational level is split into a manifold of closely spaced rotational levels (not shown), whose energies are given approximately by (14.1-5).

Dye Molecule

Organic dyes are large and complex molecules. As a result, they may undergo joint electronic, vibrational, and rotational transitions; they therefore have a vast array of energy levels that comprise both singlet (S) and triplet (T) multiplicities. In the singlet state, the spin of the excited electron is antiparallel to that of the remainder of the dye molecule, whereas in the triplet state the spins are parallel. Energy-level differences correspond to wavelengths covering broad swaths of the optical and ultraviolet regions. Figure 14.1-8 portrays the structure of Rhodamine-6G, which becomes an ion when dissolved in a solution of water or alcohol, along with an idealization of its energy-level structure. This particular dye served as a welcome medium during the formative years of laser development since it lased in the yellow region of the spectrum and could be tuned over a reasonable wavelength range. The organic dye laser is briefly discussed in Sec. 16.3E.



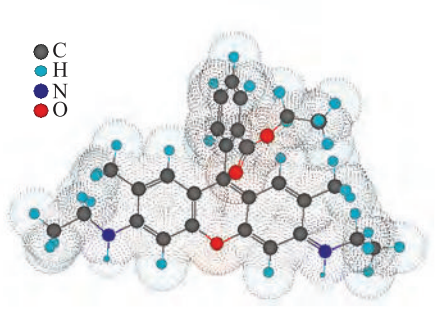


Figure 14.1-8 Structure of the Rhodamine-6G ion, which has the chemical formula $C_{28}H_{31}N_2O_3^+$. The schematic at left illustrates a laser transition between two singlet manifolds with slightly different configurations, as indicated by their horizontal offset. Vibrational and rotational energy levels are represented by thick and thin horizontal lines, respectively.

D. Solids

The atoms (or molecules) of solids lie in close proximity to each other and typically coalesce into a periodic arrangement comprising a **crystal lattice**. The strength of the forces holding the atoms together is roughly of the same magnitude as the forces that bind atoms into molecules. Consequently, the energy levels of solids are determined not only by the potentials associated with individual atoms, but also by the potentials associated with neighboring lattice atoms. Though noncrystalline solids, such as glasses and plastics, have orderly structures similar to those of crystals, they extend only over a short range.

Four principal types of bonding occur in ordinary solids: ionic, covalent, metallic, and molecular. **Ionic solids** (such as CaF_2) comprise a crystalline array of positive and negative ions held together by electrostatic attraction. Since there are no free electrons to carry current, these materials are insulators. They are generally transparent in the visible region of the spectrum since their bandgaps usually lie in the ultraviolet (Fig. 5.5-1). **Covalent solids**, like covalently bound molecules, consist of atoms bound by shared valence electrons. They are often insulators and can be transparent (such as

diamond) or opaque (such as graphite) in the visible region. Covalent solids can also be semiconductors (such GaAs), which are opaque in the visible and transparent in the infrared (Fig. 5.5-1). **Metallic solids** have delocalized valence electrons that are shared by all of the positive ions, moving in their combined potential. The ability of the electrons to wander through metallic crystals is responsible for their high electrical conductivities. Metals strongly reflect light and are opaque in the visible. **Molecular solids** (or **van der Waals solids**) contain small, non-polar covalent molecules held together by van der Waals forces, which are far weaker than those involved in other kinds of binding.

Energy bands. It is instructive to examine how the energy levels of an isolated atom are modified as it comes into close contact with neighboring atoms in the course of forming a crystal lattice. Isolated atoms and molecules (e.g., those in gases) exhibit discrete energy levels (see Figs. 14.1-1, 14.1-2, 14.1-7, 14.1-8, for example). Each individual atom in a collection of such identical isolated atoms has an identical set of discrete energy levels. As these atoms are brought into proximity to form a solid, exchange interactions (arising from the quantum-mechanical requirement of indistinguishability for identical particles), along with the presence of fields of varying strengths from neighboring atoms, become increasingly important. The initially sharp energy levels associated with the valence electrons of isolated atoms gradually broaden into collections of numerous densely spaced energy levels that form energy bands. This process is illustrated in Fig. 14.1-9, where electron energy levels are illustrated schematically for two isolated atoms (a); for a molecule containing two such atoms (b); and for a rudimentary 1D lattice comprising five such atoms (c). The lowest-lying energy levels remain sharp because the electrons in the inner subshells are shielded from the influence of nearby atoms, but the initially sharp energy levels associated with the outer atomic electrons become bands as the atoms enter into close proximity and degeneracies are removed by Stark splitting.

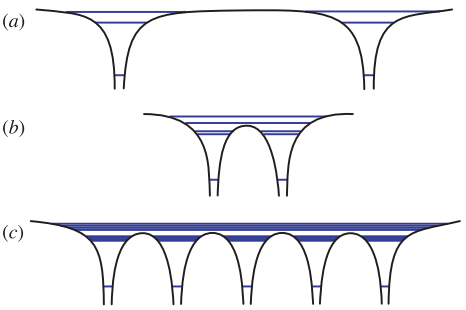


Figure 14.1-9 Schematic energy levels for: (a) two isolated atoms; (b) the same two atoms after having been brought into close contact and forming a diatomic molecule; and (c) five identical atoms in close proximity having formed a rudimentary 1D crystal.

This picture is elaborated in Fig. 14.1-10, where we schematically compare the energy levels of an isolated atom and three different kinds of solids that comprise lattices of such atoms: a metal, a semiconductor, and an insulator. The lowest-lying energy levels of these solids, denoted in this example by the electron configurations $1s$, $2s$, and $2p$, resemble those of the isolated atom because the inner electrons are shielded from interatomic forces. In contrast, the discrete higher energy levels of the atomic valence electrons, denoted $3s$ and $3p$ here, are split into densely packed energy bands in the solids. The lowest-lying unoccupied, or partially occupied, energy band is called the **conduction band** while the highest-lying fully occupied energy band is known as the **valence band**. These two bands are separated by a **forbidden band**, with an energy extent E_g known as the **bandgap energy**. As with electrons in individual atoms, the Pauli exclusion principle applies to the electrons in solids so that the lowest-lying energy bands are occupied first.

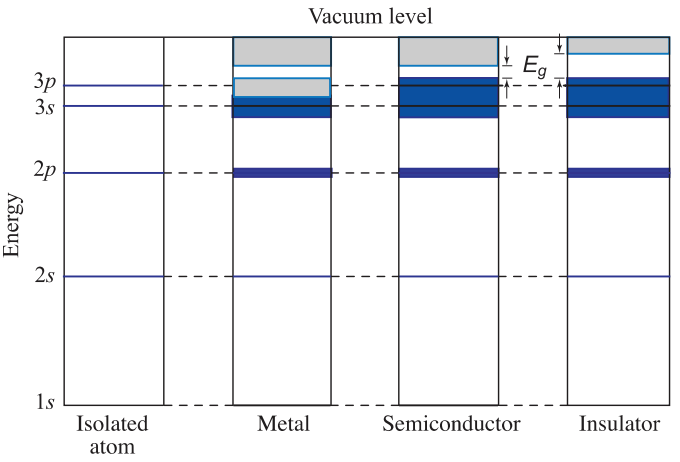


Figure 14.1-10 Broadening of the discrete energy levels of an isolated atom into energy bands when atoms in close proximity form a solid. Fully occupied bands are darkly shaded, unoccupied bands are lightly shaded, and partially occupied bands are both lightly and darkly shaded. The forbidden band is shown as white. Typical values for the room-temperature conductivity σ for metals, semiconductors, and insulators are 10^{8} $(\Omega\text{-m})^{-1}$, 10^{-4}–10^{5} $(\Omega\text{-m})^{-1}$, and 10^{-10} $(\Omega\text{-m})^{-1}$, respectively.

Metals, semiconductors, and insulators. Metals comprise the greatest preponderance of elements in the periodic table (Fig. 14.1-3). They have a partially occupied conduction band at all temperatures (lightly and darkly shaded region in Fig. 14.1-10). The availability of many unoccupied states in this band is responsible for their high electrical conductivity (Sec. 8.2A). Semimetals, in contrast, have overlapping valence and conduction bands.

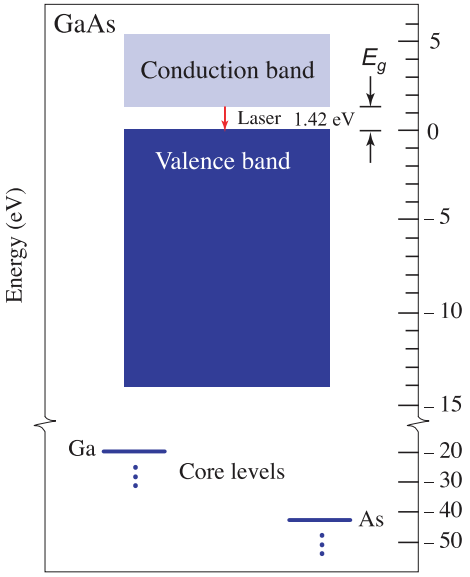
Intrinsic semiconductors have an occupied valence band (dark shading in Fig. 14.1-10) and an unoccupied conduction band (light shading) at $T = 0°$ K. Since there are no available free states in the valence band, and no electrons in the conduction band, the conductivity of an ideal intrinsic semiconductor at $T = 0°$ K is zero. As the temperature of the semiconductor rises above absolute zero, however, an increasing number of electrons from the valence band gain sufficient thermal energy to enter the conduction band and thereby to contribute to the conductivity of the material.

Insulators also have a fully occupied valence band (dark shading in Fig. 14.1-10) and an unoccupied conduction band (light shading). They are distinguished from semiconductors by their larger bandgap energy, which is typically greater than about 3 eV. As an example, the bandgap energy for silicon (a semiconductor) is $E_g \approx 1.1$ eV whereas that for diamond (an insulator) is $E_g \approx 5.5$ eV. Above absolute zero, fewer electrons in insulators have the requisite thermal energy to surmount the bandgap energy and contribute to the conductivity of the material. It should be pointed out, however, that the degree of band overlap also plays a role in determining whether a material is a metal, semiconductor, or insulator.

Semiconductors

Semiconductors find widespread use in photonics. They serve as sources such as light-emitting diodes and laser diodes, waveguides, modulators, switches, and detectors, and play many other important roles as well. We proceed to provide a brief introduction to the energy levels of inorganic bulk semiconductors, quantum wells, quantum wires, and quantum dots. A more extensive exposition relating to the properties of semiconductors, including organic semiconductors, is provided in Sec. 17.1.

Bulk semiconductors. The binary semiconductor GaAs was early on found to be useful in photonics. This material takes the form of a zincblende structure comprising two face-centered-cubic lattices, one of Ga atoms and the other of As atoms, displaced from each other by $1/4$ the length of a body diagonal (Fig. 14.1-11). Four molecules of GaAs are present in the conventional cell, which is a cube. Each atom is surrounded by four atoms of the opposite type, equally spaced and located at the corners of a regular tetrahedron.



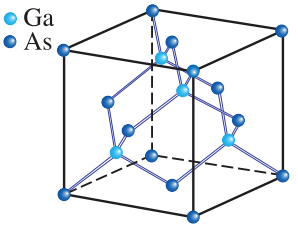


Figure 14.1-11 The semiconductor GaAs takes the form of a zincblende crystal structure comprising two face-centered-cubic lattices, one of Ga and the other of As. The higher energy levels are closely spaced and form bands. The zero of energy is (arbitrarily) defined at the top edge of the valence band. The GaAs laser diode operates on the electron transition between the conduction and valence bands, in the near-infrared region of the spectrum.

Semiconductors have many closely spaced allowed electron energy levels that take the form of bands, as displayed in Fig. 14.1-11 for GaAs. The bandgap energy E_g, which is the energy separating the valence and conduction bands, is 1.42 eV at room temperature. The Ga and As ($3d$) core levels are quite sharp, as displayed in Fig. 14.1-11. The valence band of GaAs is formed from the $4s$ and $4p$ levels (in analogy with the schematic in Fig. 14.1-10).

Quantum wells. Crystal-growth techniques such as molecular-beam epitaxy and vapor-phase epitaxy can be used to grow materials with specially designed band structures. In semiconductor **quantum-well structures**, the energy bandgap is engineered to vary with position in a specified manner, leading to materials with unique electronic and optical properties. An example is the **multiquantum-well structure** illustrated in Fig. 14.1-12. This particular structure comprises ultrathin (2- to 15-nm-thick) layers of GaAs alternating with thin (20-nm-thick) layers of AlGaAs. The bandgap of the GaAs is smaller than that of the AlGaAs. For motion perpendicular to the layer, the allowed energy levels for electrons in the conduction band, and for holes in the valence band, are discrete and well separated, like those of the square-well potential in quantum mechanics (Exercise 17.1-5); the lowest energy levels are shown schematically in each of the quantum wells in Fig. 14.1-12.

The AlGaAs barrier regions can also be made ultrathin (< 1 nm), in which case the electrons in adjacent wells can readily couple to each other via quantum-mechanical tunneling, whereupon the discrete energy levels broaden into miniature bands called **minibands**. The material is then called a **superlattice structure** because the associated

lattice is "super to" (i.e., larger than) that of the atomic crystal lattice, which gives rise to minibands rather than the natural full-size energy bands associated with the atomic lattice. Quantum wells are discussed further in Secs. 17.1G, 18.2D, and 18.4A.

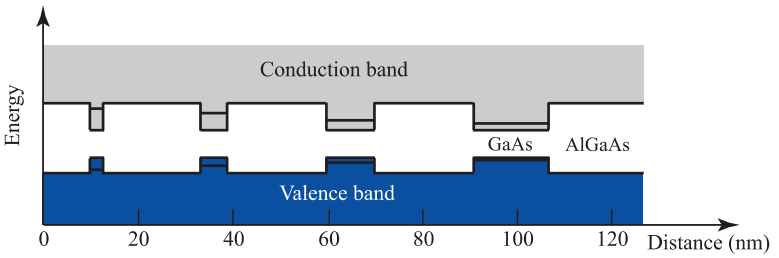


Figure 14.1-12 Quantized energy levels in a single-crystal AlGaAs/GaAs multiquantum-well structure. The well widths can be periodic or arbitrary (as shown).

Quantum wires. A semiconductor material that takes the form of a thin wire surrounded by a material of wider bandgap is known as a **quantum wire**. The wire acts as a potential well that narrowly confines electrons (and holes) in the two lateral directions but not in the direction along the axis of the wire. Quantum wires are readily made from III–V and II–VI semiconductors such as InP and CdSe, respectively; they usually have rectangular or circular cross section. Nanotubes and nanowires fabricated from a broad variety of materials can behave as quantum wires. **Carbon nanotubes** are cylindrical carbon molecules with diameters of one or a few nm in which the carbon molecules organize themselves into thin hollow ropes held together by van der Waals forces. Single- or multiwalled nanotubes exhibit unique optical, mechanical, and electrical properties. They can behave as semiconductors or highly conductive metals, depending on the details of their structure. There are a multitude of uses for carbon nanotubes in photonics, ranging from filaments for incandescent light sources to photovoltaic detectors. Quantum wires are discussed further in Secs. 17.1G and 18.4B.

Quantum dots. Also known as **nanocrystals** or quantum boxes, **quantum dots** are semiconductor particles whose dimensions typically fall in the range of 1 to 50 nm. Quantum dots can be fabricated from many different kinds of semiconductors and in many geometrical shapes (e.g., cubes, spheres, hemispheres, disks, and pyramids), depending on the growth conditions. They are often embedded in semiconductor materials with larger bandgaps or in glasses or polymers. Robust techniques for growing quantum dots continue to be developed. When fabricated using molecular-beam epitaxy (MBE) or chemical-vapor deposition (CVD), they can assume the form of disk-shaped structures, in which the electron motion is restricted to a plane and is characterized by 2D atomic-like shell structures not unlike those associated with the toroidal resonators considered in Sec. 11.4B. Quantum dots can also be created via electron-beam lithography; a pattern is etched onto a semiconductor chip and conducting metal is deposited onto the pattern. Quantum dots can also be readily grown in a beaker using wet chemistry. Self-assembled quantum dots, with typical dimensions in the range of 10–50 nm, are formed from colloidal nanocrystals provided in liquid suspension or dispersed in a plastic composite. Chemical synthesis yields near-perfect crystalline clusters that range from several hundred to several tens-of-thousands of atoms and assume various shapes, again depending on the growth conditions. They can be deposited onto substrates or incorporated directly into devices designed to accommodate them. Self-assembly can also be achieved by means of epitaxial synthesis, which can yield

strained quantum-dot layers designed to improve device characteristics. Arrays and self-assembled stacks of quantum dots are readily fabricated.

The sizes of quantum dots, and thus the number of atoms they contain, varies over a broad range; a 10-nm cube of GaAs contains some 40 000 atoms. All electrons belong to the dot as a whole; the number of electrons can be as small as just a few or as large as millions. The energy levels of a quantum dot are those of its excitons, namely the electron–hole pairs generated within, and confined to, the dot. As with atoms, a series of sharp energy levels results from tight electron confinement; quantum dots are in fact often called **artificial atoms**. Unlike atoms, however, a quantum dot fabricated from a given material has the property that its energy levels are strongly dependent on its size. Much as with the energy levels of an electron in a quantum well (Exercise 17.1-5), tight confinement in a small quantum dot corresponds to large energy-level differences and short transition wavelengths. The wavelength of an emitted photon consequently decreases along with the size of the quantum dot. The color of light elicited from a CdSe quantum dot by photoexcitation, for example, can be gradually tuned from the red region of the spectrum for a 5-nm-diameter dot to the violet region for a 1.5-nm-diameter dot; the trend is illustrated in Fig. 14.1-13. The photoexcitation wavelength is arbitrary, as long as it is shorter than the emission wavelength. Quantum dots fabricated from InP luminesce in the near infrared, whereas those fabricated from InAs emit across the 1300–1600-nm silica-fiber-based telecommunications band. Photoexcited Si quantum dots also emit over a broad spectral range that extends from the infrared to the visible (Example 17.2-2). Quantum dots can also be fabricated from organic compounds.

Figure 14.1-13 Photoluminescence from colloidal CdSe quantum dots (with oleylamine surface capping molecules) dispersed in n-hexane, in response to ultraviolet excitation at $\lambda_o = 365$ nm. Quantum-confinement effects allow the emission color to be tuned with quantum-dot size (courtesy Dong-Kyun Seo, Arizona State University).

Quantum dots overcoated with a semiconductor material of higher-bandgap are known as *core–shell quantum dots*, whereas those overcoated with multiple semiconductors of alternating higher and lower bandgaps are known as *quantum-well–quantum dots*. Such overcoatings can substantially improve the tunability and photoluminescence efficiency of the nanostructure. Ordered arrangements of quantum dots, known as **quantum-dot solids**, can be grown by a number of methods, including the self-assembly of nanocrystals into a close-packed configuration. In the same way that tunneling can occur in multiquantum-well superlattices, so too can it occur in quantum-dot solids known as **nanocrystal superlattices**.

When quantum-dot structures are brought into contact with electrodes, they can serve as miniature photonic devices. By constructing arrays of quantum dots of different sizes in specially designed configurations, they can sustain currents and operate over broad, or specially chosen, wavelength ranges. Quantum dots are useful as spectral tags in biological, commercial, and military applications. As discussed in Chapters 18 and 19, they also find use in a broad array of photonic devices such as light-emitting diodes, semiconductor optical amplifiers, laser diodes, single-photon sources, memory elements, photodetectors, solar cells, flat-panel displays, backlighting sources, and as absorbers in materials where it is desirable to filter out ultraviolet light. Quantum dots are discussed further in Secs. 17.1G and 18.4C.

14.2 OCCUPATION OF ENERGY LEVELS

As indicated earlier, each atom or molecule in a collection continuously undergoes random transitions among its different energy levels. These transitions are characterized by the rules of statistical physics. Temperature is the principal determinant of both the average behavior and the fluctuations in energy-level occupancy.

A. Boltzmann Distribution

Consider a collection of distinguishable objects, such as atoms or molecules that form a dilute gas. Each atom is in one of its allowed energy levels $E_1, E_2, \ldots$. If the system is in thermal equilibrium at temperature T (i.e., if the atoms are kept in contact with a thermal bath maintained at temperature T and their motion reaches a steady state in which the fluctuations are, on average, invariant to time), the probability $P(E_m)$ that an arbitrary atom is in energy level E_m is given by the Boltzmann distribution

$$P(E_m) \propto \exp(-E_m/kT), \qquad m = 1, 2, 3, \ldots, \tag{14.2-1}$$

where k is the Boltzmann constant. The coefficient of proportionality is chosen such that $\sum_m P(E_m) = 1$. The occupation probability $P(E_m)$ vs. E_m is an exponentially decreasing function of E_m, as displayed in Fig. 14.2-1.

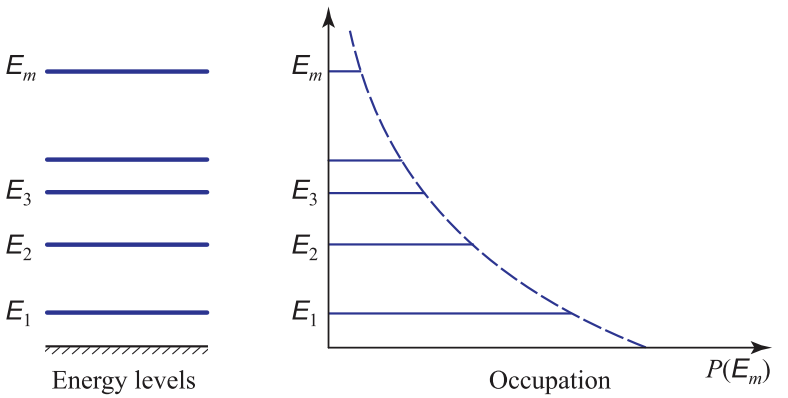


Figure 14.2-1 The Boltzmann distribution $P(E_m)$, plotted on the abscissa, specifies the probability that energy level E_m of an arbitrary atom is occupied; it is an exponentially decreasing function of E_m.

The origin of Boltzmann distribution can be understood by considering a system of many identical entities that share a fixed total energy E. The entities are isolated from their surroundings but are in thermal equilibrium, exchanging energy among themselves via a bath at temperature T. The divisions of energy are taken to be distinguishable if they involve different energy states, and all possible divisions of the total energy are assumed to occur with equal probability. If one of the entities takes a large share of the total energy, less is available for the remaining constituents so there are fewer possible divisions. Consequently, large energies are less probable than small energies. A quantitative description is provided by considering two entities. The probability of finding one with energy E_1 and the other with energy E_2 is the product $P(E_1)P(E_2)$ because they are independent. If the sum of the energies of the two entities is fixed at the value $E_1 + E_2$, then $P(E_1)P(E_2)$ must be a function of $(E_1 + E_2)$, which uniquely specifies an exponential function. The equipartition energy

kT for the two degrees of freedom associated with a harmonic mode leads directly to the Boltzmann distribution.

Consider the Boltzmann distribution in the context of a large number of atoms N. If N_m is the number of atoms occupying energy level E_m, the fraction $N_m/N \approx P(E_m)$. If N_1 atoms occupy level 1 and N_2 atoms occupy a higher level 2, the population ratio is, on average,

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{kT}\right). \tag{14.2-2}$$

This quantity depends on the temperature T. At $T = 0°$ K, all atoms are in the lowest energy level (ground state). As the temperature increases the populations of the higher energy levels grow. Under equilibrium conditions, the average population of a given energy level is always greater than that of a higher-lying level. This condition need not hold under non-equilibrium conditions, however, in which case a higher energy level can have a greater average population than a lower energy level. This latter state of affairs, known as a **population inversion**, provides the basis for laser action (Sec. 15.1A).

It was assumed in the foregoing that there is a unique way in which an atom can find itself in one of its energy levels. It is sometimes the case, however, that two or more states (e.g., different states of angular momentum) correspond to the same energy. To account for such degenerate states, (14.2-2) can be written in the more general form

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} \exp\left(-\frac{E_2 - E_1}{kT}\right), \tag{14.2-3}$$

where the degeneracy parameters g_2 and g_1 represent the numbers of states corresponding to the energy levels E_2 and E_1, respectively.

B. Fermi–Dirac Distribution

Fermions subject to the Pauli exclusion principle, such as electrons with overlapping wavefunctions in a multielectron atom or in a semiconductor, obey Fermi–Dirac statistics. The probability of occupancy of a state of energy E is then described by the **Fermi–Dirac distribution** (or **Fermi function**),

$$f(E) = \frac{1}{\exp\left[(E - E_f)/kT\right] + 1}, \tag{14.2-4}$$

where E_f is called the **Fermi energy** (Fig. 14.2-2).

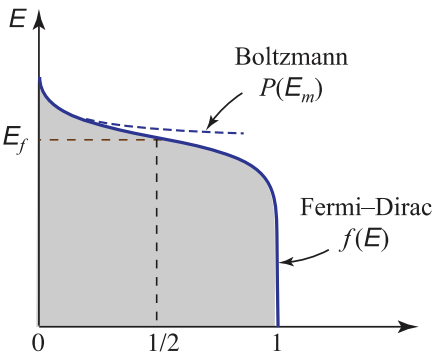


Figure 14.2-2 The Fermi–Dirac distribution $f(E)$, plotted on the abscissa, represents the probability of occupancy of a state of energy E. This distribution is applicable for a system containing particles with overlapping wavefunctions in which the Pauli exclusion principle applies. The Fermi–Dirac distribution is well approximated by the Boltzmann distribution $P(E_m)$ for $E \gg E_f$, where the occupancy probability is low.

The occupancy probability decreases monotonically with increasing E, and falls to a value of $1/2$ at the Fermi energy $E = E_f$. States for which $f(E) = 1$ are definitely occupied. It is important to recognize that the Fermi–Dirac distribution $f(E)$ is a sequence of probabilities with values lying between 0 and 1 for all values of E, rather than a probability density function. However, for $E \gg E_f$, the occupancy probability is low and (14.2-4) reveals that the Fermi–Dirac distribution then reduces to the Boltzmann probability distribution

$$P(E) \propto \exp(-E/kT), \tag{14.2-5}$$

as illustrated in Fig. 14.2-2. This condition is generally applicable for valence electrons in the outer subshells of atoms and ions so that optically active electrons are populated in accordance with the Boltzmann distribution. The Fermi function is considered in more detail in Chapter 17.

Bosons such as photons (Sec. 13.2C) and atoms in a Bose–Einstein condensate (Sec. 14.3F) obey the Bose–Einstein distribution.

14.3 INTERACTIONS OF PHOTONS WITH ATOMS

A. Interaction of Single-Mode Light with an Atom

An atom may emit (create) or absorb (annihilate) a photon by undergoing a downward or upward transition between pairs of its energy levels, while conserving energy in the process. This section is devoted to describing the laws that govern such emissions and absorptions. The interactions of photons with electrons and holes in semiconductors is considered in Sec. 17.2.

Interaction Between an Atom and an Electromagnetic Mode

Consider the energy levels E_1 and E_2 of an atom placed in an optical resonator of volume V that can sustain a number of electromagnetic modes. We are particularly interested in the interaction between the atom and the photons of a *prescribed* radiation mode of frequency $\nu \approx \nu_0$, where $h\nu_0 = E_2 - E_1$, since photons of this energy match the atomic energy-level difference. A formal study of such interactions relies on quantum electrodynamics; we present the key results that emerge from such an analysis below, without proof.

Three forms of interaction are possible — spontaneous emission, absorption, and stimulated emission, which we consider in turn.

Spontaneous Emission

If the atom is initially in the upper energy level, it may decay spontaneously to the lower energy level and release its energy in the form of a photon (Fig. 14.3-1). The photon energy $h\nu$ is added to the energy of the electromagnetic mode. The process is called **spontaneous emission** because the transition is independent of the number of photons that may already be in the mode.

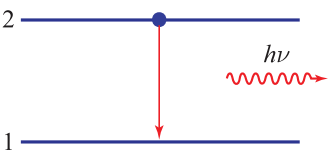


Figure 14.3-1 Spontaneous emission of a photon into the mode of frequency ν by an atomic transition from energy level 2 to energy level 1. The photon energy $h\nu \approx E_2 - E_1$.

In a cavity of volume V, the probability density (per second), or rate, for this spontaneous transition depends on ν in a way that characterizes that atomic transition,

$$p_{\text{sp}} = \frac{c}{V}\,\sigma(\nu). \tag{14.3-1}$$

(14.3-1) Spontaneous Emission into a Prescribed Mode

The quantity $\sigma(\nu)$, known as the **transition cross section**, is a function of ν centered about the atomic resonance frequency ν_0. The significance of this designation will become apparent subsequently, but it is clear that σ has dimensions of cm^2 (since the dimensions of p_{sp}, c, and V are s^{-1}, cm/s, and cm^3, respectively). In principle, $\sigma(\nu)$ can be determined from the Schrödinger equation but the calculations are generally sufficiently complex that it is usually determined empirically. Equation (14.3-1) applies separately to every mode, with a transition cross section σ that depends on the angle θ between the dipole moment of the atom and the field direction of the mode, in accordance with

$$\sigma = \sigma_{\text{max}} \cos^2\theta. \tag{14.3-2}$$

The maximum cross section σ_{max} is attained when the dipole moment and field align.

The term "probability density" signifies that the probability of an emission taking place in an incremental time interval between t and $t + \Delta t$ is simply $p_{\text{sp}}\,\Delta t$. Because it is a probability density, p_{sp} can have a numerical value greater than 1 s^{-1}, although of course $p_{\text{sp}}\Delta t$ must always be smaller than 1. Thus, if there are a large number N of such atoms, a fraction of approximately $\Delta N = (p_{\text{sp}}\,\Delta t)N$ atoms will undergo this transition within the time interval Δt. Consequently, we can write $dN/dt = -p_{\text{sp}}\,N$, indicating that the number of atoms $N(t) = N(0)\exp(-p_{\text{sp}}t)$ decays exponentially with time constant $1/p_{\text{sp}}$, as illustrated in Fig. 14.3-2.

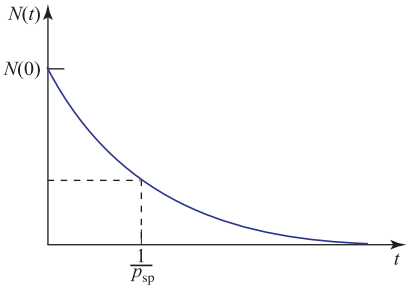


Figure 14.3-2 Spontaneous emission into a single mode results in an exponential decrease of the number of excited atoms, with time constant $1/p_{\text{sp}}$.

Absorption

If the atom is initially in the lower energy level and the radiation mode contains a photon, the photon may be annihilated and the atom concomitantly raised to the upper energy level (Fig. 14.3-3). This process, which is *induced* by the photon, is called **absorption**. It can occur only when the mode contains a photon.

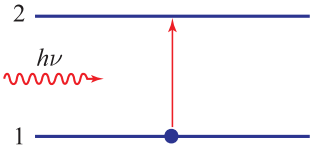


Figure 14.3-3 Absorption is a process whereby a photon of energy $h\nu$ induces the atom to undergo an upward transition from level 1 to level 2.

The probability density for the absorption of a photon from a given mode of frequency ν, in a cavity of volume V, is governed by the *same law* that governs spontaneous emission into that mode, namely

$$p_{\text{ab}} = \frac{c}{V}\sigma(\nu). \tag{14.3-3}$$

However, if there are n photons in the mode, the probability density that the atom absorbs *one* photon is n times greater since the events are mutually exclusive, i.e.,

$$\boxed{P_{\text{ab}} = n\frac{c}{V}\sigma(\nu).} \tag{14.3-4}$$

Absorption of One Photon from a Mode with n Photons

Stimulated Emission

Finally, if the atom is in the upper energy level and the mode contains a photon, the atom may be *induced* to emit another photon into the same mode. This process, known as **stimulated emission**, is the inverse of absorption. The presence of a photon in a mode of specified frequency, propagation direction, and polarization stimulates the emission of a duplicate ("clone") photon with precisely the same characteristics as the original (Fig. 14.3-4). This photon amplification process is the phenomenon that underlies the operation of laser amplifiers and lasers, as will be elucidated in subsequent chapters.

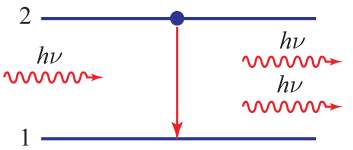


Figure 14.3-4 Stimulated emission is a process whereby a photon of energy $h\nu$ induces the atom to emit a clone photon as it undergoes a downward transition from level 2 to level 1.

The probability density p_{st} that this process occurs in a cavity of volume V is governed by the *same law* that governs spontaneous emission and absorption:

$$p_{\text{st}} = \frac{c}{V}\sigma(\nu). \tag{14.3-5}$$

If the mode originally carries n photons, the probability density that the atom is stimulated to emit an additional photon is, just as in the case of absorption,

$$\boxed{P_{\text{st}} = n\frac{c}{V}\sigma(\nu).} \tag{14.3-6}$$

Stimulated Emission of One Photon into a Mode with n Photons

Under certain circumstances, as will be elucidated in Sec. 15.2B, the effective transition cross sections $\sigma(\nu)$ specified in (14.3-4) and (14.3-6) for absorption and stimulated emission, respectively, differ; they are then denoted $\sigma_{\text{ab}}(\nu)$ and $\sigma_{\text{em}}(\nu)$, respectively, in which case $P_{\text{st}} \neq P_{\text{ab}}$. When the two cross sections are equal, however, we use a common notation for both probability densities: $W_i \equiv P_{\text{st}} = P_{\text{ab}}$.

Inasmuch as spontaneous emission is present in addition to stimulated emission, combining (14.3-1) and (14.3-6) leads to an overall probability density of the atom emitting a photon into the mode that is described by $p_{\text{sp}} + P_{\text{st}} = (n+1)(c/V)\sigma(\nu)$.

From a quantum-electrodynamic point of view, spontaneous emission may be regarded as stimulated emission induced by the zero-point fluctuations associated with the mode (Sec. 13.1A). Because the zero-point energy plays no role in absorption, however, P_{ab} is proportional to n rather than to $(n + 1)$.

The three possible interactions between an atom and a radiation mode in a cavity (spontaneous emission, absorption, and stimulated emission) obey the fundamental relations set forth above. These formulas should be regarded as the rules that govern the interactions between photons and atoms in the context of laser physics. They are to be used in conjunction with the rules of photon optics set forth in Chapter 13. We now proceed to discuss some of the consequences of these rather simple relations.

Lineshape Function and Transition Strength

It is clear from the foregoing that the transition cross section $\sigma(\nu)$ characterizes the interaction of the atom with the radiation mode. Its shape governs the relative magnitude of the interaction of the atom with photons over a range of frequencies, while its area,

$$S = \int_0^\infty \sigma(\nu)\, d\nu, \tag{14.3-7}$$

known as the **transition strength** or **oscillator strength**, represents the strength of the interaction. The area S, which has units of cm^2-Hz, can be readily separated from the shape (profile) of $\sigma(\nu)$ by defining a normalized **lineshape function** $g(\nu) = \sigma(\nu)/S$, which has unity area, $\int_0^\infty g(\nu)\, d\nu = 1$, and units of Hz^{-1}. The transition cross section can then be written in terms of its strength and profile as

$$\sigma(\nu) = Sg(\nu). \tag{14.3-8}$$

The lineshape function $g(\nu)$ is centered about the resonance frequency ν_0, where $\sigma(\nu)$ is largest, and decreases sharply as ν deviates from ν_0. Transitions are therefore most likely to occur for photons of frequency $\nu \approx \nu_0$. The width of the function $g(\nu)$ is known as the **transition linewidth** $\Delta\nu$, which is usually defined as the full-width at half-maximum (FWHM) value of $g(\nu)$ (Sec. A.2 of Appendix A). Since the area of $g(\nu)$ is unity, its width is inversely proportional to its central value,

$$\Delta\nu \propto 1/g(\nu_0). \tag{14.3-9}$$

It is also useful to define a **peak cross section** at the resonance frequency: $\sigma_0 \equiv \sigma(\nu_0)$. As illustrated in Fig. 14.3-5, the transition cross section $\sigma(\nu)$ is then characterized by four features: 1) its height σ_0; 2) its width $\Delta\nu$; 3) its area S; and 4) its profile $g(\nu)$.

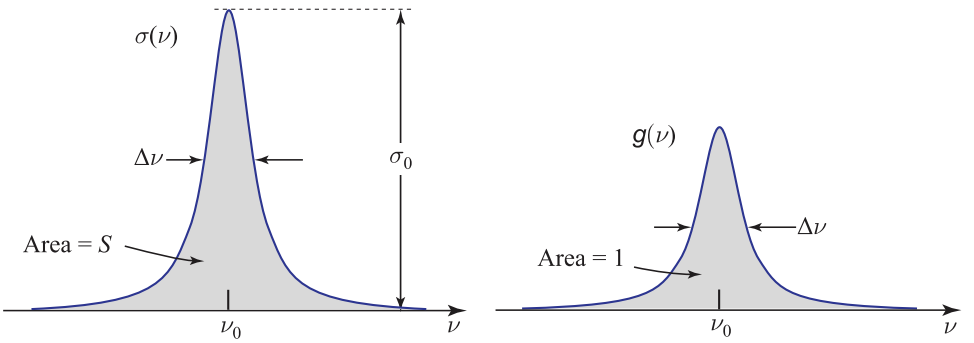


Figure 14.3-5 The transition cross section $\sigma(\nu)$ and the lineshape function $g(\nu)$.

B. Spontaneous Emission

Total Spontaneous Emission into All Modes

Equation (14.3-1) provides the probability density p_{sp} for spontaneous emission into a *prescribed* mode of frequency ν, without regard to whether the mode contains photons. As indicated in (11.3-10), the density of modes for a three-dimensional cavity increases quadratically with frequency as $M(\nu) = 8\pi\nu^2/c^3$. This quantity approximates the number of modes of frequency ν, per unit volume of the cavity per unit bandwidth, provided that the number of modes is sufficiently large so that a continuous approximation is suitable for counting the modes. An atom may spontaneously emit *one* photon of frequency ν into *any* of these modes, as shown schematically in Fig. 14.3-6.

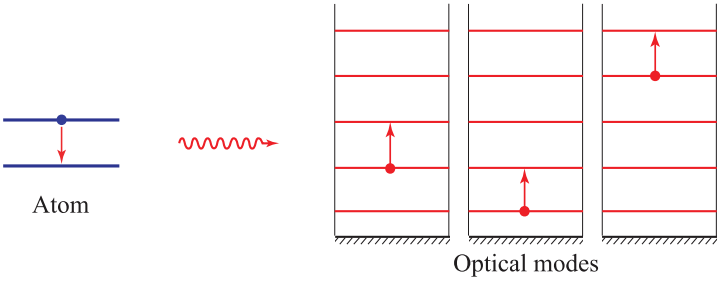


Figure 14.3-6 An atom may spontaneously emit a photon into any one (but only one) of the many optical modes with frequencies $\nu \approx \nu_0$.

The probability density for spontaneous emission into any available mode is therefore given by the probability density for spontaneous emission into a specific mode, weighted by the modal density. Since modes at each frequency have an isotropic distribution of directions, each with two polarizations, we must make use of the average transition cross section $\overline{\sigma}(\nu)$. If θ is the angle between the dipole moment of the atom and the field direction, (14.3-2) provides

$$\overline{\sigma}(\nu) = \tfrac{1}{3}\sigma_{\max} \tag{14.3-10}$$

since $\langle \cos^2\theta \rangle = \frac{1}{3}$, where $\langle \cdot \rangle$ represents an average in 3D space. The overall spontaneous-emission probability density therefore becomes

$$P_{\text{sp}} = \int_0^\infty \left[\frac{c}{V}\,\overline{\sigma}(\nu)\right][V\,M(\nu)]\,d\nu = c\int_0^\infty \overline{\sigma}(\nu)\,M(\nu)\,d\nu. \tag{14.3-11}$$

Because the function $\overline{\sigma}(\nu)$ is sharply peaked, it is narrow in comparison with the quadratic function $M(\nu) = 8\pi\nu^2/c^3$. Since $\overline{\sigma}(\nu)$ is centered about ν_0, $M(\nu)$ is approximately constant with a value $M(\nu_0)$, and can thus be removed from the integral. The probability density of spontaneous emission of one photon into *any* mode is therefore

$$P_{\text{sp}} = M(\nu_0)\,c\overline{S} = \frac{8\pi\nu_0^2\overline{S}}{c^2} = \frac{8\pi\overline{S}}{\lambda^2}, \tag{14.3-12}$$

where $\lambda = c/\nu_0$ is the wavelength of the light in the medium and $\overline{S} = \int_0^\infty \overline{\sigma}(\nu)\,d\nu$. We define a time constant t_{sp}, known as the **spontaneous lifetime** of the $2 \rightarrow 1$ transition,

such that $1/t_{\rm sp} \equiv P_{\rm sp}$, so

$$\boxed{P_{\rm sp} = \frac{1}{t_{\rm sp}}\,,} \qquad (14.3\text{-}13)$$

Spontaneous Emission of One Photon into Any Mode

which is independent of the cavity volume V. Using this, (14.3-12) provides

$$\overline{S} = \frac{\lambda^2}{8\pi t_{\rm sp}}\,, \qquad (14.3\text{-}14)$$

which enables the transition strength to be determined from an empirical measurement of the spontaneous lifetime $t_{\rm sp}$. For a simple transition, such as between the first excited state and the ground state of atomic hydrogen, we find that $t_{\rm sp} \approx 10^{-8}$ s; however, $t_{\rm sp}$ can vary over a very broad range, from femtoseconds to seconds (see, e.g., Table 15.3-1).

Equation (14.3-14) is useful because a first-principles calculation of $\overline{S}$ would require intimate knowledge about the quantum-mechanical behavior of the system, which is not always available.

EXERCISE 14.3-1

Frequency of Spontaneously Emitted Photons. Show that the probability density for an excited atom spontaneously emitting a photon of frequency between ν and $\nu + d\nu$ is $P_{\rm sp}(\nu)\,d\nu = (1/t_{\rm sp})g(\nu)\,d\nu$. Explain why the spectrum of spontaneous emission from an atom is proportional to its lineshape function $g(\nu)$ after a large number of photons have been emitted.

Relation Between Transition Cross Section and Spontaneous Lifetime

Using (14.3-14) together with the relation $\overline{\sigma}(\nu) = \overline{S}g(\nu)$ shows that the average transition cross section is related to the spontaneous lifetime and the lineshape function via

$$\boxed{\overline{\sigma}(\nu) = \frac{\lambda^2}{8\pi t_{\rm sp}}\, g(\nu),} \qquad (14.3\text{-}15)$$

Average Transition Cross Section

a relationship known as the Füchtbauer–Ladenburg equation. The average transition cross section at the central frequency ν_0 is therefore

$$\overline{\sigma}_0 \equiv \overline{\sigma}(\nu_0) = \frac{\lambda^2}{8\pi t_{\rm sp}}\, g(\nu_0). \qquad (14.3\text{-}16)$$

Because $g(\nu_0)$ is inversely proportional to $\Delta\nu$, for a given value of $t_{\rm sp}$ the peak transition cross section $\overline{\sigma}_0$ is inversely proportional to the linewidth $\Delta\nu$, in accordance with (14.3-9).

C. Stimulated Emission and Absorption

Transitions Induced by Monochromatic Light

We now consider the interaction of single-mode light with an atom when a stream of photons impinges on it, rather than when it resides in a resonator of volume V as considered above. Let monochromatic light of frequency ν, intensity I, and mean photon-flux density (photons/cm^2-s)

$$\phi = I/h\nu \tag{14.3-17}$$

interact with an atom whose resonance frequency is ν_0. We wish to determine the probability densities for stimulated emission and absorption, $W_i \equiv P_{\mathrm{st}} = P_{\mathrm{ab}}$, in this configuration.

The number of photons n involved in the interaction process is determined by constructing a volume in the form of a cylinder of base area A, height $c \times 1$ s, and volume $V = cA$. The axis of the cylinder is parallel to $\mathbf{k}$, the direction of propagation of the light. The photon flux that crosses the cylinder base is $\Phi = \phi A$ (photons/s). Because photons travel at the speed of light c, all of the photons within the volume of the cylinder cross its base within one second. It follows that, at any time, the cylinder contains $n = \phi A = \phi V/c$ photons so that

$$\phi = n\frac{c}{V}. \tag{14.3-18}$$

To determine W_i, we substitute (14.3-18) into (14.3-4) or (14.3-6) to obtain

$$\boxed{W_i = \phi\,\sigma(\nu).} \tag{14.3-19}$$

It is apparent that $\sigma(\nu)$ is the coefficient of proportionality between the probability density of an induced transition and the photon-flux density. This relationship informs us that the appellation "transition cross section" is apt: ϕ is the photon-flux density ($\mathrm{cm}^{-2}\cdot\mathrm{s}^{-1}$) while $\sigma(\nu)$ is the effective cross-sectional area of the atom (cm^2), so that $\phi\,\sigma(\nu)$ represents the probability density (s^{-1}) that a photon in the stream is "captured" by the "cross section" of the atom for the purpose of stimulated emission or absorption.

It is clear from (14.3-4), (14.3-6), and (14.3-11) that the probability densities for absorption, stimulated emission, and spontaneous emission are all proportional to $\sigma(\nu)$. As discussed above, stimulated emission involves decay only into those modes that contain photons. Though the expression for $\overline{\sigma}(\nu)$ set forth in (14.3-15) was obtained for spontaneous emission into multiple modes, it is convenient to make use of it in conjunction with (14.3-19) to determine the probability density for induced transitions as well, since t_{sp} is readily determined experimentally.

The use of the quantity $\overline{\sigma}(\nu)$ instead of $\sigma(\nu)$ in (14.3-15) is a result of averaging over the angle between the dipole moment of the atom and the field direction [see (14.3-2) and (14.3-10)]; it is appropriate for spontaneous emission into all modes. However, when such averaging is not called for, such as in the case of stimulated emission into a particular mode and a fixed angle θ, $\sigma(\nu)$ and σ_0 are used in place of $\overline{\sigma}(\nu)$ and $\overline{\sigma}_0$. Any required change in $\sigma(\nu)$ associated with averaging for a particular induced-transition configuration can be readily accommodated by modifying t_{sp}, which is then referred to as the **effective spontaneous lifetime**. For simplicity, we shall henceforth not distinguish between t_{sp} for spontaneous emission and its effective value for stimulated emission.

Transitions Induced by Broadband Light

Consider now an atom in a cavity of volume V containing multimode polychromatic light of spectral energy density $\varrho(\nu)$ (energy per unit bandwidth per unit volume) that is broadband in comparison with the atomic linewidth. The average number of photons in the frequency band from ν to $\nu + d\nu$ is $[\varrho(\nu)V/h\nu]\, d\nu$; each of these has a probability density $(c/V)\sigma(\nu)$ of initiating an atomic transition. As with spontaneous emission, the modes at each frequency are taken to be isotropically distributed in direction, each with two polarizations, so that the overall probability of absorption or stimulated emission is

$$W_i = \int_0^\infty \frac{\varrho(\nu)V}{h\nu} \left[\frac{c}{V}\,\overline{\sigma}(\nu)\right] d\nu. \tag{14.3-20}$$

Since the radiation is broadband, the function $\varrho(\nu)$ varies slowly in comparison with the sharply peaked transition cross section $\overline{\sigma}(\nu)$. We can therefore replace $\varrho(\nu)/h\nu$ under the integral with $\varrho(\nu_0)/h\nu_0$, which leads to

$$W_i = \frac{\varrho(\nu_0)}{h\nu_0}\, c \int_0^\infty \overline{\sigma}(\nu)\, d\nu = \frac{\varrho(\nu_0)}{h\nu_0}\, c\overline{S}. \tag{14.3-21}$$

Using (14.3-14), we therefore have

$$W_i = \frac{\lambda^3}{8\pi h t_{\text{sp}}}\, \varrho(\nu_0), \tag{14.3-22}$$

where $\lambda = c/\nu_0$ is the wavelength in the medium at the central frequency ν_0. Defining

$$\overline{n} = \frac{\lambda^3}{8\pi h}\, \varrho(\nu_0), \tag{14.3-23}$$

which represents the mean number of photons per mode, allows us to write (14.3-22) in the convenient form

$$\boxed{W_i = \overline{n}/t_{\text{sp}}\,.} \tag{14.3-24}$$

The interpretation of $\overline{n}$ as the mean number of photons per mode follows from the form of the ratio [see (14.3-12), (14.3-21), and (14.3-22)]

$$\frac{W_i}{P_{\text{sp}}} = \frac{\lambda^3 \varrho(\nu_0)}{8\pi h t_{\text{sp}}} \frac{1}{M(\nu_0) c\overline{S}} = \frac{\varrho(\nu_0)}{h\nu_0 M(\nu_0)}\,; \tag{14.3-25}$$

the quantity $\varrho(\nu_0)/h\nu_0$ represents the mean number of photons per unit volume in the vicinity of the frequency ν_0 while $M(\nu_0)$ is the number of modes per unit volume in the vicinity of ν_0. The probability density W_i is thus a factor of $\overline{n}$ greater than that for spontaneous emission, since each mode contains an average of $\overline{n}$ photons.

Einstein $\mathbb{A}$ and $\mathbb{B}$ Coefficients

Though Einstein did not have knowledge of (14.3-22), in 1917 he carried out an important analysis of the energy exchange between atoms and radiation that permitted him to obtain general expressions for the probability densities of spontaneous and stimulated transitions. He assumed that the atoms interacted with broadband radiation

of spectral energy density $\varrho(\nu)$, under conditions of thermal equilibrium, and obtained the following expressions:

$$P_{\text{sp}} = \mathbb{A} \tag{14.3-26}$$

$$W_i = \mathbb{B}\varrho(\nu_0). \tag{14.3-27}$$

Einstein's Postulates

The constants $\mathbb{A}$ and $\mathbb{B}$ are known as the **Einstein $\mathbb{A}$ and $\mathbb{B}$ coefficients**.

Comparison with (14.3-13) and (14.3-22) reveals that the $\mathbb{A}$ and $\mathbb{B}$ coefficients correspond to

$$\mathbb{A} = \frac{1}{t_{\text{sp}}} \tag{14.3-28}$$

$$\mathbb{B} = \frac{\lambda^3}{8\pi h t_{\text{sp}}}, \tag{14.3-29}$$

which are associated with spontaneous and stimulated transitions, respectively. The ratio is given by

$$\frac{\mathbb{B}}{\mathbb{A}} = \frac{\lambda^3}{8\pi h}. \tag{14.3-30}$$

The relation between the $\mathbb{A}$ and $\mathbb{B}$ coefficients is a result of the microscopic (rather than macroscopic) probability laws of interaction between an atom and the photons of each mode. We shall present an analysis similar to that provided by Einstein in Sec. 14.4.

EXAMPLE 14.3-1. ***Comparison Between Rates of Spontaneous and Stimulated Emission.*** Whereas the rate of spontaneous emission for an atom in the upper state is constant at $\mathbb{A} = 1/t_{\text{sp}}$, the rate of stimulated emission in the presence of broadband light, $\mathbb{B}\varrho(\nu_0)$, is proportional to the spectral energy density of the light, $\varrho(\nu_0)$. The two rates are equal when $\varrho(\nu_0) = \mathbb{A}/\mathbb{B} = 8\pi h/\lambda^3$; for larger values of the spectral energy density, the rate of stimulated emission exceeds that of spontaneous emission. If $\lambda = 1\ \mu\text{m}$, for example, $\mathbb{A}/\mathbb{B} = 1.66 \times 10^{-14}\ \text{J/m}^3\text{-Hz}$. This corresponds to an intensity spectral density $c\varrho(\nu_0) \approx 5 \times 10^{-6}\ \text{W/m}^2\text{-Hz}$ in free space. Thus, for a linewidth $\Delta\nu = 10^7$ Hz, the optical intensity at which the stimulated emission rate equals the spontaneous emission rate is $50\ \text{W/m}^2$ or $5\ \text{mW/cm}^2$.

Summary

An atomic transition may be considered in terms of its resonance frequency $\nu_0 = (E_2 - E_1)/h$, spontaneous lifetime t_{sp}, and lineshape function $g(\nu)$, which has linewidth $\Delta\nu$. The average transition cross section is

$$\overline{\sigma}(\nu) = \overline{S}g(\nu) = \frac{\lambda^2}{8\pi t_{\text{sp}}}\, g(\nu). \tag{14.3-15}$$

Spontaneous Emission

- If the atom is in the upper level and in a cavity of volume V, the probability density (per second) of emitting spontaneously into one *prescribed* mode of frequency ν is

$$p_{\text{sp}} = \frac{c}{V}\,\sigma(\nu). \tag{14.3-1}$$

- The probability density of spontaneous emission into *any* of the available modes is

$$P_{\text{sp}} = \frac{8\pi\overline{S}}{\lambda^2} = \frac{1}{t_{\text{sp}}}\,. \tag{14.3-13}$$

- The probability density of emitting into modes lying only in the frequency band between ν and $\nu + d\nu$ is $P_{\text{sp}}\,d\nu = (1/t_{\text{sp}})g(\nu)\,d\nu$.

Stimulated Emission and Absorption

- If the atom in the cavity is in the upper level and a radiation mode contains n photons of frequency ν, the probability density of emitting a photon into that mode is

$$W_i = n\,\frac{c}{V}\,\sigma(\nu). \tag{14.3-6}$$

  If the atom is instead in the lower level, and a mode contains n photons, the probability density of absorption of a photon from that mode is also given by (14.3-6).
- If instead of being in a cavity, the atom is illuminated by a monochromatic beam of light of frequency ν, with mean photon-flux density ϕ (photons per second per unit area), the probability density of stimulated emission (if the atom is in the upper level) or absorption (if the atom is in the lower level) is

$$W_i = \phi\,\sigma(\nu). \tag{14.3-19}$$

- If the light illuminating the atom is polychromatic, but narrowband in comparison with the atomic linewidth, and has a mean spectral photon-flux density ϕ_ν (photons per second per unit area per unit frequency), the probability density of stimulated emission/absorption is

$$W_i = \int \phi_\nu\,\sigma(\nu)\,d\nu. \tag{14.3-31}$$

- If the light illuminating the atom has a spectral energy density $\varrho(\nu)$ that is broadband in comparison with the atomic linewidth, the probability density of stimulated emission/absorption is

$$W_i = \mathbb{B}\varrho(\nu_0), \tag{14.3-27}$$

  where $\mathbb{B} = \lambda^3/8\pi h t_{\text{sp}}$ is the Einstein $\mathbb{B}$ coefficient.

In all of these formulas, $c = c_o/n$ is the velocity of light and $\lambda = \lambda_o/n$ is the wavelength of light in the atomic medium, and n is the refractive index.

D. Line Broadening

Because the lineshape function $g(\nu)$ plays an important role in atom–photon interactions, we provide in a brief discussion of some of the mechanisms that lead to line broadening. The same lineshape function applies for spontaneous emission, absorption, and stimulated emission.

Lifetime Broadening

Atoms can undergo transitions between energy levels by both radiative and nonradiative means. Radiative transitions are associated with photon absorption and emission, whereas nonradiative transitions permit energy transfer to take place via mechanisms such as lattice vibrations, inelastic collisions among the constituent atoms, and inelastic collisions with the walls of the vessel. Each atomic energy level has a lifetime τ, which is the inverse of the rate at which its population decays, radiatively or nonradiatively, to all lower levels.

The lifetime τ_2 of energy level 2 shown in Fig. 14.3-1, for example, represents the inverse of the rate at which the population of that level decays to level 1 and to all other lower energy levels (none of which are shown in the figure), by either radiative or nonradiative means. Since $1/t_{\text{sp}}$ is the radiative decay rate from level 2 to level 1, the overall decay rate $1/\tau_2$ must be greater, i.e., $1/\tau_2 \geq 1/t_{\text{sp}}$, thus corresponding to a shorter decay time, $\tau_2 \leq t_{\text{sp}}$. The lifetime τ_1 of level 1 is defined similarly. Clearly, if level 1 is the lowest allowed energy level (the ground state), then it will never decay and $\tau_1 = \infty$.

Lifetime broadening is, in essence, a Fourier transform effect. The lifetime τ of an energy level is related to the time uncertainty of the occupation of that level. As shown in Sec. A.1 of Appendix A, the Fourier transform of an exponentially decaying harmonic field $e^{-t/2\tau}e^{j2\pi\nu_0 t}$, which has an energy that decays as $e^{-t/\tau}$ with time constant τ, is proportional to $1/[1 + j4\pi(\nu - \nu_0)\tau]$. The full-width at half-maximum (FWHM) of the absolute square of this Lorentzian function of frequency is $\Delta\nu = 1/2\pi\tau$. This spectral uncertainty corresponds to an energy uncertainty $\Delta E = h\Delta\nu = h/2\pi\tau$. We conclude that a lifetime-broadened energy level with lifetime τ has an energy spread $\Delta E = h/2\pi\tau$, provided that the decay process can be modeled as a simple exponential. In this picture, spontaneous emission can be viewed in terms of a damped harmonic oscillator that generates an exponentially decaying harmonic function, as embodied in the Lorentz oscillator model presented in Sec. 5.5C.

Hence, if the energy spreads of levels 1 and 2 are $\Delta E_1 = h/2\pi\tau_1$ and $\Delta E_2 = h/2\pi\tau_2$, respectively, the spread in the energy difference corresponding to the transition between the two levels is

$$\Delta E = \Delta E_1 + \Delta E_2 = \frac{h}{2\pi}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right) = \frac{h}{2\pi}\frac{1}{\tau}, \tag{14.3-32}$$

where τ is the transition lifetime and $\tau^{-1} = (\tau_1^{-1} + \tau_2^{-1})$. The corresponding spread of the transition frequency, which is called the lifetime-broadening linewidth, is therefore

$$\boxed{\Delta\nu = \frac{1}{2\pi}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right).} \tag{14.3-33}$$

Lifetime-Broadening Linewidth

This spread is centered about the frequency $\nu_0 = (E_2 - E_1)/h$, and the lineshape

function has a Lorentzian profile:

$$g(\nu) = \frac{\Delta\nu/2\pi}{(\nu - \nu_0)^2 + (\Delta\nu/2)^2}\,. \tag{14.3-34}$$

Lorentzian Lineshape Function

More generally, the lifetime broadening associated with an atom or a collection of atoms may be modeled in the following manner. Each of the photons emitted in the course of a transition is represented by a wavepacket of central frequency ν_0 (the transition resonance frequency), with an exponentially decaying field envelope with decay time 2τ, which corresponds to an energy decay time equal to the transition lifetime τ. As illustrated in Fig. 14.3-7, the radiated light is taken to be a sequence of such wavepackets emitted at random times. As discussed in Example 12.1-1, this corresponds to random (partially coherent) light with a spectral intensity that is described precisely by the Lorentzian function given in (14.3-34), with $\Delta\nu = 1/2\pi\tau$.

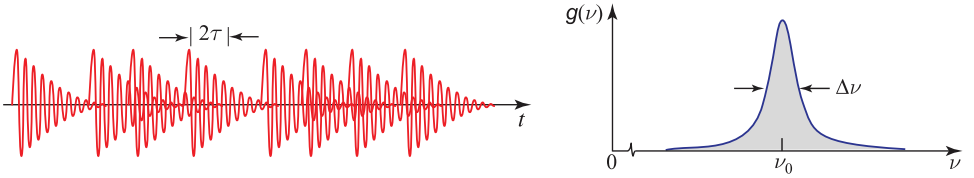


Figure 14.3-7 Wavepacket emissions at random times from a lifetime-broadened atomic system with transition lifetime τ. The light emitted has a Lorentzian spectral intensity of width $\Delta\nu = 1/2\pi\tau$.

The value of the Lorentzian lineshape function at the central frequency ν_0 is

$$g(\nu_0) = 2/\pi\Delta\nu, \tag{14.3-35}$$

so that the peak transition cross section, given by (14.3-16), becomes

$$\overline{\sigma}_0 = \frac{\lambda^2}{2\pi}\,\frac{1}{2\pi t_{\rm sp}\Delta\nu}\,. \tag{14.3-36}$$

The largest transition cross section occurs under ideal conditions when the decay is entirely radiative so that $\tau_2 = t_{\rm sp}$ and $1/\tau_1 = 0$ (which is the case when level 1 is the ground state from which no decay is possible). From (14.3-33), we then have $\Delta\nu = 1/2\pi t_{\rm sp}$, whereupon

$$\overline{\sigma}_0 = \lambda^2/2\pi\,, \tag{14.3-37}$$

indicating that the peak cross section is of the order of one square wavelength. When level 1 is not the ground state, or when nonradiative transitions are significant, we have $\Delta\nu \gg 1/2\pi t_{\rm sp}$ in which case $\overline{\sigma}_0$ can be significantly smaller than $\lambda^2/2\pi$. For example, for optical transitions in the range $\lambda = 0.1$ to 10 μm, values of $\lambda^2/2\pi$ lie between 10^{-11} and 10^{-7} cm^2, whereas observed values of σ_0 generally fall in the range between 10^{-20} and 10^{-12} cm^2 (see, e.g., Table 15.3-1).

Collision Broadening

Collisions in which energy is exchanged, called *inelastic collisions*, result in transitions between atomic energy levels. This affects the decay rates and lifetimes of all levels involved and modifies the linewidth of the radiated field considered above.

Collisions that do not involve an exchange of energy, called *elastic collisions*, also modify the linewidth of the radiated field, but in a different way. An elastic collision imparts a random phase shift to the wavefunction associated with the energy level, which in turn results in a random phase shift of the radiated field at each collision time. As illustrated in Fig. 14.3-8, a sine wave whose phase is modified by a random shift at random times (the collision times) exhibits spectral broadening. The spectrum of such a randomly dephased function can be determined by appealing to the theory of random processes. The result turns out to again be Lorentzian, with a width $\Delta\nu = f_{\text{col}}/\pi$, where f_{col} is the collision rate (mean number of collisions per second). Both lifetime and collision broadening are therefore accommodated by a Lorentzian lineshape function with an overall linewidth that is the sum of the individual linewidths:

$$\Delta\nu = \frac{1}{2\pi}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2} + 2f_{\text{col}}\right). \tag{14.3-38}$$

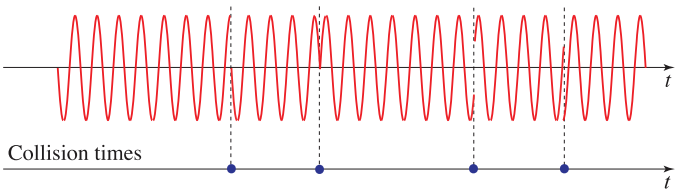


Figure 14.3-8 A sine wave interrupted at the rate f_{col} by random phase jumps has a Lorentzian spectrum of width $\Delta\nu = f_{\text{col}}/\pi$.

Inhomogeneous Broadening

Lifetime and collision broadening are examples of homogeneous broadening mechanisms in which the interacting atoms of a medium are all taken to be identical, with the same lineshape functions and center frequencies. For some media, however, different subsets of interacting atoms exhibit different behavior, either because of differences in the local environment or because of their dynamical behavior. The distinction is highlighted by examining the sublevels in the Nd^{3+}-ion manifolds discussed in Sec. 14.1B. Incorporating Nd^{3+} in a crystal such as YAG, which gives rise to **homogeneous broadening**, leads to energy sublevels that are distinct and narrow (Fig. 14.1-6). In contrast, these sublevels are smeared into bands when Nd^{3+} is embedded in a less structured material such as glass, which gives rise to **inhomogeneous broadening**. The origin of the distinction in this case is the random value of the electric field at different locations in the glass, which imparts position-dependent Stark shifts to the energy levels of the embedded Nd^{3+} ions that results in broadening. Analogously, different subsets of the excited active ions in the Ar^{+}-ion gas laser travel at different velocities and in different directions, resulting in a distribution of the lineshape-function center frequencies by virtue of the Doppler effect.

For inhomogeneously broadened media, we can define an average lineshape function

$$\overline{g}(\nu) = \langle g_\beta(\nu)\rangle, \tag{14.3-39}$$

where $\langle\cdot\rangle$ represents an average with respect to the variable β, which labels the subset of atoms with lineshape function $g_\beta(\nu)$. The average lineshape function is obtained by weighting the $g_\beta(\nu)$, which are known as *spectral packets*, by the fraction of the atomic population endowed with the property β, as pictured in Fig. 14.3-9.

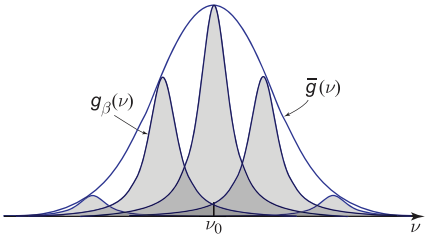


Figure 14.3-9 The average lineshape function for an inhomogeneously broadened collection of atoms.

A commonly encountered inhomogeneous broadening mechanism is Doppler broadening. As a result of the **Doppler effect**, an atom moving with velocity v along a given direction exhibits a lineshape function that is shifted by the frequency $\pm(v/c)\nu_0$ when viewed along that direction, where ν_0 is its central frequency. The shift is in the direction of higher frequency (+ sign) if the atom is moving toward the observer, and in the direction of lower frequency (− sign) if it is moving away. For an arbitrary direction of observation, the frequency shift is $\pm(v_\parallel/c)\nu_0$, where $v_\parallel$ is the component of velocity parallel to the direction of observation. Since a collection of atoms in a gas exhibits a distribution of velocities, as depicted in Fig. 14.3-10, the light they emit exhibits a range of frequencies that is known as **Doppler broadening**.

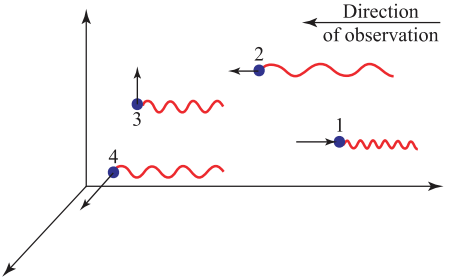


Figure 14.3-10 The frequency radiated by an atom depends on the direction of atomic motion relative to the direction of observation. Radiation from atom 1 has a higher frequency than that from atoms 3 and 4 since the atom is moving toward the observer. Radiation from atom 2 has a lower frequency since it is moving away from the observer.

For Doppler broadening, the velocity v therefore plays the role of the parameter β and $\bar{g}(\nu) = \langle g_v(\nu)\rangle$. As illustrated in Fig. 14.3-11, if $p(v)\,dv$ is the probability that the velocity of a given atom lies between v and $v + dv$, the overall inhomogeneous Doppler-broadened lineshape function is

$$\bar{g}(\nu) = \int_{-\infty}^{\infty} g\left(\nu - \nu_0 \frac{v}{c}\right) p(v)\,dv. \tag{14.3-40}$$

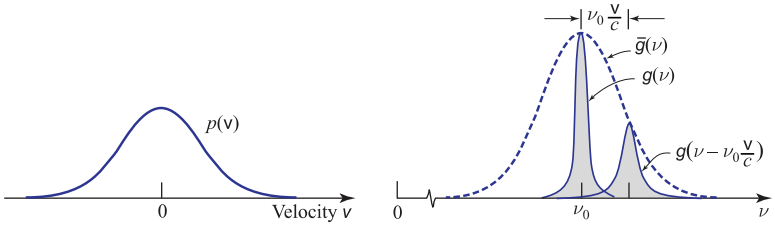


Figure 14.3-11 Velocity distribution and construction of the average lineshape function for a Doppler-broadened atomic system.

EXERCISE 14.3-2

Doppler-Broadened Lineshape Function.

(a) A collection of atoms in a gas has a component of velocity v along a particular direction that obeys the Gaussian probability density function

$$p(v) = \frac{1}{\sqrt{2\pi}\,\sigma_v} \exp\left(-\frac{v^2}{2\sigma_v^2}\right), \tag{14.3-41}$$

where $\sigma_v^2 = kT/M$ and M is the atomic mass. If each atom has a Lorentzian natural lineshape function of width $\Delta\nu$ and central frequency ν_0, derive an expression for the average lineshape function $\overline{g}\,(\nu)$.

(b) Show that if $\Delta\nu \ll \nu_0\sigma_v/c$, $\overline{g}\,(\nu)$ may be approximated by the Gaussian lineshape function

$$\overline{g}\,(\nu) = \frac{1}{\sqrt{2\pi}\,\sigma_D} \exp\left[-\frac{(\nu-\nu_0)^2}{2\sigma_D^2}\right], \tag{14.3-42}$$

where

$$\sigma_D = \nu_0\frac{\sigma_v}{c} = \frac{1}{\lambda}\sqrt{\frac{kT}{M}}\,. \tag{14.3-43}$$

The full-width at half-maximum (FWHM) Doppler linewidth $\Delta\nu_D$ is then

$$\Delta\nu_D = \sqrt{8\ln 2}\,\sigma_D \approx 2.35\,\sigma_D. \tag{14.3-44}$$

(c) Compute the Doppler linewidth $\Delta\nu_D$ for the $\lambda_o = 632.8$-nm laser transition in Ne and compare it with that for the $\lambda_o = 10.6$-μm laser transition in CO_2, assuming that $\Delta\nu \ll \nu_0\sigma_v/c$ and $T = 300°$ K. The relative atomic mass for Ne is $A_r \approx 20$, the analogous quantity for CO_2 is $\approx 12 + 16 + 16 = 44$, and the proton mass $m_p = 1.67 \times 10^{-27}$ kg.

(d) Show that the maximum value of the transition cross section for the Gaussian lineshape function in (14.3-42) is

$$\sigma_0 = \left(\frac{\lambda^2}{8\pi\,t_{sp}}\right)\overline{g}\,(\nu_0) \approx 0.94\,\frac{\lambda^2}{8\pi}\,\frac{1}{t_{sp}\,\Delta\nu_D}\,. \tag{14.3-45}$$

Compare this with (14.3-36) for the Lorentzian lineshape function.

Some atom–photon interactions exhibit broadening that is intermediate between pure homogeneous and pure inhomogeneous. Such mixed broadening can be modeled by an intermediate lineshape function such as the Voigt profile.

*E. Enhanced Spontaneous Emission

All of the results presented thus far in Sec. 14.3 are predicated on the assumption that $\Delta\nu \gg \delta\nu$, i.e., that the atomic linewidth $\Delta\nu$ is far greater than the width of an electromagnetic mode $\delta\nu$. This condition is usually, but not always, obeyed. In the opposite limit, when the atomic linewidth is far smaller than the width of an electromagnetic mode (Fig. 14.3-12), an enhancement of the spontaneous emission probability density can be achieved, particularly in high-Q microcavities, as we proceed to demonstrate. The enhancement of spontaneous emission is desirable for the operation of certain microcavity sources, as discussed in Sec. 18.5.

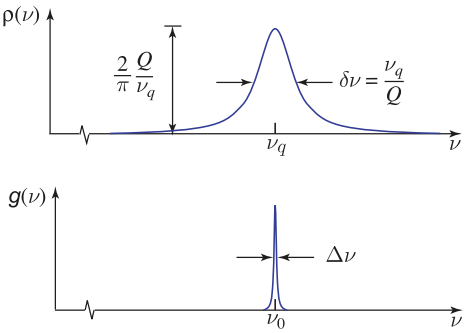


Figure 14.3-12 Spontaneous emission from an atom with normalized lineshape function $g(\nu)$ into a broader normalized Lorentzian cavity mode $\rho(\nu)$. The lineshape-function and cavity-mode center frequencies are designated by ν_0 and ν_q, respectively, while their widths are specified by $\Delta\nu$ and $\delta\nu$. We consider the case where $\nu_0 = \nu_q$ and $\Delta\nu \ll \delta\nu$.

Consider the spontaneous emission of an atom with resonance frequency ν_0 into an electromagnetic mode with center frequency $\nu_q = \nu_0$ in the regime $\Delta\nu \ll \delta\nu$, as portrayed in Fig. 14.3-12. In accordance with (14.3-11), when the dipole moment of the atom is aligned with the field direction of the mode, the probability density for spontaneous emission into a single cavity mode $\rho(\nu)$ is given by

$$P_{\rm sp}^{\rm max} = \int_0^\infty \frac{c}{V}\,\sigma_{\rm max}(\nu)\,\rho(\nu)\,d\nu \approx \frac{c}{V}\,\frac{3\lambda^2}{8\pi t_{\rm sp}}\,\rho(\nu_0)\int_0^\infty g(\nu)\,d\nu, \tag{14.3-46}$$

since $\sigma_{\rm max}(\nu) = 3\overline{\sigma}(\nu)$ and $\overline{\sigma}(\nu) = \lambda^2\,g(\nu)/8\pi t_{\rm sp}$, as provided in (14.3-10) and (14.3-15), respectively. Inasmuch as the lineshape function $g(\nu)$ is normalized, and the height of the normalized Lorentzian lineshape function of the cavity mode is $2Q/\pi\nu_q$, where $Q = \nu_q/\delta\nu$, we obtain

$$P_{\rm sp}^{\rm max} = \frac{1}{t_{\rm sp}}\,\frac{3c\lambda^2}{8\pi V}\,\frac{2Q}{\pi\nu_q} = \frac{1}{t_{\rm sp}}\cdot\frac{3}{4\pi^2}\,\frac{\lambda^3}{V}\,Q. \tag{14.3-47}$$

The net result is an enhancement of the spontaneous emission probability density relative to that in free space by a quantity known as the **Purcell factor**:

$$F_P = \frac{P_{\rm sp}^{\rm max}}{P_{\rm sp}} = \frac{3}{4\pi^2}\,\frac{\lambda^3}{V}\,Q. \tag{14.3-48}$$

Purcell Factor

The Purcell factor in (14.3-48) exhibits the following features:

- The factor of 3 is a result of the alignment of the dipole moment of the atom and the field direction of the mode.
- The quantity λ^3/V, which is the ratio of the cubed wavelength to the cavity volume, is substantially enhanced in a microcavity.
- A high value of Q, i.e., a sharp cavity mode, enhances the Purcell factor; however, as Q increases, $\delta\nu = \nu_q/Q$ decreases, so that ultimately the condition $\Delta\nu \ll \delta\nu$ is violated.

Finally, we note that as ν_0 deviates from ν_q, the height of the cavity mode at ν_0 becomes smaller and the enhancement of spontaneous emission ultimately becomes a suppression of spontaneous emission.

*F. Laser Cooling, Laser Trapping, and Atom Optics

The forces exerted by light on material media (Sec. 13.1D) can be used to reduce the velocity spread of a low-density collection of neutral atoms or ions (**laser cooling**) and to confine them to a reduced volume of space (**laser trapping**). Atomic beams, and atoms confined to traps, are used in **atom optics** and **atom interferometry** to carry out highly precise measurements. Cooled and trapped atoms in the form of **Bose–Einstein condensates**, along with trapped Fermi gases, offer fundamental insights into the quantum nature of matter.

Laser Cooling

A number of schemes exist for the laser cooling of ions and neutral atoms. Some make use of the external degrees of freedom of the atom, such as its position and momentum, while others rely on the dynamical interplay between its external and internal degrees of freedom, such as its electronic configuration and spin. Both absorptive (scattering) and dispersive (gradient or dipole) forces can play a role. Laser cooling has also been extended beyond neutral atoms toward more complex systems such as molecules and small macroscopic objects. Applications of laser cooling include precision metrology, high-resolution spectroscopy, and quantum science.

Doppler cooling. One of the simplest laser-cooling schemes, known as **Doppler cooling**, relies on a beam of atoms moving toward a narrow-linewidth laser beam whose center frequency is tuned slightly below the atomic line center. Following photon absorption by the subset of atoms whose Doppler-shifted frequency matches the photon frequency, an atom can return to the ground state via either stimulated emission or spontaneous emission. If it returns by stimulated emission, the momentum of the emitted photon is the same as that of the absorbed photon, leaving the atom with no net change of momentum. If it returns by spontaneous emission, on the other hand, the direction of the photon emission is random so that repeated absorptions and emissions result in a net decrease of the atom's momentum in the direction pointing toward the laser beam. The result of this absorptive effect is a decrease in the velocity of that subset of atoms, as shown schematically in Fig. 14.3-13. Other subsets of atoms can be cooled by changing the laser frequency or by using alternative techniques such as a broadband laser or an inhomogeneous magnetic field that modifies the atomic-line center frequency (a "Zeeman slower").

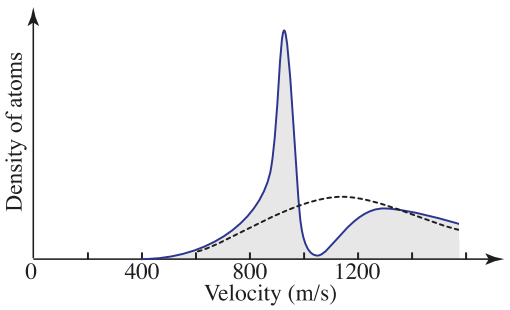


Figure 14.3-13 Velocity distribution of a beam of Na atoms (dotted curve). A laser beam of fixed frequency serves to transfer atoms in a narrow velocity range to an even narrower velocity range centered at a lower velocity, thereby Doppler cooling the atomic beam (solid curve). (Adapted from W. D. Phillips, Laser Cooling and Trapping of Neutral Atoms, *Reviews of Modern Physics*, vol. 70, pp. 721–741, 1998, Fig. 3 ©1998 by the American Physical Society.)

Optical molasses. When three pairs of orthogonally oriented, counterpropagating laser beams are used for **3D Doppler cooling** a collection of atoms, each atom encounters a velocity-dependent viscous-damping force regardless of the direction in which it moves. The laser beams are said to behave like **optical molasses** since they strongly resists the velocity spread of the atoms, though there is no restoring force to

push the atoms to the center of the apparatus. The calculated *Doppler-cooling limit* for an idealized two-level system is in the vicinity of several hundred μK.

Sisyphus cooling. Doppler cooling is widely used but other approaches allow atoms and ions to be cooled to temperatures orders of magnitude below the Doppler-cooling limit. Atoms can undergo **polarization gradient cooling** or **Sisyphus cooling** by making use of a set of orthogonal, linearly polarized, counterpropagating laser beams that produce a standing wave whose polarization varies on a subwavelength scale. Atoms in such a standing wave experience a combination of a position-dependent AC Stark shift (light shift), a gradient or dipole force, and optical pumping that allow the temperature to be reduced to the *single-photon recoil limit* associated with the atomic emission of a single photon (Exercise 13.1-2). Temperatures in the few μK regime, well below the Doppler-cooling limit, are obtained.

Evaporative cooling. It is also possible to reduce temperatures below the single-photon recoil limit. One approach for doing so, known as **evaporative cooling**, operates by gradually reducing the trap depth so that atoms with energies exceeding this depth escape, leaving behind less energetic atoms. A reduced temperature results when thermal equilibrium is reestablished by subsequent collisions. Another subrecoil cooling technique can be implemented by making use of **velocity-sensitive coherent population trapping**. Physical insight into this process can be obtained by considering the velocity of the atom in terms of a random walk that obeys Lévy statistics. Careful design of sub-recoil cooling configurations allows temperatures to be reduced to a few nK, corresponding to atomic velocities smaller than cm/s. Cooling to tens of pK has been achieved in gravity-free environments.

Laser Trapping

Laser trapping, often combined with laser cooling, can be used to confine, manipulate, and study ions, neutral atoms, molecules, dielectric particles, biological cells, and small macroscopic objects. Laser trapping and cooling is useful in quantum and nonlinear optics; a single atom and a single photon can be trapped together to implement cavity quantum electrodynamics in its most elemental configuration. It also allows light to be shed on the physics of cold-atom collisions and microfluidics and enables Bose–Einstein condensates and atom lasers to be created.

Optical tweezers. **Optical tweezers** are widely used for confining and manipulating particles that range in size from single atoms to small biological structures. Developed in the 1980s and originally called a "single-beam gradient force trap," they can be used in zero, one, two, or three dimensions. Optical tweezers make use of focused laser beams to physically contain samples via scattering and gradient forces at levels that extend from fN to pN. The particle recoil associated with Rayleigh scattering (Sec. 5.6B), and ray optics, provide suitable descriptions for the confinement mechanisms when the trapped particle sizes are, respectively, much smaller than, and much larger than, the optical wavelength. For homogeneous spherical particles, Mie scattering theory (Sec. 5.6C) provides accurate numerical results for arbitrary particle sizes and refractive indices. The confining laser beams can also impart torque to a trapped particle by transferring spin angular momentum if the beams are circularly polarized, or orbital angular momentum if they are structured. Examples of the application of optical tweezers at the extremes of small and large trapped particles are, respectively: 1) the confinement of a single Å-size neutral atom, thereby enabling it to be cooled to its three-dimensional vibrational ground state; and 2) the confinement and manipulation of μm-size objects such as living organisms and DNA molecules. Variations on the theme, such as magnetic tweezers and acoustic tweezers, have also been implemented.

Magneto-optical traps. The atoms in optical molasses may also be spatially confined by incorporating a **magneto-optical trap** (**MOT**) that makes use of an inhomogeneous magnetic field. This gives rise to Zeeman splitting of the atomic energy levels and results in a net position-dependent scattering force that directs the atoms to the location where the three pairs of laser beams intersect. The net result is cooling and trapping.

Optical lattices. As indicated above, polarization gradient cooling leads to the localization of atoms at a spatial scale finer than the wavelength of light. In essence, the standing waves produced by counterpropagating laser beams in 1D, 2D, or 3D behave as **optical lattices** comprising potential wells in which atoms are trapped and well-localized by the interaction of the AC electric field and the induced atomic dipole moment (AC Stark effect). The well depth of the optical lattice can be modified by altering the laser power, while its periodicity can be tuned by changing the laser wavelength or the relative angle between the laser beams. Moreover, an auxiliary laser field can facilitate trapping in a lattice potential whose features are far smaller than the wavelength of the lasers. Limiting the atomic positions in this manner achieves laser cooling and trapping in a lattice. Optical lattices are important for attaining motional control of collections of single neutral atoms. They can be used, for example, to hold atoms whose optical transition can serve as a highly accurate atomic clock (e.g., fermionic ^{87}Sr), or they can be loaded with deterministic arrays of atoms. Optical lattices have allowed simplified artificial crystals to be generated, whose properties are tunable, thereby bringing clarity to fundamental features of importance in condensed-matter physics.

Atom Optics

Atom optics is concerned with the study of beams of moving neutral atoms and the **matter waves** associated with them. As with optical waves, matter waves exhibit phenomena such as reflection, refraction, diffraction, scattering, and interference. The role of the optical wavelength λ is played by the de Broglie wavelength λ_{dB}, which is related to the momentum p of the atoms by $\lambda_{\text{dB}} = h/p$, where h is Planck's constant. This expression is the same as that provided in (13.1-11) for photons; however, the atomic momentum $p = mv$ is orders of magnitude greater than the photon momentum, so that $\lambda_{\text{dB}} \ll \lambda$. As an example, the de Broglie wavelength for thermal Na atoms, $\lambda_{\text{dB}} = h/\sqrt{3kTm} \approx 20$ pm (see Exercise 13.1-2), is some 30 000 times smaller than the wavelength of visible light. The coherence length for a thermal atomic beam is also short, $l_c \approx 100$ pm.

Atom interferometry. Atom interferometry can be conducted with atoms in beams, magneto-optical traps, or Bose–Einstein condensates. Cooled and trapped atoms are important components in the arsenal of atom optics since they offer reduced atomic momentum and uncertainty. Ultracold atoms can exhibit de Broglie wavelengths as long as 1 μm and coherence lengths that can extend up to $l_c \approx 10$ μm. Matter-wave interferometry is thus akin to optical interferometry with partially coherent light (Sec. 12.2). Optical transitions in single atoms confined in optical lattices often serve as atomic clocks. Atom interferometers are useful for the precise measurement of acceleration, rotation, and gravity, as well as for the determination of various atomic and material properties. Unlike photons, however, atoms can interact strongly so that matter-wave interferometry is often nonlinear.

Bose–Einstein Condensates and Atom Amplifiers

Bose–Einstein condensates. A gas comprising bosonic atoms in thermal equilibrium can be treated as a quantum collection of indistinguishable particles described by the Bose–Einstein distribution presented in (13.2-20) and Fig. 13.2-7, much as for a photon gas in thermal equilibrium. A **Bose–Einstein condensate** (**BEC**) can be

formed when a bosonic-atom gas is cooled to a temperature sufficiently low that the particle kinetic energy decreases to a negligible value and λ_{dB} becomes comparable to the interatomic separation. The atomic wavepackets then overlap sufficiently so that they can be considered as condensing into a single quantum state that minimizes the system's free energy (the gas must be sufficiently dilute so that it does not condense into a liquid or solid as it is cooled). Particles in that state then act collectively as a coherent wave. A gas of ^{23}Na atoms cooled below about 1 μK, for example, can form a BEC that contains somewhere between 10^2 and 10^8 atoms and has a spatial extent of tens of μm. The interference exhibited by two separate Bose–Einstein condensates clearly illustrates the wavelike properties of matter on a macroscopic scale. The mathematical description of a BEC in terms of quantized matter waves is not unlike that of quantized electromagnetic waves in a nonlinear refractive medium.

Miniaturization has led to the production of BECs on specially designed electronic microchips known as **atom chips**. Cooled atoms extracted from optical molasses in an MOT are transferred to a **magnetic trap** on the chip created by electric currents. Evaporative cooling is instigated by radiowaves generated by the chip, causing a BEC to be formed. An atom-chip BEC is being installed in the International Space Station, where the absence of gravity is expected to enable temperatures to be reduced to the pK level. BECs provide a path to the generation of synthesized forms of quantum matter and are expected to offer atom interferometry with unprecedented accuracy.

Atom amplifiers. An **atom amplifier** relies on increasing the number of atoms in a beam passed through a BEC atomic cloud that serves as the active medium. The atom amplifier converts atomic waves in the active medium into atomic waves that assume the same quantum state as the wave to be amplified. One way to achieve this is by pumping the BEC with a laser whose photons are scattered by atoms in the BEC at the precise angles that permits recoiling condensate atoms to augment the input matter wave, while conserving momentum and energy. Such devices could in principle be used to improve the performance of atom interferometers. However, since atoms are conserved and cannot be created on demand, amplified atom beams are typically weak. The laser amplifier, in contrast, relies on photons, which are readily created in large numbers on demand, thereby allowing amplified laser light to be made highly intense. Another distinction between laser and BEC amplifiers is that the laser active medium comprises noninteracting photons in a distinctly nonequilibrium state whereas the BEC active medium comprises interacting atoms in thermodynamic equilibrium.

14.4 THERMAL LIGHT

Under conditions of thermal equilibrium, and in the absence of other external energy sources, a universal form of radiation known as **thermal light** or **blackbody radiation** is emitted from a blackbody (these objects are so-named because they absorb all of the light incident on them). In this section we determine the properties of thermal light by examining the interactions among a collection of photons and atoms in thermal equilibrium. We also show how the thermal light emitted from an object can be used to image it.

A. Thermal Equilibrium Between Photons and Atoms

A macroscopic rate-equation approach that balances spontaneous emission, absorption, and stimulated emission, under conditions of thermal equilibrium, leads to the spectral energy density of thermal light. The point of departure for our analysis is (14.3-13) and (14.3-24), which govern spontaneous emission and induced transitions, respectively, in

the presence of broadband light. Consider a cavity of unit volume whose walls consist of large numbers of atoms with two energy levels, denoted 1 and 2, that are separated by an energy difference $h\nu$. The cavity, which is maintained at temperature T, supports broadband radiation that can be observed through a small hole. Let $N_1(t)$ and $N_2(t)$ represent the numbers of atoms per unit volume occupying energy levels 1 and 2, at time t, respectively. Since some of the atoms are initially in level 2, as ensured by the finite temperature, spontaneous emission creates radiation in the cavity. This radiation in turn can induce absorption and stimulated emission. The three processes coexist and it is assumed that steady-state (equilibrium) conditions are attained. We assume that an average of $\overline{n}$ photons occupies *each* of the radiation modes whose frequencies lie within the atomic linewidth, as established in (14.3-24).

We first consider spontaneous emission alone. The probability that a single atom in the upper level undergoes spontaneous emission into any of the modes, within the time increment from t to $t+\Delta t$, is $P_{\rm sp}\Delta t = \Delta t/t_{\rm sp}$. There are $N_2(t)$ such atoms so that the average number of emitted photons within Δt is $N_2(t)\Delta t/t_{\rm sp}$. This is also the number of atoms that depart from level 2 during the time interval Δt. Hence, the (negative) rate of increase of $N_2(t)$ arising from spontaneous emission is described by the differential equation

$$\frac{dN_2}{dt} = -\frac{N_2}{t_{\rm sp}}. \tag{14.4-1}$$

The solution, $N_2(t) = N_2(0)\exp(-t/t_{\rm sp})$, is an exponentially decaying function of time, as displayed in Fig. 14.4-1. Given sufficient time, the number of atoms in the upper level N_2 decays to zero with time constant $t_{\rm sp}$. The energy is carried off by the spontaneously emitted photons.

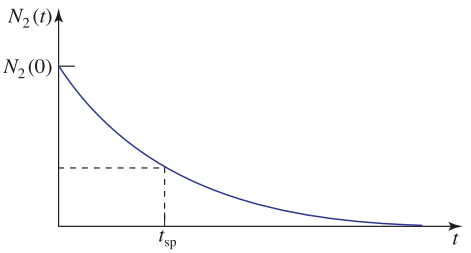


Figure 14.4-1 Decay of the upper-level population caused by spontaneous emission alone.

We now incorporate absorption and stimulated emission, which contribute to changes in the populations. Since there are N_1 atoms capable of absorption, the rate of increase of the population of atoms in the upper energy level arising from absorption is, based on (14.3-24),

$$\frac{dN_2}{dt} = N_1 W_i = \frac{\overline{n}\, N_1}{t_{\rm sp}}. \tag{14.4-2}$$

Similarly, stimulated emission gives rise to a (negative) rate of increase of atoms in the upper state, expressed as

$$\frac{dN_2}{dt} = -N_2 W_i = -\frac{\overline{n}\, N_2}{t_{\rm sp}}. \tag{14.4-3}$$

The rates of atomic absorption and stimulated emission are both proportional to $\overline{n}$, the average number of photons in each mode.

Combining (14.4-1), (14.4-2), and (14.4-3) to accommodate spontaneous emission, absorption, and stimulated emission together, yields the rate equation

$$\boxed{\frac{dN_2}{dt} = -\frac{N_2}{t_{sp}} + \frac{\overline{n}\,N_1}{t_{sp}} - \frac{\overline{n}\,N_2}{t_{sp}}\,.} \qquad \begin{matrix}(14.4\text{-}4)\\ \text{Rate Equation}\end{matrix}$$

This equation ignores transitions into or out of level 2 that arise from other effects, such as interactions with other energy levels, nonradiative transitions, and external sources of excitation. Steady-state operation demands that $dN_2/dt = 0$, which leads to

$$\frac{N_2}{N_1} = \frac{\overline{n}}{1+\overline{n}}\,. \tag{14.4-5}$$

Clearly, $N_2/N_1 \leq 1$. If we now make use of the fact that the atoms are in thermal equilibrium, (14.2-2) dictates that their populations obey the Boltzmann distribution:

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{kT}\right) = \exp\left(-\frac{h\nu}{kT}\right). \tag{14.4-6}$$

Substituting (14.4-6) into (14.4-5) leads to a mean number of photons per mode near frequency ν given by

$$\overline{n} = \frac{1}{\exp(h\nu/kT) - 1}\,. \tag{14.4-7}$$

The foregoing derivation is predicated on the interaction of two energy levels coupled by absorption, as well as by stimulated and spontaneous emission, at a frequency near ν. The applicability of (14.4-7) is, however, far broader. This may be understood by considering a cavity whose walls are made of solid materials that possess a continuum of energy levels at all energy separations, and therefore all values of ν. Atoms in the walls spontaneously emit into the cavity. The emitted light subsequently interacts with the atoms in the walls, giving rise to absorption and stimulated emission. If the walls are maintained at temperature T, the combined system of atoms and radiation reaches thermal equilibrium, whatever the nature of the walls and the shape of the cavity.

Equation (14.4-7) is identical to (13.2-21) — the expression for the mean photon number in a mode of thermal light for which the occupation of the modal energy levels follows the distribution $p(n) \propto \exp(-nh\nu/kT)$. This indicates a self-consistency in our analysis. Photons interacting with atoms in thermal equilibrium at temperature T are themselves in thermal equilibrium at the same temperature T (Sec. 13.2C). A collection of such photons is often termed a "photon gas."

B. Blackbody Radiation Spectrum

Based on the discussion provided in Sec. 14.4A, the average energy $\overline{E}$ of a radiation mode is simply $\overline{n}\,h\nu$, where $\overline{n}$ is given by (14.4-7), which leads to

$$\boxed{\overline{E} = \frac{h\nu}{\exp(h\nu/kT) - 1}\,.} \qquad \begin{matrix}(14.4\text{-}8)\\ \text{Average Energy of a}\\ \text{Mode in Thermal Equilibrium}\end{matrix}$$

The dependence of $\overline{E}$ on ν, which is identical to that set forth in (13.2-24), is portrayed in Fig. 14.4-2.

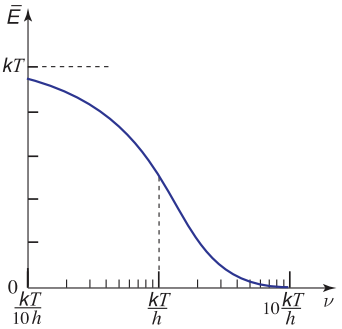


Figure 14.4-2 Semilogarithmic plot of the average energy $\overline{E}$ of an electromagnetic mode in thermal equilibrium at temperature T, as a function of the mode frequency ν. At $T = 300°$ K, $kT/h = 6.25$ THz, which corresponds to a wavelength of 48 μm.

Multiplying the average energy per mode $\overline{E}$ by the 3D modal density $M(\nu) = 8\pi\nu^2/c^3$ provided in (11.3-10) gives rise to a spectral energy density $\varrho(\nu) = M(\nu)\overline{E}$ (energy per unit bandwidth per unit cavity volume) that takes the form

$$\varrho(\nu) = \frac{8\pi h\nu^3}{c^3}\,\frac{1}{\exp(h\nu/kT) - 1}\,. \tag{14.4-9}$$

Spectral Energy Density for Blackbody Radiation

This formula, known as the **blackbody radiation spectrum** or **Planck spectrum**, is plotted in Fig. 14.4-3 as a function of frequency on double-linear coordinates. A plot with temperature as a parameter is shown in the iconic graph provided in Fig. 14.4-4.

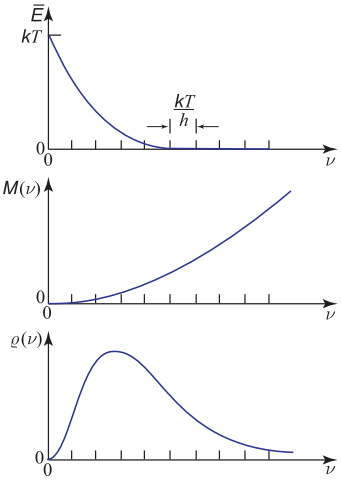


Figure 14.4-3 Frequency dependence of the energy per mode $\overline{E}$, the density of modes $M(\nu)$, and the spectral energy density $\varrho(\nu) = M(\nu)\overline{E}$ for blackbody radiation, on double-linear coordinates.

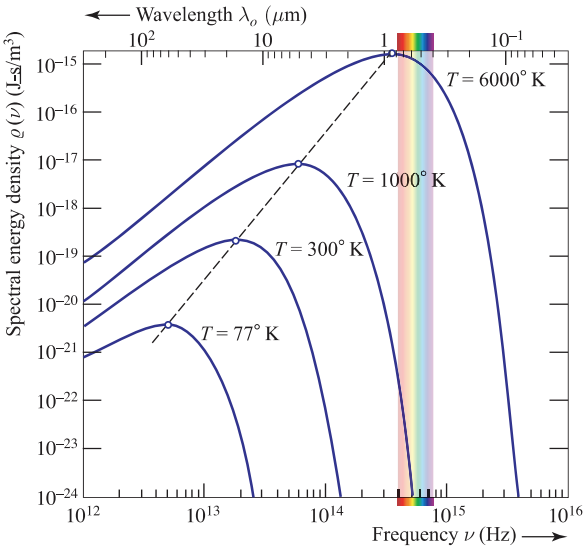


Figure 14.4-4 Dependence of the blackbody spectral energy density $\varrho(\nu)$ on frequency for several different temperatures, on double-logarithmic coordinates.

As the blackbody temperature is altered, the mean number of photons in the cavity changes in accordance with (14.4-7) since photons can emerge from, or disappear into, the walls of the cavity. Though they are bosons, the photons in a blackbody cavity are not conserved and thus do not form a Bose–Einstein condensate. Bosonic atoms, on the other hand, are conserved so they can form a BEC when the temperature is reduced below a critical value (Sec. 14.3F).

The spectrum of blackbody radiation played a central role in the discovery of the photon nature of light. Based on classical electromagnetic theory, the modal density for a three-dimensional cavity was long known to be given by $M(\nu) = 8\pi\nu^2/c^3$, as provided in (11.3-10). Moreover, the equipartition law in classical statistical mechanics specified that the average energy per mode was constant at $\overline{E} = kT$, independent of the modal frequency. This led to a theoretical expression for the blackbody spectrum, $\varrho(\nu) = M(\nu)\overline{E} = 8\pi\nu^2 kT/c^3$, known as the **Rayleigh–Jeans formula**, but it failed to agree with experiment in a significant way. Max Planck resolved the dilemma in 1900 by observing that imposing quantization on the allowed energies of the atoms in the walls of the cavity led instead to (14.4-9), a result that agreed with experiment. Einstein subsequently built on Planck's approach and proposed that the quantization be imposed directly on the energy of the electromagnetic radiation, which led to the concept of the photon. The Rayleigh–Jeans formula is recovered from (14.4-9) in the classical limit $h\nu \ll kT$ by making use of the approximation $\exp(h\nu/kT) \approx 1 + h\nu/kT$.

EXERCISE 14.4-1

Frequency of Maximum Blackbody Energy Density. Using the blackbody radiation law $\varrho(\nu)$, show that the frequency $\nu_{\rm p}$ at which the spectral energy density is maximum satisfies the equation $3(1 - e^{-x}) = x$, where $x = h\nu_{\rm p}/kT$. Find x approximately and determine $\nu_{\rm p}$ at $T = 300^\circ$ K.

The total power radiated by a blackbody increases with temperature as T^4, a result known as the **Stefan–Boltzmann law** (Prob. 14.4-7). Although they are not perfect blackbodies, planets and stars emit light with a spectral character that approximately follows (14.4-9). The temperature of the sun and earth, which can be estimated using the Stefan–Boltzmann law, turn out to be $T \approx 5800^\circ$ K and $T \approx 300^\circ$ K, respectively.

Thermography

The blackbody spectral energy-density formula (14.4-9) is useful for generating maps (images) of the temperature distribution of thermal objects. This is achieved by using a camera that is sensitive in the wavelength region of the object's thermal emissions (Fig. 14.4-4). Hot objects, such as the sun, emit most strongly in the visible region, whereas objects of moderate temperature, such as the earth and humans, typically radiate in the mid-infrared region. Cold objects radiate in the far-infrared. The imaging of thermal objects by means of their self-radiation is known as **thermography**. Thermographic cameras contain an array of photodetectors sensitive in a particular region of the spectrum (Sec. 19.5). The technique is often used in the wavelength region $0.7\ \mu\text{m} \leqslant \lambda_o \leqslant 300\ \mu\text{m}$, corresponding to 12° K $\leqslant T \leqslant 5200^\circ$ K. Though thermography is facilitated at higher temperatures by the T^4 dependence of the total radiated power, the representative images presented in Fig. 14.4-5 illustrate the broad range of temperatures that can be accessed.

Thermography is used to garner information about objects and scenes that exhibit temperature variations. Different local temperatures are typically displayed as false colors. The technique finds use in industrial applications, such as monitoring the overheating of circuit boards and the evolution of oil spills. It is of assistance in search-

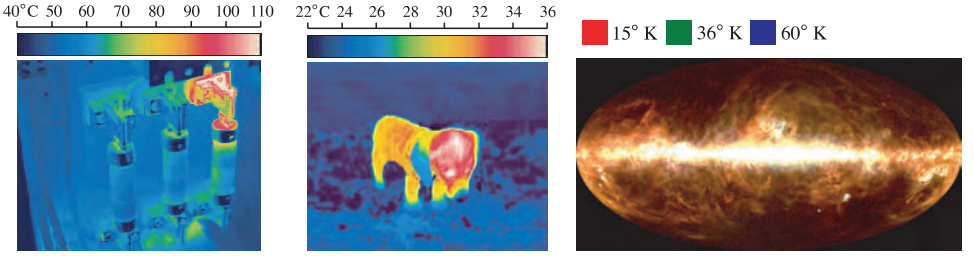


Figure 14.4-5 Representative thermographic images in different temperature regions for use in industrial-systems analysis (left); search and rescue (center); and cosmology (right).

and-rescue missions for humans and animals, even when they are concealed in dense foliage at night. Thermography is also used in clinical medicine since skin-surface temperature is a diagnostic for blood-flow blockages and tumors. Environmental applications include fire-fighting and forestry. The technique is invaluable in astronomy and cosmology since it allows astronomical objects, such as cooler red stars and red giants, to be imaged in the near infrared; planets, comets, and asteroids to be seen in the mid-infrared; and central galactic regions and emissions from cold dust to be imaged in the far-infrared.

14.5 LUMINESCENCE AND SCATTERING

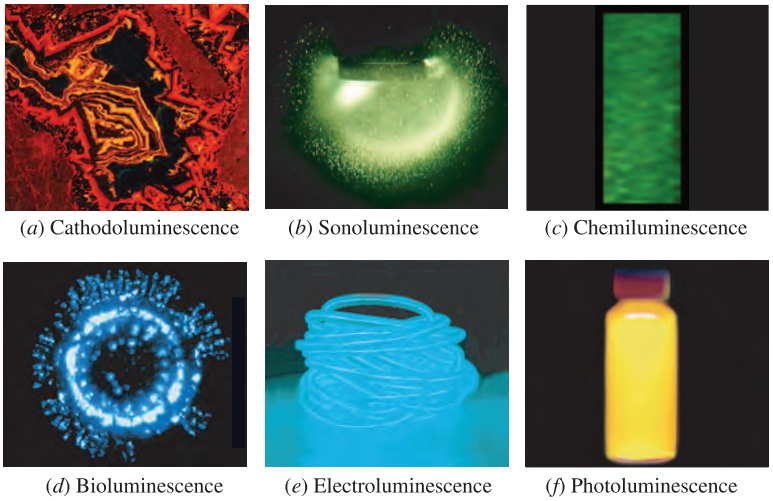
Thermal excitation is not the only external source of energy that can result in the emission of light from a material system that is excited to a higher energy level and then decays back to its ground state. Other sources of excitation, such as electron beams, chemical reactions, and electric fields, can also result in light being emitted. So too can excitation in the form of one or more photons via a process known as photoluminescence. Nonthermal radiators that operate on this principle are generically called **luminescent radiators** and the radiation process is known as **luminescence**.

While photoluminescence involves the absorption and subsequent emission of photons, light can also scatter from a material system in a resonant or nonresonant manner. Linear and nonlinear scattering, such as Rayleigh and Raman scattering, respectively, play important roles in optics and photonics.

A. Forms of Luminescence

The form of the luminescence is classified according to the source of excitation, as indicated by the examples provided below and illustrated in Fig. 14.5-1. It is also worthy of mention that most lasers operate by amplifying luminescence radiation via stimulated emission; the initial source of light is most often photoluminescence, collision-induced luminescence, or electroluminescence (Chapters 15–18).

Cathodoluminescence. The emission of light from a material as a result of excitation by energetic electrons. Examples are the images at the face of a cathode-ray tube or an image intensifier, which are induced in phosphors at the screen by the electrons. Cathodoluminescence is frequently used for assaying the composition of a material because of two salutary features: 1) the depth of penetration into the sample can be modified by changing the electron energy, and 2) different material components give rise to emission at different wavelengths. The light is termed **betaluminescence** when the exciting electrons are the result of nuclear beta decay.



(*a*) Cathodoluminescence (*b*) Sonoluminescence (*c*) Chemiluminescence

(*d*) Bioluminescence (*e*) Electroluminescence (*f*) Photoluminescence

Figure 14.5-1 (*a*) Cathodoluminescence from a mineral sample reveals the presence of zoned calcite and saddle dolomite in boxwork breccia. The edge dimension is 1.3 mm and the electron energy is 22 keV (courtesy Charles M. Onasch, Bowling Green State University). (*b*) Multibubble sonoluminescence created by an ultrasonic horn immersed in liquid (courtesy Kenneth S. Suslick, University of Illinois at Urbana–Champaign). (*c*) Chemiluminescence from a lightstick. (*d*) The deep-sea scyphomedusa *Atolla vanhoeffeni* (diameter $\approx$ 3 cm) is abundant throughout the world and produces bioluminescence when disturbed (courtesy Edith A. Widder, Ocean Research & Conservation Association). (*e*) The electric field across a pair of parallel wires held at different potentials elicits electroluminescence from a powdered material coating them. (*f*) Photoluminescence from colloidal CdSe quantum dots dispersed in hexane following illumination by ultraviolet light (see Fig. 14.1-13).

Sonoluminescence. The emission of light induced by acoustic cavitation, namely the formation, growth, and collapse of bubbles in a liquid irradiated with high-intensity sound or ultrasound. The light comprises brief flashes (duration $\approx$ 100 ps) emitted when the collapsing bubbles reach minimum size ($\approx$ 1 μm). Sonoluminescence is generally observed from clouds of bubbles although it is possible to generate single-bubble sonoluminescence with a stable period and position by trapping single bubbles in an acoustic standing wave.

Chemiluminescence. The emission of light via a chemical reaction. Chemiluminescence is observed under those relatively rare circumstances when the reaction between two or more chemicals releases sufficient energy to populate the excited state of a reaction product. Lightsticks, used for illumination in underwater and military environments, are an example: they glow when the seal between two compartments containing different chemicals is broken and the chemicals are permitted to mix. The color of the emitted light is determined by the dye incorporated in the chemical mixture. Chemiluminescence is responsible for the operation of chemical lasers (Sec. 16.3E).

Bioluminescence. Chemiluminescence produced by living organisms such as fireflies and jellyfish. Bioluminescence provides a means of communication; indeed, some organisms such as fireflies synchronize their flashes. Most deep-sea marine organisms produce bioluminescence as a matter of course, often in the blue–green region of the spectrum where seawater is most transparent, but also at other wavelengths. An important technique employed in biology makes use of genetic engineering to insert a jellyfish gene that expresses the green fluorescent protein (GFP) adjacent to a gene in another species that expresses a protein under study. When this latter protein is generated, it is automatically attached to the bioluminescent indicator protein, thereby allowing the effects of the protein of interest to be optically tracked *in vivo*.

Electroluminescence. The emission of light resulting from the application of an electric field to a material. An important example is **injection electroluminescence**, which occurs when an electric current is injected into a forward-biased semiconductor p–n junction such as that used in a light-emitting diode. The recombination of electrons from the conduction band with holes from the valence band results in the emission of photons (Sec. 18.1A).

Photoluminescence. The emission of light by a sample following the absorption of optical photons. An example is the glow emitted by some materials after exposure to ultraviolet light. Photoluminescence, which is discussed in greater detail in the next section, is a useful tool for investigating the properties of semiconductor materials. It underlies the operation of white-light LEDs and many lasers. Photoluminescence is termed **radioluminescence** when the exciting photons are in the X-ray or gamma-ray region.

Fluorescence and Phosphorescence

Luminescence that appears within a brief time following excitation is also called **fluorescence**; typical fluorescence lifetimes lie in the range of picoseconds to microseconds. In the context of an organic material such as a dye molecule (Fig. 14.1-8), fluorescence arises when the radiative transitions are spin-allowed, i.e., when they take place between states with the same multiplicity (singlet$\rightarrow$singlet or triplet$\rightarrow$triplet). Luminescence that is delayed following excitation is also called **phosphorescence**; typical phosphorescence lifetimes are milliseconds or longer. For an organic material, phosphorescence arises when the radiative transitions are spin-forbidden (e.g., triplet $\rightarrow$ singlet).

B. Photoluminescence

Single-Photon Photoluminescence

Photoluminescence occurs when a system excited to a higher energy level by the absorption of a photon spontaneously decays to a lower energy level, emitting a photon in the process. To conserve energy, the emitted photon cannot have more energy than the exciting photon. Several examples of transitions that lead to photoluminescence are depicted in Fig. 14.5-2. Nonradiative downward transitions can participate in the process, as shown by the dashed vertical lines in Figs. 14.5-2(b) and (c). Intermediate downward nonradiative transitions, followed by upward nonradiative transitions, can also occur, as illustrated in Fig. 14.5-2(d). Photoluminescence occurs naturally in many materials, including inorganic molecules and crystals, noble gases, aromatic molecules, and semiconductors.

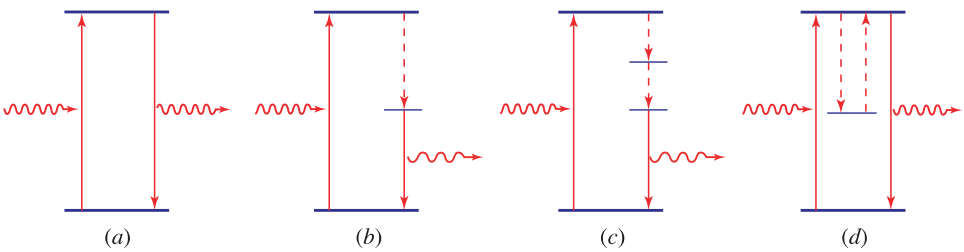


Figure 14.5-2 Single-photon photoluminescence from materials with different energy-level structures. Solid and dashed vertical lines represent radiative and nonradiative transitions, respectively.

Photoluminescence has many uses, including:

- The *reduction* of photon frequency, e.g., the conversion of an ultraviolet photon into a visible photon.
- The *conversion* of an ultraviolet photon into a pair of visible photons, a process known as **quantum cutting**.
- The *delay* of photon emission by storage of the excited electron in a long-lived intermediate state such as a trap.
- The *generation* of metameric white light by phosphor-conversion LEDs, wherein blue LED photons irradiate a yellow phosphor; the combination of blue and yellow appears white to the eye (Sec. 18.1F).
- The *use* as a seed for optically pumped lasers.

Multiphoton Photoluminescence

Photoluminescence can also occur when a system is excited to a higher energy level by the absorption of more than one photon, followed by the emission of a single photon via spontaneous emission to a lower energy level. The exciting photons can have the same, or different, energies and the emitted photon can have an energy greater than that of one of the exciting photons.

Multiphoton fluorescence microscopy. Two or more photons of the same energy can conspire to raise a material to a higher energy level, where it undergoes photoluminescence (fluorescence), as shown schematically in Figs. 14.5-3(a) and (b). Two-photon fluorescence, illustrated in Fig. 14.5-3(a), is the basis of an imaging technique known as **two-photon microscopy (2PM)** [originally called two-photon laser scanning microscopy (TPLSM)]. A photoluminescent compound, known as a **fluorophore**, is attached to specific locations in a biological sample, for example by chemistry, viral injection, or genetic engineering. A pair of photons (each of energy $h\nu_1$) arriving at the location of the fluorophore can be absorbed and result in the emission of a single fluorescence photon (of energy $h\nu_2 > h\nu_1$), thereby providing **structural imaging** of the fluorophore locations within the sample, as a function of time. The emission takes place within the lifetime of the fluorophore, which typically is in the range of nanoseconds. Making use of an activity-sensitive fluorophore allows **functional imaging** to be concomitantly carried out.

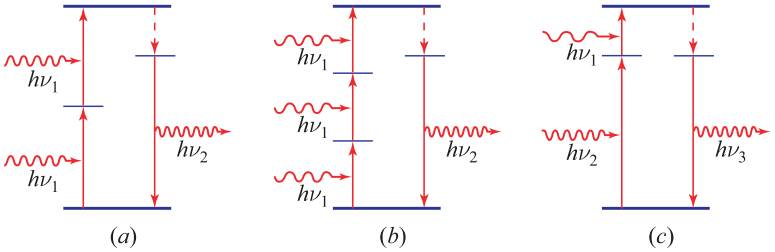


Figure 14.5-3 (a) Two-photon fluorescence. (b) Three-photon fluorescence. (c) Up-conversion fluorescence. Nonradiative relaxation (dashed vertical lines) is presumed to take part in all of these examples. Other scenarios are also possible.

The probability of observing two independently arriving photons at a given position and time is the square of observing a single such photon. Thus, by virtue of (13.1-15), the two-photon absorption rate at position $\mathbf{r}$ and time t, as well as the fluorescence-photon emission rate, behaves as a quadratic function of the incident intensity, i.e., it is proportional to $I^2(\mathbf{r}, t)$. An advantage of 2PM derives from this quadratic dependence:

a focused excitation beam results in absorption that is localized to the immediate vicinity of the focal point since two-photon absorption occurs preferentially at locations where the intensity is greatest. In comparison with ordinary (single-photon) fluorescence microscopy [Fig. 14.5-2(b)], the region from which fluorescence is observed is thereby sharpened, yielding enhanced resolution; this is also accompanied by a reduction of the background light arising from out-of-focus fluorescence. Another important advantage of 2PM in the domain of biology is the increased excitation wavelength, which penetrates more deeply into biological tissue. To ensure that the peak intensity is sufficiently high to engender two-photon absorption, and that the average intensity is sufficiently low to avoid damage to delicate tissue, the excitation is usually provided by a mode-locked laser that generates femtosecond-duration optical pulses with high peak power and low average power.

Multiphoton microscopy (MPM) operates in much the same way as two-photon microscopy, except that k independent photons, rather than two, conspire to effect each absorption, so that the fluorescence-photon emission rate varies as $I^k(\mathbf{r}, t)$. In particular, **three-photon microscopy (3PM)** is a valuable technique for carrying out *in-vivo*, noninvasive, high-resolution imaging in brain tissue. In comparison with 2PM, the $I^3(\mathbf{r}, t)$ behavior of the fluorescence-photon emission rate, together with the longer excitation wavelength [compare Figs. 14.5-3(a) and (b)], result in deeper penetration into brain tissue and improved performance. Three-photon microscopy requires optical pulses of sufficiently high energy, short duration, and long wavelength. Optimal wavelengths for brain-tissue imaging lie in the vicinity of 1300 nm for blue and green fluorophores and 1700 nm for orange and red fluorophores. Structural and functional imaging of populations of neurons deep within the intact mammalian brain has been effected by using a calcium-sensitive GFP fluorophore that is sensitive to neural activity, in conjunction with 1300-nm ultrafast pulses generated by a noncollinear optical parametric amplifier. Deep-brain 3PM imaging has also been conducted by making use of a red fluorophore excited by 1675-nm optical solitons generated in a photonic-crystal rod (Example 23.5-3).

3D multiphoton microlithography. A approach complementary to multiphoton fluorescence microscopy is useful for fabricating micro-objects. A lens delivers femtosecond-duration high-intensity optical pulses to a particular location in a specially designed transparent polymeric material. The intensity of the light is sufficient to effect multiphoton polymerization only in the vicinity of the focal region of the lens; its intensity before reaching the focal region is insufficient to polymerize the intervening material. Moving the focal point of the lens about allows any desired three-dimensional microstructure to be written. As an additional benefit, in practice the strong thresholding behavior of the polymerization nonlinearity serves to further increase the resolution of the microstructure (see, e.g., Sec. 7.3B).

Up-conversion fluorescence. Multiphoton photoluminescence can also take place when the two photons that conspire to excite the system have different energies, as illustrated in Fig. 14.5-3(c). This scheme is useful for converting infrared photons to visible ones. An infrared photon of low energy ($h\nu_1$) teams up with a more energetic auxiliary photon ($h\nu_2$) to excite a system such as a single ion or atom, which then produces a luminescence photon at or near the sum energy ($h\nu_3 = h\nu_1 + h\nu_2$).

Up-conversion fluorescence via sequential absorption can be observed most easily in materials containing traps that can store the electron elevated by the first photon for a time sufficient for the second photon to arrive and boost the system to its upper state. Phosphors doped with rare-earth ions such as Er^{3+} are often used. In some materials, the traps can be charged to their intermediate state in minutes by exposing the material to daylight or fluorescent light, thereby providing the auxiliary photons

of energy $h\nu_2$. The arrival of an infrared signal photon of energy $h\nu_1$ then releases an electron from the trap, which results in the emission of a visible luminescence photon with an energy at or near $h(\nu_1 + \nu_2)$. Up-conversion fluorescence can also occur via more complex processes, such as collective emission from two nearby ions that have both been excited.

A practical up-conversion-fluorescence device used in the laboratory often takes the form of a small reflective or transmissive card with an active area of about 5 cm $\times$ 5 cm, known as an **infrared sensor card**. Upconverting powder is laminated between a pair of stiff transparent plastic sheets to form the card. Upconverting powder can also be dispersed in a block of polymer for three-dimensional viewing. Though the conversion efficiency of these devices is typically quite small, they are nevertheless useful for visually viewing the spatial distribution of an infrared beam, such as that produced by an infrared laser. The relative infrared detection sensitivity and the visible emission spectral intensity of a commercially available card are portrayed in Fig. 14.5-4.

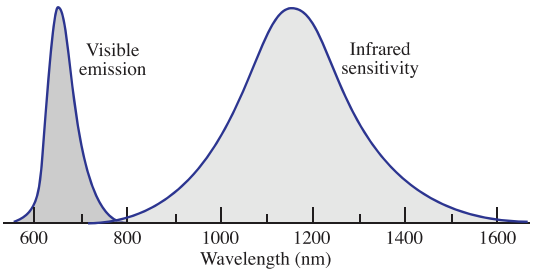


Figure 14.5-4 Relative infrared detection sensitivity and relative visible spectral intensity for up-conversion fluorescence from a commercially available infrared sensor card.

C. Scattering

Photoluminescence, as considered in Sec. 14.5B, involves the resonant absorption of a photon via a transition between the ground state and a real excited state; the subsequent relaxation of the excited state back to the ground state results in the emission of a luminescence photon. Absorption and subsequent re-emission from a real upper state are the defining characteristics of luminescence, fluorescence, and phosphorescence.

Light scattering processes can involve transitions that occur via virtual states. Since they are often nonresonant interactions, light can be scattered over a broad range of frequencies. We consider in turn three scattering processes of importance in optics and photonics, as portrayed in Fig. 14.5-5: Rayleigh, Raman, and Brillouin scattering. Scattering is inherent and unavoidable, and usually undesirable, but it also proves useful for providing information about the characteristics of the scattering medium and for creating useful sources of light.

Rayleigh Scattering

Rayleigh scattering is a process whereby a medium causes an incident photon to change direction. It entails an energy-conserving (elastic) interaction so that the scattered photon has the same energy as the incident photon, as schematized in Fig. 14.5-5(a). Rayleigh scattering occurs in gases, liquids, and solids. It is engendered by variations in a medium that are finer than the wavelength of light, such as random density fluctuations in air or random refractive-index inhomogeneities in glass (Sec. 10.3A). It can also be brought about by the presence of particles whose sizes are much smaller than the wavelength of light, such as electrons, atoms, molecules, or nanoparticles. As discussed in Sec. 5.6B, the scattered intensity is proportional to ν^4, and thus to $1/\lambda_o^4$, where ν and λ_o are the frequency and wavelength of the illumination, respectively. Short wavelengths thus undergo greater scattering than long wavelengths; Rayleigh

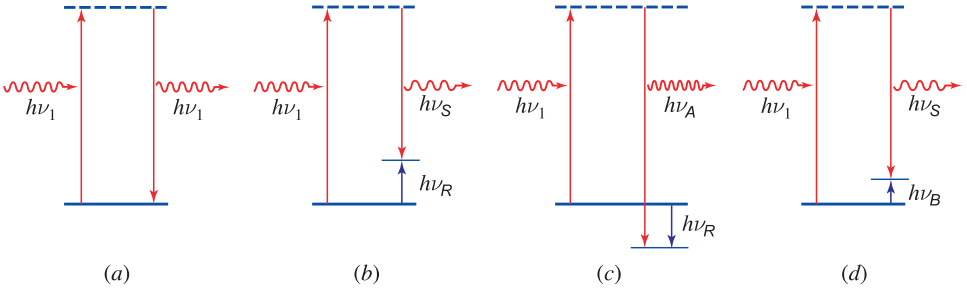


Figure 14.5-5 Several forms of light scattering: (*a*) Rayleigh; (*b*) Raman (Stokes); (*c*) Raman (anti-Stokes); and (*d*) Brillouin. Dashed horizontal lines indicate virtual states and therefore nonresonant scattering.

scattering is responsible for the blue color of the sky. Scattering from spherical particles larger than $\approx \lambda_o/10$ is known as **Mie scattering**; its strength does not depend strongly on the wavelength of the illumination. Mie scattering is responsible, as an example, for the white glare around a light source in the presence of mist or fog (Sec. 5.6C).

Raman Scattering

Raman scattering is a process by means of which a photon of frequency $h\nu_1$, following an inelastic interaction with a material, emerges either at a lower frequency $h\nu_S = h\nu_1 - h\nu_R$ (Stokes scattering) or at a higher frequency $h\nu_A = h\nu_1 + h\nu_R$ (anti-Stokes scattering), as displayed in Figs. 14.5-5(*b*) and (*c*), respectively. Raman scattering occurs in gases, liquids, and solids. Unlike Rayleigh scattering, Raman scattering is inelastic; the alteration of photon frequency is brought about by an exchange of energy $h\nu_R$ with a rotational and/or vibrational mode of a molecule or solid. In **Stokes scattering**, the photon imparts energy to the material system, whereas the reverse occurs in **anti-Stokes scattering**. In general, the spectrum of light scattered from a material contains a Rayleigh-scattered component, at the incident frequency, together with red-shifted and blue-shifted sidebands corresponding to inelastically scattered Stokes and anti-Stokes components, respectively. Though the sideband power is typically weak for nonresonant interactions, lying a factor of 10^{-7} below that of the incident light, **Raman scattering** is useful for characterizing materials. In crystalline materials, the vibrational spectrum is generally discrete and the Raman lines are narrow. Glasses, in contrast, have broad vibrational spectra that in turn give rise to broad Raman spectra. **Brillouin scattering**, portrayed in Fig. 14.5-5(*d*), is similar to Raman scattering except that the exchange of energy $h\nu_B$ takes place with acoustic, rather than vibrational, modes of the medium.

Stimulated Raman Scattering

Stimulated Raman scattering (**SRS**) can take place when a signal photon enters a nonlinear optical medium together with a pump photon of higher frequency (inset in Fig. 15.3-7). The signal photon stimulates the emission of a second signal photon, which is obtained by Stokes-shifting the pump photon so that its frequency precisely matches that of the input signal photon. The surplus energy of the pump photon is transferred to the vibrational modes of the medium. The process bears some similarity to stimulated emission, but the Raman interaction is a third-order nonlinear optical process (Sec. 22.3B).

Stimulated Raman scattering is useful for making optical amplifiers (Sec. 15.3D) and lasers (Sec. 16.3C). Raman amplifiers and Raman lasers have the distinct merit that the bandwidth over which they can operate is governed by the vibrational spectrum of the material rather than by the linewidth of a transition. The vibrational spectrum of

glass is particularly broad, so that a length of glass optical fiber can serve as a fiber amplifier or fiber laser that is tunable over a range of hundreds of nm. Raman optical amplifiers and Raman fiber lasers are used in dense wavelength-division-multiplexed optical fiber communication systems (Sec. 25.3C).

Stimulated Raman scattering is also useful as a spectroscopic tool since it can reveal the underlying vibrational characteristics of a material. The sensitivity of Raman-based spectroscopy can be enhanced by making use of **coherent anti-Stokes Raman scattering** (**CARS**), which uses two pump lasers whose frequency difference is resonant with the vibrational frequency of the material under investigation, thereby increasing the efficiency of wave mixing.

In another important application, Raman processes are useful for generating broadband light in optical fibers. A pump gives rise to Raman-scattered spontaneous emission that is amplified via stimulated Raman scattering as the light and pump propagate through the fiber. For an optical fiber of sufficient length, and a pump of sufficient strength, the resulting Raman spectrum initiates yet further Raman frequency conversion, resulting in the production of broadband (supercontinuum) light (Sec. 23.5C). Stabilization of the process can be achieved by making use of a resonator. **Stimulated Brillouin scattering** is similar to SRS except that acoustic vibrations, rather than molecular vibrations, are involved.

READING LIST

Quantum Mechanics

See also the reading list in Chapter 13.

P. R. Berman, *Introductory Quantum Mechanics: A Traditional Approach Emphasizing Connections with Classical Physics*, Springer-Verlag, 2018.

M. G. Raymer, *Quantum Physics: What Everyone Needs to Know*, Oxford University Press, 2017.

J. Greensite, *An Introduction to Quantum Theory*, IOP Publishing, 2017.

S. Weinberg, *Lectures on Quantum Mechanics*, Cambridge University Press, 2nd ed. 2015.

L. E. Ballentine, *Quantum Mechanics: A Modern Development*, World Scientific, 2nd ed. 2015.

L. Susskind and A. Friedman, *Quantum Mechanics: The Theoretical Minimum*, Basic/Perseus, 2014.

E. D. Commins, *Quantum Mechanics: An Experimentalist's Approach*, Cambridge University Press, 2014.

D. A. B. Miller, *Quantum Mechanics for Scientists and Engineers*, Cambridge University Press, 2008.

L. D. Landau and E. M. Lifshitz, *Quantum Mechanics*, Addison–Wesley, 1958.

P. A. M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, 4th ed. 1958.

Atomic, Molecular, and Solid-State Physics

See also the reading lists in Chapters 15–18.

M. Dresselhaus, G. Dresselhaus, S. B. Cronin, and A. G. S. Filho, *Solid State Properties: From Bulk to Nano*, Springer-Verlag, 2018.

H. Friedrich, *Theoretical Atomic Physics*, Springer-Verlag, 4th ed. 2017.

J. B. Ketterson, *The Physics of Solids*, Oxford University Press, 2016.

P. Hofmann, *Solid State Physics: An Introduction*, Wiley–VCH, 2nd ed. 2015.

J. Lucas, P. Lucas, T. Le Mercier, A. Rollat, and W. Davenport, *Rare Earths: Science, Technology, Production and Use*, Elsevier, 2015.

C. Kittel, *Introduction to Solid State Physics*, Wiley, 8th ed. 2012.

C. Cohen-Tannoudji and D. Guéry-Odelin, *Advances In Atomic Physics: An Overview*, World Scientific, 2011.

P. W. Atkins and R. S. Friedman, *Molecular Quantum Mechanics*, Oxford Univ. Press, 5th ed. 2011.

D. Budker, D. F. Kimball, and D. P. DeMille, *Atomic Physics: An Exploration Through Problems and Solutions*, Oxford University Press, 2nd ed. 2008.

M. Schwoerer and H. C. Wolf, *Organic Molecular Solids*, Wiley–VCH, 2007.

W. Demtröder, *Molecular Physics: Theoretical Principles and Experimental Methods*, Wiley–VCH, 2005.

M. Born, *Atomic Physics*, Blackie & Son, 1935, 8th ed. 1969; Dover, reissued 1989.

D. ter Haar, *The Old Quantum Theory*, Pergamon, 1967 [contains English translations of key early papers by Planck, Einstein, Rutherford, and Bohr].

E. U. Condon and G. H. Shortley, *The Theory of Atomic Spectra*, Cambridge University Press, 1935.

Interaction of Radiation and Matter

See also the reading lists in Chapters 15–18.

J. Weiner and F. Nunes, *Light-Matter Interaction: Physics and Engineering at the Nanoscale*, Oxford University Press, 2nd ed. 2017.

P. van der Straten and H. J. Metcalf, *Atoms and Molecules Interacting with Light: Atomic Physics for the Laser Era*, Cambridge University Press, 2016.

G. Grynberg, A. Aspect, and C. Fabre, *Introduction to Quantum Optics: From the Semi-Classical Approach to Quantized Light*, Cambridge University Press, 2010.

M. Planck, *Planck's Columbia Lectures: Abridged and Unabridged Versions*, with commentary by W. Vlasak, Adaptive Enterprises, 2005.

J. H. van Vleck and D. L. Huber, Absorption, Emission, and Linebreadths: A Semihistorical Perspective, *Reviews of Modern Physics*, vol. 49, pp. 939–959, 1977.

E. M. Purcell, Spontaneous Emission Probabilities at Radio Frequencies, *Proceedings of the American Physical Society*, Cambridge, MA, April 25–27, 1946, Abstract B10, (*Physical Review*, vol. 69, p. 681, June 1946).

M. Göppert-Mayer, Über Elementarakte mit zwei Quantensprüngen, *Annalen der Physik*, vol. 9, pp. 273–294, 1931.

A. Einstein, Zur Quantentheorie der Strahlung, *Physikalische Zeitschrift*, vol. 18, pp. 121–128, 1917 [Translation: On the Quantum Theory of Radiation, in D. ter Haar, *The Old Quantum Theory*, Pergamon, 1967].

Laser Cooling and Trapping, Atom Optics, and Bose–Einstein Condensates

G. E. Marti, R. B. Hutson, A. Goban, S. L. Campbell, N. Poli, and J. Ye, Imaging Optical Frequencies with 100 μHz Precision and 1.1 μm Resolution, *Physical Review Letters*, vol. 120, 103201, 2018.

C. W. Gardiner and P. Zoller, *The Quantum World of Ultra-Cold Atoms and Light. Book III: Ultra-Cold Atoms*, World Scientific, 2017.

J. L. Bohn, A. M. Ray, and J. Ye, Cold Molecules: Progress in Quantum Engineering of Chemistry and Quantum Matter, *Science*, vol. 357, pp. 1002–1010, 2017.

P. H. Jones, O. M. Maragò, and G. Volpe, *Optical Tweezers: Principles and Applications*, Cambridge University Press, 2015.

M. J. Padgett, J. Molloy, and D. McGloin, eds., *Optical Tweezers: Methods and Applications*, CRC Press/Taylor & Francis, 2010.

A. D. Cronin, J. Schmiedmayer, and D. E. Pritchard, Optics and Interferometry with Atoms and Molecules, *Reviews of Modern Physics*, vol. 81, pp. 1051–1129, 2009.

V. Letokhov, *Laser Control of Atoms and Molecules*, Oxford University Press, 2007.

A. Ashkin, *Optical Trapping and Manipulation of Neutral Particles Using Lasers: A Reprint Volume with Commentaries*, World Scientific, 2006.

F. Bardou, J.-P. Bouchaud, A. Aspect, and C. Cohen-Tannoudji, *Lévy Statistics and Laser Cooling: How Rare Events Bring Atoms to Rest*, Cambridge University Press, 2002.

W. Ketterle, Nobel Lecture: When Atoms Behave as Waves: Bose–Einstein Condensation and the Atom Laser, *Reviews of Modern Physics*, vol. 74, pp. 1131–1151, 2002.

E. A. Cornell and C. E. Wieman, Nobel Lecture: Bose–Einstein Condensation in a Dilute Gas, the First 70 Years and Some Recent Experiments, *Reviews of Modern Physics*, vol. 74, pp. 875–893, 2002.

P. Meystre, *Atom Optics*, Springer-Verlag, 2001.

H. J. Metcalf and P. van der Straten, *Laser Cooling and Trapping*, Springer-Verlag, 1999.

W. D. Phillips, Laser Cooling and Trapping of Neutral Atoms, *Reviews of Modern Physics*, vol. 70, pp. 721–741, 1998.

C. N. Cohen-Tannoudji, Manipulating Atoms with Photons, *Reviews of Modern Physics*, vol. 70, pp. 707–719, 1998.

S. Chu, The Manipulation of Neutral Particles, *Reviews of Modern Physics*, vol. 70, pp. 685–706, 1998.

A. Einstein, Quantentheorie des einatomigen idealen Gases, *Sitzungberichte der Preussischen Akademie der Wissenschaften*, vol. XXII, pp. 261–267, 1924; vol. I, pp. 3–14, 1925.

Thermal Physics, Thermography, and Statistical Physics

J. P. Casquilho and P. I. C. Teixera, *Introduction to Statistical Physics*, Cambridge University Press, 2015.

H. J. W. Müller-Kirsten, *Basics of Statistical Physics*, World Scientific, 2nd ed. 2013.

M. J. R. Hoch, *Statistical and Thermal Physics: An Introduction*, CRC Press/Taylor & Francis, 2011.

A. L. Wasserman, *Thermal Physics: Concepts and Practice*, Cambridge University Press, 2011.

M. Vollmer and K.-P. Möllmann, *Infrared Thermal Imaging: Fundamentals, Research and Applications*, Wiley–VCH, 2010.

L. E. Reichl, *A Modern Course in Statistical Physics*, Wiley–VCH, 3rd ed. 2009.

C. Kittel, *Elementary Statistical Physics*, Wiley, 1958; Dover, reissued 2004.

Luminescence, Multiphoton Fluorescence, and Scattering

K. Yasui, *Acoustic Cavitation and Bubble Dynamics*, Springer-Verlag, 2018.

S. Martini and S. H. D. Haddock, Quantification of Bioluminescence from the Surface to the Deep Sea Demonstrates Its Predominance as an Ecological Trait, *Scientific Reports*, vol. 7, 45750, 2017.

U. Kubitscheck, ed., *Fluorescence Microscopy*, Wiley–VCH, 2nd ed. 2017.

D. G. Ouzounov, T. Wang, M. Wang, D. D. Feng, N. G. Horton, J. C. Cruz-Hernández, Y.-T. Cheng, J. Reimer, A. S. Tolias, N. Nishimura, and C. Xu, *In Vivo* Three-Photon Imaging of Activity of GCaMP6-Labeled Neurons Deep in Intact Mouse Brain, *Nature Methods*, vol. 14, pp. 388–390, 2017.

P. Lecoq, A. Gektin, and M. Korzhik, *Inorganic Scintillators for Detector Systems: Physical Principles and Crystal Engineering*, Springer-Verlag, 2nd ed. 2016.

M. Gaft, R. Reisfeld, and G. Panczer, *Modern Luminescence Spectroscopy of Minerals and Materials*, Springer-Verlag, 2nd ed. 2015.

X. Chen, Y. Liu, and D. Tu, *Lanthanide-Doped Luminescent Nanomaterials: From Fundamentals to Bioapplications*, Springer-Verlag, 2014.

D. M. Jameson, *Introduction to Fluorescence*, CRC Press/Taylor & Francis, 2014.

K. Wang, N. G. Horton, and C. Xu, Going Deep: Brain Imaging with Multiphoton Microscopy, *Optics & Photonics News*, vol. 24, no. 11, pp. 32–39, 2013.

N. G. Horton, K. Wang, D. Kobat, C. G. Clark, F. W. Wise, C. B. Schaffer, and C. Xu, *In Vivo* Three-Photon Microscopy of Subcortical Structures within an Intact Mouse Brain, *Nature Photonics*, vol. 7, pp. 205–209, 2013.

O. Shimomura, *Bioluminescence*, World Scientific, revised ed. 2012.

J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*, Springer-Verlag, 3rd ed. 2006.

B. R. Masters, *Confocal Microscopy and Multiphoton Excitation Microscopy: The Genesis of Live Cell Imaging*, SPIE Optical Engineering Press, 2006.

B. R. Masters, ed., *Selected Papers on Multiphoton Excitation Microscopy*, SPIE Optical Engineering Press (Milestone Series Volume 175), 2003.

D. A. Long, *The Raman Effect: A Unified Treatment of the Theory of Raman Scattering by Molecules*, Wiley, 2002.

M. J. Damzen, V. I. Vlad, V. Babin, and A. Mocofanescu, *Stimulated Brillouin Scattering: Fundamentals and Applications*, Institute of Physics, 2002.

B. H. Cumpston, S. P. Ananthavel, S. Barlow, D. L. Dyer, J. E. Ehrlich, L. L. Erskine, A. A. Heikal, S. M. Kuebler, I.-Y. S. Lee, D. McCord-Maughon, J. Qin, H. Röckel, M. Rumi, X.-L. Wu, S. R. Marder, and J. W. Perry, Two-Photon Polymerization Initiators for Three-Dimensional Optical Data Storage and Microfabrication, *Nature*, vol. 398, pp. 51–54, 1999.

M. J. Weber, ed., *Selected Papers on Phosphors, Light Emitting Diodes, and Scintillators: Applications of Photoluminescence, Cathodoluminescence, Electroluminescence, and Radioluminescence*, SPIE Optical Engineering Press (Milestone Series Volume 151), 1998.

W. Denk, J. H. Strickler, and W. W. Webb, Two-Photon Laser Scanning Fluorescence Microscopy, *Science*, vol. 248, pp. 73–76, 1990.

B. J. Berne and R. Pecora, *Dynamic Light Scattering: With Applications to Chemistry, Biology, and Physics*, Wiley, 1976; Dover, reissued 2000.

PROBLEMS

14.3-3 **Comparison of Stimulated and Spontaneous Emission.** An atom with two energy levels corresponding to a transition with characteristics $\lambda_o = 0.7\ \mu\text{m}$, $t_{\text{sp}} = 3\ \text{ms}$, $\Delta\nu = 50\ \text{GHz}$, and Lorentzian lineshape, is placed in a resonator of volume $V = 100\ \text{cm}^3$ and refractive index $n = 1$. Two radiation modes (one at the center frequency ν_0 and the other at $\nu_0 + \Delta\nu$) are each excited with 1000 photons. Determine the probability density for stimulated emission (or absorption). If N_2 such atoms are excited to energy level 2, determine the time constant for the decay of N_2 due to stimulated *and* spontaneous emission. Find the number of photons that should be present (instead of 1000) in order for the decay rate due to stimulated emission to equal that due to spontaneous emission?

14.3-4 **Spontaneous Emission into Prescribed Modes.**

(a) Consider a 1-μm^3 cubic cavity containing a medium of refractive index $n = 1$. Consulting Sec. 11.4, determine the mode numbers (q_1, q_2, q_3) of the lowest- and next-higher-frequency modes. Show that these frequencies are 260 and 367 THz, respectively.

(b) Now consider a single excited atom in the cavity when it contains zero photons. Let p_{sp1} be the probability density (s^{-1}) that the atom spontaneously emits a photon into the $(2, 1, 1)$ mode, and let p_{sp2} be the probability density that the atom spontaneously emits a photon with frequency 367 THz. Determine the ratio $p_{\text{sp2}}/p_{\text{sp1}}$.

14.4-2 **Rate Equations for Broadband Radiation.** A resonator of unit volume contains atoms with two energy levels whose associated transition has resonance frequency ν_0 and linewidth $\Delta\nu$. There are N_1 and N_2 atoms in the lower and upper levels, 1 and 2, respectively, and a total of $\overline{n}$ photons in each of the modes within a broad frequency band surrounding ν_0. Photons are lost from the resonator at a rate $1/\tau_p$ as a result of imperfect reflection at the cavity walls. Assuming that there are no nonradiative transitions between levels 2 and 1, write the rate equations for N_2 and $\overline{n}$.

14.4-3 **Inhibited Spontaneous Emission.** Consider a two-dimensional blackbody radiator (e.g., a square plate of area A) in thermal equilibrium at temperature T.

(a) Determine the density of modes $M(\nu)$ and the spectral energy density (i.e., the energy in the frequency range between ν and $\nu + d\nu$ per unit area) of the emitted radiation $\varrho(\nu)$ (see Sec. 11.3).

(b) Find the probability density of spontaneous emission P_{sp} for an atom located in a cavity that permits radiation only in two dimensions. Such a cavity may be made, for example, by using photonic-crystal omnidirectional reflectors above and below a slab.

14.4-4 **Comparison of Stimulated and Spontaneous Emission in Blackbody Radiation.** Find the temperature of a thermal-equilibrium blackbody cavity emitting a spectral energy density $\varrho(\nu)$ when the rates of stimulated and spontaneous emission from the atoms in the cavity walls are equal at $\lambda_o = 1\ \mu\text{m}$.

14.4-5 **Wien's Law.** Show that the wavelength spectral energy density $\varrho_\lambda(\lambda)$ [$\varrho_\lambda(\lambda)\,d\lambda$ is the energy per unit volume in the wavelength region between λ and $\lambda + d\lambda$] is given by

$$\varrho_\lambda(\lambda) = \frac{8\pi hc}{\lambda^5}\,\frac{1}{\exp(hc/\lambda kT) - 1}\,,$$

so that $\varrho_\lambda(\lambda)/\varrho_\nu(\nu) = c/\lambda^2$. Show also that the wavelength λ_{p} at which the spectral energy density is maximum satisfies the equation $5(1 - e^{-y}) = y$, where $y = hc/\lambda_{\text{p}}kT$, demonstrating the validity of the relationship $\lambda_{\text{p}}T = \text{constant}$ (Wien's law) is satisfied.

Find $\lambda_p T$ approximately. Show that $\lambda_p \neq c/\nu_p$, where ν_p is the frequency at which the blackbody energy density $\varrho(\nu)$ is maximum (Exercise 14.4-1). The shapes and peak locations of density functions are dependent on the representation chosen.

14.4-6 **Spectral Energy Density of One-Dimensional Blackbody Radiation.** Consider a one-dimensional blackbody radiator of length L in thermal equilibrium at temperature T.

(a) Determine the density of modes $M(\nu)$ (number of modes per unit frequency per unit length) in one dimension.

(b) Using the average density $\overline{E}$ of a mode of frequency ν, determine the spectral energy density (i.e., the energy in the frequency range between ν and $\nu + d\nu$ per unit length) of the blackbody radiation $\varrho(\nu)$. Sketch $\varrho(\nu)$ versus ν.

14.4-7 **Stefan–Boltzmann Law.** Use the conventional expression for the spectral energy density for blackbody radiation provided in (14.4-9) to confirm that the total power radiated is proportional to T^4, in accord with the Stefan–Boltzmann law. Determine the proportionality constant. *Hint:* $\int_0^\infty dx\, x^3/(e^x - 1) = \pi^4/15$.

*14.5-1 **Statistics of Cathodoluminescence Light.** Consider a beam of electrons impinging on the phosphor of a cathode-ray tube. Let $\overline{m}$ be the mean number of electrons striking a unit area of the phosphor in unit time. If the number m of electrons arriving in a fixed time is random with a Poisson distribution, and the number of photons emitted per electron is also Poisson distributed but with mean $\overline{G}$, find the overall distribution $p(n)$ of the emitted cathodoluminescence photons. The result is known as the **Neyman Type-A distribution**.[†] Determine expressions for the mean $\overline{n}$ and the variance σ_n^2. *Hint:* Use conditional probability.

[†] See M. C. Teich, Role of the Doubly Stochastic Neyman Type-A and Thomas Counting Distributions in Photon Detection, *Applied Optics*, vol. 20, pp. 2457–2467, 1981.

CHAPTER

15

LASER AMPLIFIERS

15.1 THEORY OF LASER AMPLIFICATION 622
- A. Gain and Bandwidth
- B. Phase Shift

15.2 AMPLIFIER PUMPING 626
- A. Rate Equations
- B. Pumping Schemes

15.3 REPRESENTATIVE LASER AMPLIFIERS 636
- A. Ruby
- B. Neodymium-Doped Glass
- C. Erbium-Doped Silica Fiber
- D. Raman Fiber Amplifiers
- E. Tabulation of Selected Laser Transitions

15.4 AMPLIFIER NONLINEARITY 645
- A. Saturated Gain in Homogeneously Broadened Media
- *B. Saturated Gain in Inhomogeneously Broadened Media

*15.5 AMPLIFIER NOISE 651

Charles H. Townes (1915–2015)

Nikolai G. Basov (1922–2001)

Aleksandr M. Prokhorov (1916–2002)

Townes, Basov, and Prokhorov developed the principle of **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation (**LASER**). They received the Nobel Prize for this work in 1964.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

A **coherent optical amplifier** is a device that increases the amplitude of an optical field while maintaining its phase. If the optical field at the input to such an amplifier is monochromatic, the output will also be monochromatic with the same frequency. The output amplitude is increased relative to the input while the phase remains unchanged or is shifted by a fixed amount. In contrast, an **incoherent optical amplifier** increases the intensity of an optical wave without preserving its phase. Coherent optical amplifiers play an important role in applications ranging from the amplification of weak optical pulses, such as those that have traveled through a long length of optical fiber, to the generation of high-intensity optical pulses, such as those required to achieve laser fusion. Understanding the operation of optical amplifiers serves as a prelude to understanding the operation of optical oscillators, considered in Chapter 16.

The underlying principle for achieving the coherent amplification of light is **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation, which forms the acronym **LASER**. Stimulated emission (Sec. 14.3) allows a photon in a given mode to induce an atom with an electron in an upper energy level to undergo a transition to a lower energy level and, in the process, to emit a clone photon into the same mode as the initial photon. A clone photon has the same frequency, direction, and polarization as the initial photon. These two photons in turn serve to stimulate the emission of two additional photons, and so on, while preserving these properties. The result is **laser amplification**. Because stimulated emission can occur only when the photon energy is nearly equal to the energy difference between the upper and lower energy levels, the light-amplification process is restricted to a band of frequencies determined by the atomic-transition linewidth. Though the presentation throughout this chapter is couched in terms of "atoms" and "atomic energy levels," these appellations are to be more broadly understood as "active medium" and "laser energy levels," respectively.

Light transmitted through matter in thermal equilibrium is attenuated. This is because absorption by a large population of atoms in the lower energy level is more prevalent than stimulated emission by the smaller population of atoms in the upper level. An essential ingredient for achieving laser amplification is the presence of a greater number of atoms in the upper energy level than in the lower level, a nonequilibrium situation (Sec. 14.2). Achieving such a population inversion requires a source of power, called a **pump**, that excites the atoms from the lower to the higher energy level, as illustrated in Fig. 15.0-1.

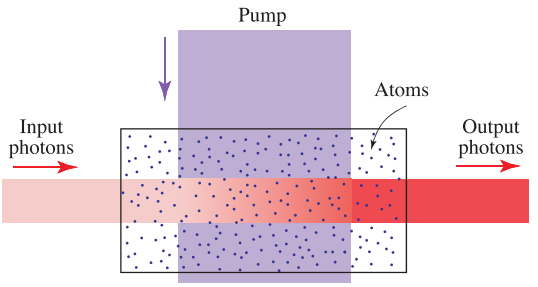


Figure 15.0-1 The laser amplifier. An external power source (the pump) excites the active medium (represented by a collection of atoms) so that it has a population inversion. Photons interact with the atoms. When stimulated emission is more prevalent than absorption, the medium acts as a coherent laser amplifier.

Laser amplification differs in a number of respects from electronic amplification. Electronic amplifiers rely on devices in which small changes of an injected electric current (or applied voltage) result in large changes in the rate of flow of charge carriers, such as electrons and holes in a semiconductor field-effect transistor (FET) or bipolar junction transistor. Tuned electronic amplifiers make use of a resonant circuit or resonator (e.g., a capacitor/inductor or a metal cavity — see Fig. 11.0-2) to limit the gain of the amplifier to the band of frequencies of interest. In contrast, atomic,

molecular, ionic, and solid-state laser amplifiers rely on differences in their allowed energy levels to provide the principal frequency-selection mechanism. These entities act as natural resonators that select the frequency of operation and bandwidth of the amplifier. Optical resonators are often used to provide auxiliary frequency tuning.

The properties of an ideal optical or electronic coherent amplifier are illustrated schematically in Fig. 15.0-2(*a*). It comprises a linear system that increases the amplitude of the input signal by a fixed factor, the amplifier gain. A sinusoidal input leads to a sinusoidal output at the same frequency, but with a larger amplitude. The gain of the ideal amplifier is constant for all frequencies within the amplifier spectral bandwidth. The amplifier may impart to the input signal a phase shift that varies linearly with frequency, corresponding to a time delay at the output with respect to the input (Sec. B.1 of Appendix B).

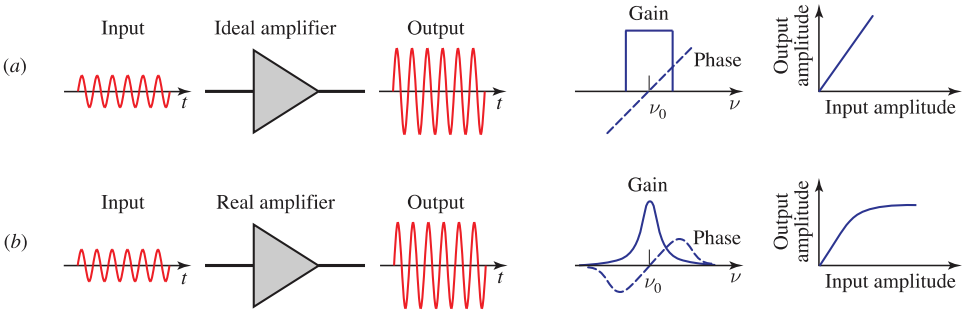


Figure 15.0-2 (*a*) An ideal amplifier is linear. It serves to increase the amplitude of a signal (whose frequencies lie within its bandwidth) by a constant gain factor, and possibly introduces a linear phase shift. (*b*) A real amplifier typically has a gain and phase shift that are functions of frequency. For large values of the input, the output signal saturates and the real amplifier exhibits nonlinearity.

A real coherent amplifier, on the other hand, delivers a gain and phase shift that are frequency-dependent, typically in the manner illustrated in Fig. 15.0-2(*b*). The gain and phase shift determine the transfer function of the amplifier. For a sufficiently large input amplitude, a real amplifier generally exhibits saturation, a form of nonlinear behavior in which the output amplitude does not increase in proportion to the input amplitude. Saturation introduces harmonic components into the output, provided that the amplifier bandwidth is sufficiently broad to allow them to pass. A real amplifier also introduces noise, so that a random fluctuating component is present at the output, regardless of the input.

An amplifier is thus characterized by the following features:

- Gain
- Bandwidth
- Phase shift
- Power source
- Nonlinearity and gain saturation
- Noise

This Chapter

In this chapter, we discuss the features listed above in turn. In Sec. 15.1 the theory of laser amplification is developed, and expressions for the amplifier gain, spectral bandwidth, and phase shift are obtained. The mechanisms by means of which a power source pumps the active medium and achieves a population inversion are examined in Sec. 15.2. A number of representative laser amplifiers are considered in Sec. 15.3. Sections 15.4 and 15.5 are devoted to nonlinearity and noise in the amplification process, respectively. The material in this chapter relies on the exposition of the interaction of photons with atoms set forth in Sec. 14.3.

15.1 THEORY OF LASER AMPLIFICATION

A monochromatic optical plane wave traveling in the z direction with frequency ν, electric field $\mathcal{E}(z) = \text{Re}\{E(z)\exp(j2\pi\nu t)\}$, complex amplitude $E(z)$, intensity $I(z) = |E(z)|^2/2\eta$, and photon-flux density $\phi(z) = I(z)/h\nu$ (photons per second per unit area) will interact with an atomic medium, provided that the atoms of the medium have two energy levels whose energy difference nearly matches the photon energy $h\nu$. The numbers of atoms per unit volume in the lower and upper energy levels are denoted N_1 and N_2, respectively. The wave is amplified with a gain coefficient $\gamma(\nu)$ (per unit length) and undergoes a phase shift $\varphi(\nu)$ (per unit length). We proceed to determine expressions for $\gamma(\nu)$ and $\varphi(\nu)$. Positive $\gamma(\nu)$ corresponds to amplification; negative $\gamma(\nu)$ corresponds to attenuation.

A. Gain and Bandwidth

In accordance with Sec. 14.3, three forms of photon–atom interaction take place. If the atom is in the lower energy level, the photon may be absorbed. If it is in the upper energy level, a clone photon may be emitted by the process of stimulated emission. These two processes lead to attenuation and amplification, respectively. The third form of interaction, spontaneous emission, in which an atom in the upper energy level emits a photon independently of the presence of other photons, is responsible for amplifier noise (Sec. 15.5).

The probability density (s^{-1}) that an unexcited atom absorbs a single photon is, according to (14.3-19) and (14.3-15),

$$W_i = \phi\,\sigma(\nu), \tag{15.1-1}$$

where $\sigma(\nu) = (\lambda^2/8\pi t_{\text{sp}})\,g(\nu)$ is the transition cross section at the frequency ν, $g(\nu)$ is the normalized lineshape function, t_{sp} is the effective spontaneous lifetime for stimulated emission, and λ is the wavelength of light in the medium. The probability density for stimulated emission is the same as that for absorption.

Gain Coefficient

The average density of absorbed photons (number of photons per unit time per unit volume) is N_1W_i. Similarly, the average density of clone photons generated as a result of stimulated emission is N_2W_i. The net number of photons gained per second per unit volume is therefore NW_i, where $N = N_2 - N_1$ is the population density difference, which is often simply referred to as the population difference. If N is positive, a **population inversion** exists, in which case the medium can act as an amplifier and the photon-flux density of a wave passing through the medium can increase. If N is negative, the medium acts as an attenuator and the photon-flux density decreases. If $N = 0$, the medium is transparent.

Since the incident photons travel in the z direction, the stimulated-emission photons also travel in that direction, as illustrated in Fig. 15.1-1. An external pump providing a population inversion ($N > 0$) then causes the photon-flux density $\phi(z)$ to increase with z. Because emitted photons stimulate further emissions, the growth at any position z is proportional to the population at that position; $\phi(z)$ thus increases exponentially.

To demonstrate this process explicitly, consider an incremental cylinder of length dz and unit area, as shown in Fig. 15.1-1. If $\phi(z)$ and $\phi(z) + d\phi(z)$ are the photon-flux densities entering and exiting the incremental cylinder, respectively, then $d\phi(z)$ must be the photon-flux density emitted from within the cylinder. This incremental number of photons per unit area per unit time, $d\phi(z)$, is simply the number of photons gained

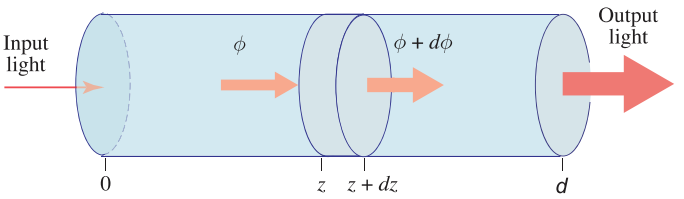


Figure 15.1-1 The photon-flux density ϕ (photons/cm^2-s) entering an incremental cylinder containing excited atoms grows to $\phi + d\phi$ after traveling a distance dz.

per unit time per unit volume, NW_i, multiplied by the thickness of the cylinder dz:

$$d\phi = NW_i\, dz. \tag{15.1-2}$$

With the help of (15.1-1), (15.1-2) can be written in the form of a differential equation,

$$\frac{d\phi(z)}{dz} = \gamma(\nu)\,\phi(z), \tag{15.1-3}$$

where

$$\boxed{\gamma(\nu) = N\sigma(\nu) = N\frac{\lambda^2}{8\pi t_{\text{sp}}}\, g(\nu).} \tag{15.1-4}$$

Gain Coefficient

The coefficient $\gamma(\nu)$ represents the net gain in the photon-flux density per unit length of the medium. The solution of (15.1-3) is the exponentially increasing function

$$\phi(z) = \phi(0)\exp[\gamma(\nu)\, z]. \tag{15.1-5}$$

Since the optical intensity $I(z) = h\nu\phi(z)$, (15.1-5) can also be written in terms of I as

$$I(z) = I(0)\exp[\gamma(\nu)\, z]. \tag{15.1-6}$$

Thus, $\gamma(\nu)$ also represents the gain in the intensity per unit length of the medium.

The amplifier gain coefficient $\gamma(\nu)$ is seen to be proportional to the population difference $N = N_2 - N_1$. Though N was taken to be positive in the example provided above, the derivation is valid whatever the sign of N. In the absence of a population inversion, N is negative ($N_2 < N_1$) and so too is the gain coefficient. The medium will then attenuate (rather than amplify) light traveling in the z direction, in accordance with the exponentially decreasing function $\phi(z) = \phi(0)\exp[-\alpha(\nu)\, z]$, where the attenuation coefficient $\alpha(\nu) = -\gamma(\nu) = -N\sigma(\nu)$. Hence, a medium in thermal equilibrium provides attenuation and cannot provide laser amplification.

Gain

For an interaction region of total length d (Fig. 15.1-1), the overall gain of the laser amplifier $G(\nu)$ is defined as the ratio of the photon-flux density at the output to that at the input, i.e., $G(\nu) = \phi(d)/\phi(0)$, so that

$$\boxed{G(\nu) = \exp[\gamma(\nu)d].} \tag{15.1-7}$$

Amplifier Gain

Bandwidth

The dependence of the gain coefficient $\gamma(\nu)$ on the frequency of the incident light ν is contained in its proportionality to the lineshape function $g(\nu)$, as given in (15.1-4). The latter is centered about the atomic resonance frequency $\nu_0 = (E_2 - E_1)/h$, where E_2 and E_1 are the atomic energy levels, and is of width $\Delta\nu$. Since stimulated emission and absorption are governed by the atomic transition, the laser amplifier is a resonant device, with a resonance frequency and bandwidth determined by the lineshape function of the atomic transition. The linewidth $\Delta\nu$ is measured either in units of frequency (Hz) or in units of wavelength (nm), which are related by $\Delta\lambda = |\Delta(c_o/\nu)| = (c_o/\nu^2)\Delta\nu = (\lambda_o^2/c_o)\Delta\nu$. Thus, a linewidth $\Delta\nu = 1$ THz at $\lambda_o = 0.6\ \mu$m corresponds to $\Delta\lambda = 1.2$ nm.

If the lineshape function is Lorentzian, for example, (14.3-34) provides

$$g(\nu) = \frac{\Delta\nu/2\pi}{(\nu - \nu_0)^2 + (\Delta\nu/2)^2}\,. \tag{15.1-8}$$

The gain coefficient is then also Lorentzian with the same width, i.e.,

$$\gamma(\nu) = \gamma(\nu_0)\frac{(\Delta\nu/2)^2}{(\nu - \nu_0)^2 + (\Delta\nu/2)^2}\,, \tag{15.1-9}$$

as illustrated in Fig. 15.1-2, where $\gamma(\nu_0) = N(\lambda^2/4\pi^2 t_{\rm sp}\Delta\nu)$ is the gain coefficient at the central frequency ν_0.

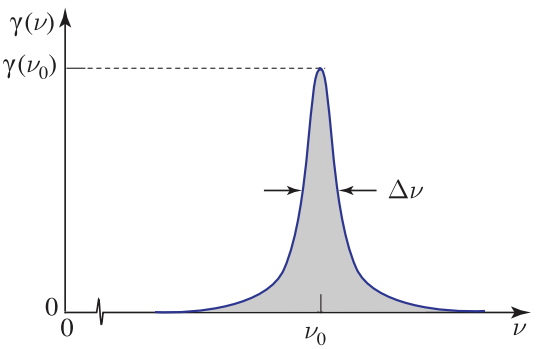


Figure 15.1-2 Gain coefficient $\gamma(\nu)$ of a Lorentzian-lineshape resonant laser amplifier.

EXERCISE 15.1-1

Attenuation and Gain in a Ruby Laser Amplifier.

(a) Consider a ruby (Cr^{3+}:Al_2O_3) crystal with two energy levels separated by an energy difference corresponding to a free-space wavelength $\lambda_o = 694.3$ nm. The transition has a Lorentzian lineshape function of width $\Delta\nu = 330$ GHz, a spontaneous lifetime $t_{\rm sp} = 3$ ms, and ruby has a refractive index $n = 1.76$ (Table 15.3-1). If $N_1 + N_2 = N_a = 10^{22}$ cm^{-3}, determine the population difference $N = N_2 - N_1$ and the attenuation coefficient at the line center $\alpha(\nu_0)$ under conditions of thermal equilibrium at $T = 300°$ K, so that the Boltzmann distribution (Sec. 14.2) is obeyed.

(b) What value should the population difference N assume to achieve a gain coefficient $\gamma(\nu_0) = 0.5$ cm^{-1} at the central frequency?

(c) How long should the ruby crystal be to provide an overall gain of 4 at the central frequency when $\gamma(\nu_0) = 0.5$ cm^{-1}?

B. Phase Shift

Because the gain of the resonant medium is frequency dependent, the medium is dispersive (Sec. 5.5) and a frequency-dependent phase shift must be associated with its gain. The phase shift imparted by the laser amplifier can be determined by considering the interaction of light with matter in terms of the electric field rather than the photon-flux density or the intensity, as we have done in the foregoing. We proceed instead with an alternative approach, in which the mathematical properties of a causal system are used to determine the phase shift. For homogeneously broadened media, the phase-shift coefficient $\varphi(\nu)$ (phase shift per unit length of the amplifier medium) is related to the gain coefficient $\gamma(\nu)$ by the Hilbert transform (Sec. B.1 of Appendix B), so that knowledge of $\gamma(\nu)$ at all frequencies uniquely determines $\varphi(\nu)$.

The optical intensity and the complex amplitude of the field are related by $I(z) = |E(z)|^2/2\eta$. Since $I(z) = I(0)\exp[\gamma(\nu)\, z]$ in accordance with (15.1-6), the field complex amplitude must obey the relation

$$E(z) = E(0)\exp\left[\tfrac{1}{2}\gamma(\nu)\, z\right]\exp[-j\varphi(\nu)\, z]\,, \tag{15.1-10}$$

where $\varphi(\nu)$ is the phase-shift coefficient. The field complex amplitude evaluated at $z + \Delta z$ is therefore

$$\begin{aligned} E(z+\Delta z) &= E(0)\exp\left[\tfrac{1}{2}\gamma(\nu)\,(z+\Delta z)\right]\exp\left[-j\varphi(\nu)\,(z+\Delta z)\right] \\ &= E(z)\exp\left[\tfrac{1}{2}\gamma(\nu)\,\Delta z\right]\exp[-j\varphi(\nu)\,\Delta z] \\ &\approx E(z)\left[1+\tfrac{1}{2}\gamma(\nu)\,\Delta z - j\varphi(\nu)\,\Delta z\right], \end{aligned} \tag{15.1-11}$$

where we have made use of a Taylor-series to approximate the exponential functions. The incremental change in the electric field, $\Delta E(z) = E(z+\Delta z) - E(z)$, therefore satisfies the equation

$$\frac{\Delta E(z)}{\Delta z} = E(z)\left[\tfrac{1}{2}\gamma(\nu) - j\varphi(\nu)\right]. \tag{15.1-12}$$

This incremental amplifier may thus be regarded as a linear system whose input and output are $E(z)$ and $\Delta E(z)/\Delta z$, respectively, and whose transfer function is

$$H(\nu) = \tfrac{1}{2}\gamma(\nu) - j\varphi(\nu). \tag{15.1-13}$$

Because this incremental amplifier represents a physical system, it must be causal. But the real and imaginary parts of the transfer function of a linear causal system are related by the Hilbert transform (Sec. B.1 of Appendix B). It follows that $-\varphi(\nu)$ is the Hilbert transform of $\frac{1}{2}\gamma(\nu)$ so that the amplifier phase shift coefficient is determined by its gain coefficient. A simple example is provided by a Lorentzian atomic lineshape function with narrow width $\Delta\nu \ll \nu_0$, for which the gain coefficient $\gamma(\nu)$ is given by (15.1-9). The corresponding phase shift coefficient $\varphi(\nu)$ is provided in (B.1-13) of Sec. B.1,

$$\boxed{\varphi(\nu) = \frac{\nu - \nu_0}{\Delta\nu}\,\gamma(\nu).} \tag{15.1-14}$$

Phase-Shift Coefficient (Lorentzian Lineshape)

The Lorentzian gain and phase-shift coefficients are plotted in Fig. 15.1-3 as functions of frequency. At resonance, the gain coefficient is maximum and the phase-shift coefficient is zero. The phase-shift coefficient is negative for frequencies below resonance and positive for frequencies above resonance.

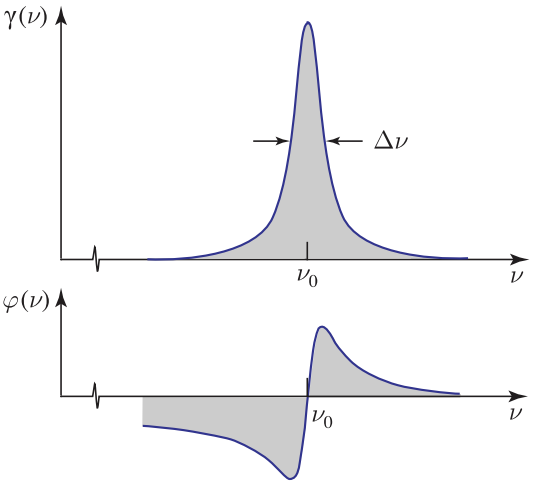


Figure 15.1-3 Gain coefficient $\gamma(\nu)$ and phase-shift coefficient $\varphi(\nu)$ for a laser amplifier with a Lorentzian line-shape function.

15.2 AMPLIFIER PUMPING

Like other amplifiers, laser amplifiers require an external source of power to provide the energy required to augment the input signal. The pump supplies this power via a mechanism that excites the electrons in the atoms, causing them to move from lower to higher atomic energy levels. To achieve amplification, the pump must provide a population inversion on the transition of interest ($N = N_2 - N_1 > 0$). The mechanics of pumping often involves the use of ancillary energy levels. As an example, the pumping of atoms from level ① into level ②, to achieve amplification on the ②→① transition, might be most readily achieved by pumping atoms from level ① into level ③ and then relying on lifetime-based decay from level ③ to populate level ②.

As discussed at the end of Sec. 15.2B, pumping may be achieved in many ways, e.g., by optical, electrical, or chemical means. To attain the **continuous-wave (CW)** operation of a laser amplifier, the rates of excitation and decay of the various energy levels participating in the process must be balanced to maintain a steady-state inverted population on the ②→① transition.

A. Rate Equations

The equations that describe the rates of change of the population densities N_1 and N_2 as a result of pumping, as well as radiative and nonradiative transitions, are called **rate equations**. They are not unlike the first-order differential equations presented in Sec. 14.4, but selective external pumping is now part of the process so that thermal-equilibrium conditions no longer prevail.

Consider the schematic energy-level diagram of Fig. 15.2-1. We focus on levels ① and ②, which have overall lifetimes τ_1 and τ_2, respectively, which accommodate transitions to lower energy levels. The lifetime of level ② has two contributions — τ_{21} is associated with decay from ② to ① while τ_{20} is associated with decay from ② to all other lower levels. When several modes of decay are possible, the overall transition rate is a sum of the component transition rates. Since the rates are inversely proportional to the decay times, the reciprocals of the decay times must be summed:

$$\tau_2^{-1} = \tau_{21}^{-1} + \tau_{20}^{-1}. \tag{15.2-1}$$

Multiple modes of decay therefore shorten the overall lifetime (i.e., they render the

decay more rapid). Aside from the radiative spontaneous emission component t_{sp} in τ_{21}, a nonradiative contribution τ_{nr} may also be present (arising, for example, from a depopulating collision of the atom with the wall of the container), so that

$$\tau_{21}^{-1} = t_{sp}^{-1} + \tau_{nr}^{-1}. \tag{15.2-2}$$

If an unpumped system such as that illustrated in Fig. 15.2-1 is allowed to reach steady state, the population densities N_1 and N_2 will vanish by virtue of all of the electrons having ultimately decayed to lower energy levels.

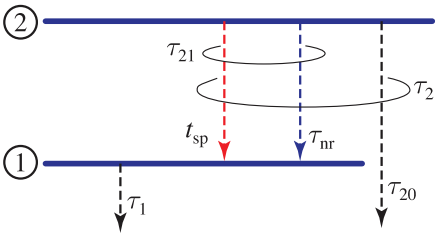


Figure 15.2-1 Energy levels ① and ②, along with their decay times.

Steady-state populations of levels ① and ② can be maintained, however, if the energy levels above level ② are continuously excited by pumping and ultimately populate level ②, as shown in the more realistic energy-level diagram of Fig. 15.2-2. Pumping serves to bring atoms out of level ① and into level ②, at rates R_1 and R_2 (per unit volume per unit time), respectively, as shown in simplified form in Fig. 15.2-3. As a result, levels ① and ② can achieve nonzero steady-state populations.

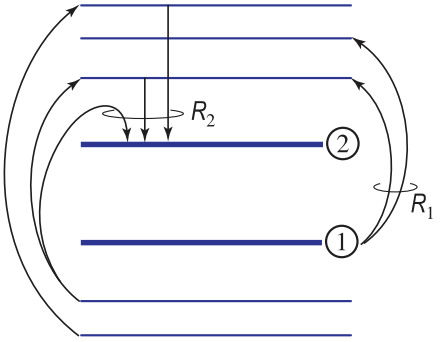


Figure 15.2-2 Energy levels ① and ②, together with surrounding higher and lower energy levels, in the presence of pumping.

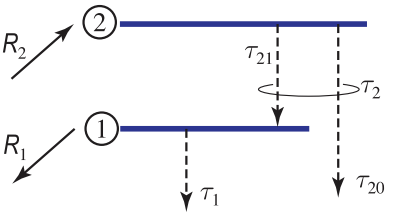


Figure 15.2-3 Energy levels ① and ② and their decay times. By means of pumping, the population density of level ② is increased at the rate R_2 while that of level ① is decreased at the rate R_1.

We now proceed to write the rate equations for this system both in the absence, and in the presence, of amplifier radiation (the radiation resonant with the ②→① transition).

Rate Equations in the Absence of Amplifier Radiation

The rates of increase of the population densities of levels ② and ① arising from pumping and decay are, respectively,

$$\frac{dN_2}{dt} = R_2 - \frac{N_2}{\tau_2} \tag{15.2-3}$$

$$\frac{dN_1}{dt} = -R_1 - \frac{N_1}{\tau_1} + \frac{N_2}{\tau_{21}}. \tag{15.2-4}$$

Under steady-state conditions ($dN_1/dt = dN_2/dt = 0$), (15.2-3) and (15.2-4) can be solved for N_1 and N_2, which provides a population difference $N = N_2 - N_1$ given by

$$N_0 = R_2\tau_2\left(1 - \frac{\tau_1}{\tau_{21}}\right) + R_1\tau_1\,, \tag{15.2-5}$$

Steady-State Population Difference (Absence of Amplifier Radiation)

where the symbol N_0 represents the steady-state population difference N in the absence of amplifier radiation.

In accordance with (15.1-4), a large gain coefficient requires a large population difference, i.e., a large positive value of N_0. Equation (15.2-5) demonstrates that this may be achieved by:

- Large R_1 and R_2.
- Long τ_2 (but t_{sp}, which contributes to τ_2 through τ_{21}, must be sufficiently short so as to make the radiative transition rate large, as will be seen subsequently).
- Short τ_1 if $R_1 < (\tau_2/\tau_{21})R_2$.

The physical rationales that underlie these conditions make good sense. The upper level should be strongly pumped and decay slowly so that it retains its population. The lower level should strongly depump so that it quickly disposes of its population. Ideally, it is desirable to have $\tau_{21} \approx t_{sp} \ll \tau_{20}$ so that $\tau_2 \approx t_{sp}$ and $\tau_1 \ll t_{sp}$. Under these conditions, (15.2-5) simplifies to

$$N_0 \approx R_2 t_{sp} + R_1\tau_1. \tag{15.2-6}$$

In the absence of depumping ($R_1 = 0$), or when $R_1 \ll (t_{sp}/\tau_1)R_2$, this result further simplifies to

$$N_0 \approx R_2 t_{sp}. \tag{15.2-7}$$

EXERCISE 15.2-1

Optical Pumping. Assume that $R_1 = 0$ and that R_2 is realized by exciting atoms from the ground state $E = 0$ to level ② using photons of frequency E_2/h absorbed with transition probability W. Assume that $\tau_2 \approx t_{sp}$ and $\tau_1 \ll t_{sp}$ so that in steady state $N_1 \approx 0$ and $N_0 \approx R_2 t_{sp}$. If N_a is the total population of levels ⓪, ①, and ②, show that $R_2 \approx (N_a - 2N_0)W$, so that the population difference is $N_0 \approx N_a t_{sp} W/(1 + 2t_{sp}W)$.

Rate Equations in the Presence of Amplifier Radiation

The presence of radiation near the resonance frequency ν_0 enables transitions between levels ② and ① to take place via stimulated emission and absorption. These processes are characterized by the probability density $W_i = \phi\,\sigma(\nu)$, as provided in (15.1-1) and illustrated in Fig. 15.2-4. At this juncture, we assume for simplicity that $\sigma(\nu)$, and therefore W_i, are identical for absorption and emission. This assumption is not valid in general, however, as will be discussed subsequently.

Extending the rate equations (15.2-3) and (15.2-4) to include the population loss and gain associated with amplifier radiation yields

$$\frac{dN_2}{dt} = R_2 - \frac{N_2}{\tau_2} - N_2 W_i + N_1 W_i \tag{15.2-8}$$

$$\frac{dN_1}{dt} = -R_1 - \frac{N_1}{\tau_1} + \frac{N_2}{\tau_{21}} + N_2 W_i - N_1 W_i\,. \tag{15.2-9}$$

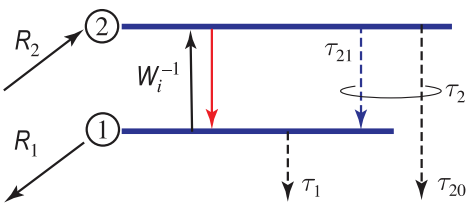


Figure 15.2-4 Population densities N_1 and N_2 of atoms in energy levels ① and ②, respectively, are determined by three processes: decay (at rates $1/\tau_1$ and $1/\tau_2$, respectively, which includes the effects of spontaneous emission), depumping and pumping (at rates R_1 and R_2, respectively), and absorption and stimulated emission (at the same rate W_i with corresponding time constant W_i^{-1}).

The population density of level ② is decreased by stimulated emission from level ② to level ① and increased by absorption from level ① to level ②. The spontaneous emission contribution is contained in τ_{21}. The population densities are generally specified in units of $\text{cm}^{-3} \cdot \text{s}^{-1}$. Under steady-state conditions ($dN_1/dt = dN_2/dt = 0$), (15.2-8) and (15.2-9) are readily solved for N_1 and N_2, which leads to a population difference $N = N_2 - N_1$ given by

$$N = \frac{N_0}{1 + \tau_s W_i}, \tag{15.2-10}$$

Steady-State Population Difference (Presence of Amplifier Radiation)

where N_0 is the steady-state population difference in the absence of amplifier radiation provided in (15.2-5). The characteristic time τ_s, given by

$$\tau_s = \tau_2 + \tau_1 \left(1 - \frac{\tau_2}{\tau_{21}}\right), \tag{15.2-11}$$

Saturation Time Constant

is always positive since $\tau_2 \leq \tau_{21}$.

In the absence of amplifier radiation, $W_i = 0$ so that (15.2-10) reverts to $N = N_0$, as expected. Because τ_s is positive, the steady-state population difference in the presence of amplifier radiation always has a smaller absolute value than in its absence, i.e., $|N| \leq |N_0|$. If the radiation is sufficiently weak so that $\tau_s W_i \ll 1$ (the **small-signal approximation**), it suffices to take $N \approx N_0$. As the amplifier radiation becomes stronger, W_i increases and ultimately $N \to 0$, regardless of the initial sign of N_0, as shown in Fig. 15.2-5. This result emerges because stimulated emission and absorption dominate the interaction when W_i is very large and they have equal probability densities. It is apparent that even very strong radiation cannot convert a negative population difference into a positive one, nor *vice versa*. The quantity τ_s plays the role of a **saturation time constant**, as is evident from Fig. 15.2-5.

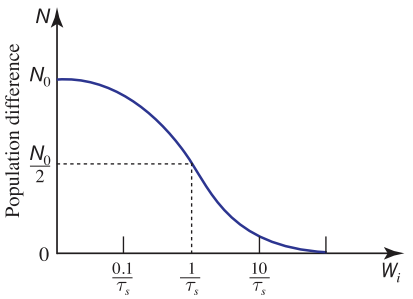


Figure 15.2-5 Depletion of the steady-state population difference $N = N_2 - N_1$ as the rate of absorption and stimulated emission W_i increases. When $W_i = 1/\tau_s$, N is reduced by a factor of 2 from its value when $W_i = 0$.

EXERCISE 15.2-2

Saturation Time Constant. Show that if $t_{sp} \ll \tau_{nr}$ (where τ_{nr} is the nonradiative part of the lifetime τ_{21} of the ②→① transition) and $\tau_1 \ll t_{sp} \ll \tau_{20}$, then $\tau_s \approx t_{sp}$.

B. Pumping Schemes

We proceed to examine a number of pumping schemes used to achieve laser amplification via a population inversion on the ②→① transition. We consider four pumping systems in turn: 1) four-level pumping; 2) three-level pumping; 3) quasi-three-level pumping; and 4) in-band pumping. With the exception of three-level pumping, all of these schemes are widely used in practice.

Four-Level Pumping

In the **four-level pumping** arrangement, displayed in Fig. 15.2-6, level ① lies above the ground state (designated as the lowest energy level ⓪). In thermal equilibrium, level ① will be virtually unpopulated provided that $E_1 \gg kT$, a situation that is, of course, desirable since it enhances the population inversion.

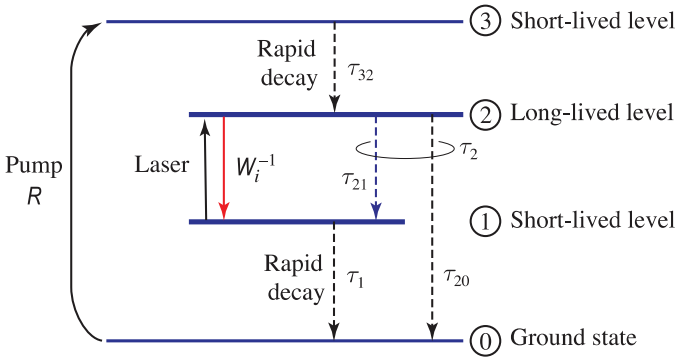


Figure 15.2-6 Energy levels and decay rates for a four-level system. The four levels are drawn from a multitude of levels (not shown). The rate of pumping into level ③, and out of level ⓪, are taken to be the same. In a four-level system, level ① is assumed to be unpopulated in thermal equilibrium. A quasi-three-level system has the same configuration except that level ① is sufficiently close to the ground state ⓪ that it retains population in thermal equilibrium. The majority of laser amplifiers and lasers encountered in practice are four-level or quasi-three-level systems.

Pumping is achieved by making use of an energy level (or collection of energy levels) that lies above level ②; we designate this as level ③. The ③→② transition has a short lifetime (decay occurs rapidly) so there is little population accumulation in level ③. Level ② is long-lived, so that it accumulates population, whereas level ① is short-lived so that it sheds population; a population inversion is thereby established between levels ② and ①. All told, four energy levels are involved in the process but the optical interaction of interest takes place between levels ② and ①.

An external source of energy (e.g., photons at frequency $\nu \approx E_3/h$) pumps atoms from level ⓪ to level ③ at a rate R. If the decay from level ③ to level ② is sufficiently rapid, it may be taken to be instantaneous, whereupon pumping to level ③ is equivalent to pumping to level ② at the rate $R_2 = R$. The situation is then the same as that shown in Fig. 15.2-4 so that the expressions in (15.2-10) and (15.2-11) apply. With respect to the value of N_0 to be used in these expressions, atoms are neither pumped into nor out of level ① in the four-level system, so that $R_1 = 0$. In the absence of amplifier

radiation ($W_i = \phi = 0$), therefore, the steady-state population difference is given by (15.2-5) with $R_1 = 0$, i.e.,

$$N_0 = R\tau_2\left(1 - \frac{\tau_1}{\tau_{21}}\right). \tag{15.2-12}$$

Also, in most four-level systems, the nonradiative decay component in the ②→① transition is negligible ($t_{\rm sp} \ll \tau_{\rm nr}$) and $\tau_1 \ll t_{\rm sp} \ll \tau_{20}$ (Exercise 15.2-2), so that

$$N_0 \approx Rt_{\rm sp} \tag{15.2-13}$$

$$\tau_s \approx t_{\rm sp}\,, \tag{15.2-14}$$

whereupon (15.2-10) becomes

$$N \approx \frac{Rt_{\rm sp}}{1 + t_{\rm sp}W_i}\,. \tag{15.2-15}$$

Implicit in the preceding derivation is the assumption that the pumping rate R is independent of the population difference $N = N_2 - N_1$. This is not always the case, however, because the population densities of the ground state and level ③, N_g and N_3 respectively, are related to N_1 and N_2 by

$$N_g + N_1 + N_2 + N_3 = N_a, \tag{15.2-16}$$

where the total atomic density in the system, N_a, is a constant. If the pumping involves a transition between the ground state and level ③ with transition probability W, then $R = (N_g - N_3)W$. If levels ① and ③ are short-lived, then $N_1 \approx N_3 \approx 0$, whereupon $N_g + N_2 \approx N_a$ so that $N_g \approx N_a - N_2 \approx N_a - N$. Under these conditions, the pumping rate can be approximated as

$$R \approx (N_a - N)W, \tag{15.2-17}$$

which reveals that the pumping rate is a linearly decreasing function of the population difference N and is thus clearly not independent of it. This arises because the population inversion established between levels ② and ① reduces the number of atoms available to be pumped. Substituting (15.2-17) into (15.2-15) and reorganizing terms leads to

$$N \approx \frac{t_{\rm sp}N_aW}{1 + t_{\rm sp}W + t_{\rm sp}W_i}\,. \tag{15.2-18}$$

Finally, then, the population difference can be written in the generic form of (15.2-10):

$$\boxed{N = \frac{N_0}{1 + \tau_s W_i}\,,} \tag{15.2-19}$$

where N_0 and τ_s, rather than being expressed as (15.2-13) and (15.2-14), become

$$\boxed{N_0 \approx \frac{t_{\rm sp}N_aW}{1 + t_{\rm sp}W},} \tag{15.2-20}$$

$$\boxed{\tau_s \approx \frac{t_{\rm sp}}{1 + t_{\rm sp}W}\,.} \tag{15.2-21}$$

For weak pumping ($W \ll 1/t_{\rm sp}$), $N_0 \approx t_{\rm sp}N_aW$ is proportional to the pumping transition probability density W, and $\tau_s \approx t_{\rm sp}$, so that (15.2-13) and (15.2-14) reemerge. However, as the pumping strength increases, N_0 decreases and ultimately saturates, while τ_s decreases.

Three-Level Pumping

A **three-level pumping** arrangement, in contrast, makes use of the ground state ($E_1 = 0$) as the lower laser level ①, as depicted in Fig. 15.2-7. Again, an auxiliary third level (designated ③) is involved and the ③→② decay is rapid so that there is no buildup of population in level ③. The ③→① decay is slow ($\tau_{32} \ll \tau_{31}$) so that the pumping serves to populate level ②, the upper laser level, which is long-lived and therefore accumulates population. Atoms are pumped from level ① to level ③ (e.g., by absorbing light at frequency $\nu \approx E_3/h$) at a rate R; the fast (nonradiative) decay effectively pumps level ② at the rate $R_2 = R$. The thermally excited population of level ② is assumed to be negligible.

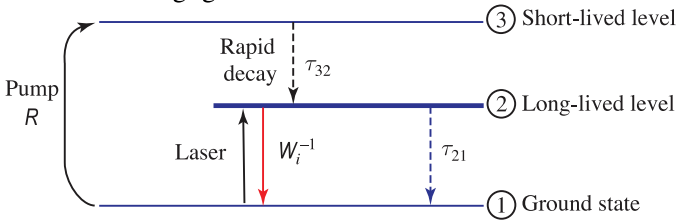


Figure 15.2-7 Energy levels and decay rates for a three-level system. A multitude of other energy levels exist, but they are not germane to the considerations at hand. It is assumed that the rate of pumping into level ③ is the same as the rate of pumping out of level ①.

It is not difficult to see that under rapid ③→② decay, the three-level system displayed in Fig. 15.2-7 is a special case of the system shown in Fig. 15.2-4 (provided that R is independent of N) with the parameters

$$R_1 = R_2 = R, \qquad \tau_1 = \infty, \qquad \tau_2 = \tau_{21}. \tag{15.2-22}$$

To avoid algebraic problems in connection with the value $\tau_1 = \infty$, rather than substituting these special values into (15.2-10) and (15.2-11), we return to the original rate equations (in the presence of radiation), (15.2-8) and (15.2-9). In steady state, both of these equations provide the same result:

$$0 = R - \frac{N_2}{\tau_{21}} - N_2 W_i + N_1 W_i \,. \tag{15.2-23}$$

It is not possible to determine both N_1 and N_2 from a single equation relating them. However, knowledge of the total atomic density N_a in levels ①, ②, and ③ provides an auxiliary condition that does permit N_1 and N_2 to be determined. Since τ_{32} is very short, level ③ retains a negligible steady-state population; all of the atoms that are raised to it immediately decay to level ②. Thus,

$$N_1 + N_2 \approx N_a, \tag{15.2-24}$$

which enables us to solve (15.2-23) for N_1 and N_2 and thereby to determine the population difference $N = N_2 - N_1$ and the saturation time τ_s. The result may be cast in the usual form of (15.2-10), $N = N_0/(1 + \tau_s W_i)$, where now

$$N_0 = 2R\,\tau_{21} - N_a \tag{15.2-25}$$

$$\tau_s = 2\tau_{21} \,. \tag{15.2-26}$$

When nonradiative decay from level ② to level ① is negligible (i.e., $t_{\rm sp} \ll \tau_{\rm nr}$), τ_{21} may be replaced by $t_{\rm sp}$, whereupon

$$N_0 \approx 2R\,t_{\rm sp} - N_a \tag{15.2-27}$$

$$\tau_s \approx 2t_{\rm sp} \,. \tag{15.2-28}$$

The dependence of the pumping rate R on the population difference N can be included in the analysis of the three-level system by writing $R = (N_1 - N_3)W$, $N_3 \approx 0$, and $N_1 = \frac{1}{2}(N_a - N)$, from which we obtain $R \approx \frac{1}{2}(N_a - N)W$. Substituting this in the principal equation $N = (2Rt_{sp} - N_a)/(1 + 2t_{sp}W_i)$, and reorganizing terms, we can write the population difference in the usual form,

$$N = \frac{N_0}{1 + \tau_s W_i}, \tag{15.2-29}$$

but now with

$$N_0 \approx \frac{N_a(t_{sp}W - 1)}{1 + t_{sp}W}, \tag{15.2-30}$$

$$\tau_s \approx \frac{2t_{sp}}{1 + t_{sp}W}. \tag{15.2-31}$$

Just as with the results for the four-level scheme presented in (15.2-20) and (15.2-21), N_0 and τ_s saturate as the pumping transition probability W increases.

Comparison of Three- and Four-Level Pumping

It is of interest to compare (15.2-27) and (15.2-28) for three-level pumping with the analogous results (15.2-13) and (15.2-14) for four-level pumping. Attaining a population inversion ($N > 0$ and therefore $N_0 > 0$) in the three-level system requires a pumping rate $R > N_a/2t_{sp}$. Thus, just to make the population density N_2 equal to N_1 (i.e., $N_0 = 0$) requires a substantial pump power density, $E_3 N_a/2t_{sp}$. The large population in the ground state (which is the lowest laser level) is an inherent obstacle to achieving a population inversion in a three-level system. This impediment is avoided in a four-level system, where level ① is normally empty since τ_1 is short. The saturation time constant $\tau_s \approx t_{sp}$ for the four-level pumping scheme is half that for the three-level scheme.

EXERCISE 15.2-3

Pumping Power in Three- and Four-Level Systems.

(a) Determine the pumping transition probability W required to achieve a zero population difference in a three- and a four-level laser amplifier.

(b) If the pumping transition probability $W = 2/t_{sp}$ in the three-level system, and $W = 1/2t_{sp}$ in the four-level system, show that $N_0 = N_a/3$. Compare the pumping powers required to achieve this population difference.

Quasi-Three-Level Pumping

Quasi-three-level pumping is a scheme intermediate between four-level and three-level pumping. As displayed in Fig. 15.2-6, its configuration is identical to that of four-level pumping; the sole distinction is that the lower laser level ① is sufficiently close to the ground state ⓪ that it retains some population in thermal equilibrium. As with the four-level system, pumping is achieved via auxiliary level ③; the ③→② transition is rapid; the upper laser level ② is long-lived; a population inversion is established between levels ② and ①; and the lower laser level ① is short-lived. However, because of the residual population in level ①, reabsorption at the transition frequency makes the task of achieving a population inversion more challenging than in four-level pumping. Many lasers operate via quasi-three-level pumping.

In-Band Pumping

It is not possible to create a steady-state population inversion using an idealized two-level system and direct optical pumping. This is a consequence of the fact that the pump creates as much stimulated emission as absorption so that, at best, the populations can equalize; this can be readily understood from the rate-equation analysis discussed in Prob. 15.2-4. To circumvent this limitation, pumping schemes typically rely on an auxiliary level ③ to funnel population to the upper laser level ②, as discussed earlier in connection with four-level, three-level, and quasi-three-level pumping.

However, narrow isolated energy levels, such as those portrayed in Figs. 15.2-6 and 15.2-7, are idealizations. Though they are suitable for formulating laser rate equations and for understanding how three- and four-level pumping schemes operate, the energy levels in real materials often comprise collections of sublevels. An example is provided by the well-known Nd^{3+}:YAG laser transition at 1.064 μm portrayed in Fig. 14.1-5. As shown in far greater detail in Fig. 14.1-6, the levels designated by the term symbols $^4I_{9/2}$, $^4I_{11/2}$, and $^4F_{3/2}$, corresponding to laser levels ⓪, ①, and ②, respectively, consist of manifolds that contain collections of multiple sublevels.

The presence of these sublevels within each level makes it possible to nest the pump band ③ and upper laser level ② within a single manifold, and the lower laser level ① and ground state ⓪, within another. This scheme, illustrated in Fig. 15.2-8, is known as **in-band pumping**. It is also referred to as **quasi-two-level pumping** since the pair of manifolds superficially resembles a two-level system. Pumping and laser action are readily implemented at distinct frequencies in this pumping scheme. Since the frequency of the pump ν_p is greater than that of the laser ν_s, as depicted in Fig. 15.2-8, the pump preferentially raises ions to the higher sublevels of the upper manifold. The active ions quickly thermalize (over a time scale of ps), and fall to the lower sublevels where they are available for laser-induced stimulated emission. They are then immune to pump-induced stimulated emission as well as to laser-induced reabsorption since the higher sublevels of the lower manifold are only weakly populated at normal temperatures. Since the processes of absorption and emission access different sublevels within the manifolds, the transition cross sections for the two processes generally differ, and therefore so do the probability densities W_i for absorption and stimulated emission. The **effective absorption cross section** and **effective emission cross section** are denoted $\sigma_{\rm ab}(\nu)$ and $\sigma_{\rm em}(\nu)$, respectively.

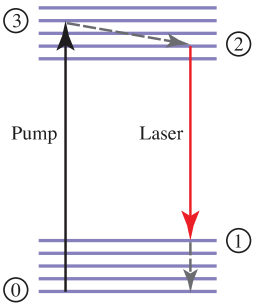


Figure 15.2-8 In-band pumping. The pump band ③ and the upper laser level ② are nested within the upper manifold while the lower laser level ① and the ground state ⓪ are nested within the lower manifold. The pump and laser photons have energies $h\nu_p$ and $h\nu_s$, respectively. The absorption spectrum is on the high-frequency side of the emission spectrum so that a population inversion can be implemented by strong pumping at a frequency higher than the peak of the emission spectrum. The effective absorption and emission cross sections, $\sigma_{\rm ab}(\nu)$ and $\sigma_{\rm em}(\nu)$, respectively, generally differ.

In-band pumping is widely used. It can be employed in four-level systems such as Nd^{3+}:YAG and Nd^{3+}:glass, where the lower laser level is well above the ground state, as well as in quasi-three-level systems such as Yb^{3+}:YAG, where the lower laser level is near the ground state. In the latter case, the energies of the pump and laser photons are quite close to each other so that the fraction of pump photon energy lost in the course of generating a lower-frequency laser photon is small. This fraction, known as the **quantum defect**, is given by

$$q = 1 - \nu_s/\nu_p. \tag{15.2-32}$$

The smaller the quantum defect q, the higher the pumping efficiency. Though in-band pumping enjoys the benefit of a small quantum defect, the gain of such systems can be adversely affected by the presence of pump-induced stimulated emission and signal-induced reabsorption, as discussed above.

Pumping Methods

Many techniques are available for pumping laser amplifiers and lasers, the most common of which make use of electrical and optical means. Electrical pumping can be implemented by passing a current, or an electron or ion beam, through a medium. In laser diodes, electrical pumping is realized by injecting charge carriers in the form of electrons and holes (Sec. 18.3). Solid insulating laser media such as crystals, doped glasses, and doped ceramics usually make use of optical pumping; indeed, lasers are often used to pump laser amplifiers and other lasers. The efficiency of pumping varies considerably, depending on the form of pumping and the system under consideration. Electrical pumping is generally highly efficient, particularly for laser diodes, whereas optical pumping can be quite inefficient if an appreciable fraction of pump photons are lost and/or if their energy is not fully utilized (e.g., if the quantum defect q is large).

Several common methods of electrical and optical pumping are illustrated schematically in Fig. 15.2-9.

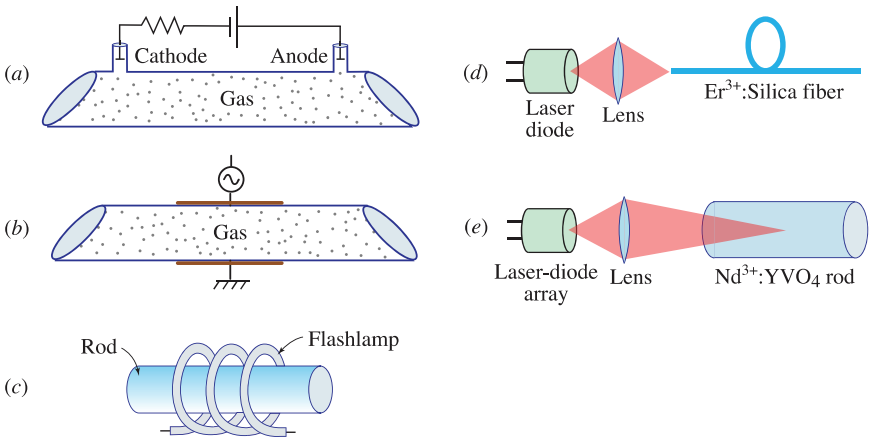


Figure 15.2-9 Examples of electrical and optical pumping. (a) Direct current (DC) is often used to pump gas lasers. The current may be passed either along the laser axis, creating a longitudinal discharge, or transversely to it. (b) Radio-frequency (RF) discharges are also used for pumping gas lasers. (c) Xe flashlamps or Kr CW arc lamps can be used to optically pump ruby and rare-earth-doped solid-state laser amplifiers and lasers. (d) Semiconductor laser diodes, which are themselves electrically pumped, are widely used for pumping Er^{3+}:silica-fiber laser amplifiers and other rare-earth-doped fiber amplifiers. (e) An array of laser diodes is generally used for the efficient optical pumping of Nd^{3+}:YVO_4 and other solid-state laser amplifiers and lasers, as well as for fiber lasers.

Pumping schemes that excite various forms of luminescence radiation are discussed in Sec. 14.5A. As will become apparent in Sec. 16.3, a variety of schemes aside from electrical and optical pumping find use in laser amplifiers and lasers. Chemical lasers make use of chemical reactions that lead to reaction products in excited states. Atomic X-ray lasers and amplifiers rely on focused laser-beam heating to create a plasma of ionized atoms in excited states. Relativistic electron beams serve as pumps for free-electron laser amplifiers and lasers. Nuclear pumping makes use of a stream of high-energy X-rays or other particles derived from nuclear reactions, radioisotopes, or a nuclear detonation.

15.3 REPRESENTATIVE LASER AMPLIFIERS

Laser amplification can take place in a great variety of materials. The energy-level diagrams for a number of representative atoms, ions, molecules, solids, and doped dielectrics that exhibit laser action are displayed in Sec. 14.1. Practical laser systems usually involve a plethora of interacting energy levels that influence the population densities associated with the transition of interest, ②→①, as illustrated in Fig. 15.2-2. Nevertheless, the essential principles of laser-amplifier operation may be codified in terms of the pumping schemes set forth in Sec. 15.2.

This is illustrated by several laser systems that we consider in turn: the three-level ruby laser amplifier, the four-level neodymium-doped glass laser amplifier, and the quasi-three-level erbium-doped silica-fiber laser amplifier. The neodymium-doped glass and erbium-doped silica-fiber amplifiers are also amenable to in-band pumping. Most laser amplifiers and lasers operate on the basis of four-level, quasi-three-level, or in-band pumping schemes, but three-level ruby is of interest for historical and didactic reasons. All three of the laser amplifiers discussed here are doped dielectrics and hence are optically pumped. We also consider an optically pumped amplifier that operates on the basis of stimulated Raman scattering, rather than stimulated emission. All of these amplifiers also operate as laser oscillators (Sec. 16.3). The semiconductor optical amplifier, which is almost always electrically pumped, is described in Sec. 18.2.

Many laser amplifiers are used as **power amplifiers** (also called **postamplifiers**), in which the amplifier is used to increase the power of a high-quality, but low-power laser oscillator called a **seed laser** or **master-oscillator**. Such systems, known as **master-oscillator power-amplifiers (MOPAs)**, offer a number of advantages that will be elucidated in Sec. 16.3B. They find use in applications such as cable television (CATV), where strong, clean signals must be generated before being fanned out into multiple fiber channels. MOPAs comprise various combinations of different types of lasers and amplifiers, including diode-pumped solid-state, semiconductor, and fiber devices. In the special case when a fiber amplifier serves as the power amplifier, a MOPA is also referred to as a **master-oscillator fiber-amplifier (MOFA)**.

Other laser amplifiers are used as **in-line amplifiers** (also called **line amplifiers**). An example is provided by the erbium-doped silica-fiber amplifier, which serves to boost optical signals in-line as they traverse long-haul optical fiber communication links (Sec. 25.1C). Yet another amplifier configuration is the **optical preamplifier**, which boosts a signal before it is sent to another amplifier or to a photodetector (Fig. 25.1-5). Laser amplifiers are often operated in the saturation regime to reduce noise (Sec. 15.4).

A. Ruby

Ruby is a dielectric medium with refractive index $n \approx 1.76$. It consists of sapphire (Al_2O_3) in which chromium ions (Cr^{3+}) replace a small percentage ($\approx 0.05\%$) of the aluminum ions (Sec. 14.1B). Though ruby was the first material in which laser action was observed (see page 657), it is rarely used and hence serves principally as a didactic example.

As in most materials that support laser action, stimulated emission can take place on a variety of transitions in ruby. The energy levels pertinent to the well-known red ruby-laser transition are displayed in Fig. 15.3-1 (which is a more fully annotated version of Fig. 14.1-4). Operation is on the basis of a *three-level* pumping scheme. Level ① is the ground state and lower laser level. Level ②, the upper laser level, comprises a pair of closely spaced, discrete sublevels that are not resolved in Fig. 15.3-1; the transition from the lower of these sublevels (known as R_1) to the ground state is responsible for the ruby-red laser radiation at $\lambda_o = 694.3$ nm. Level ③ comprises

two broad pump bands, centered at approximately 550 nm (green) and 400 nm (violet), that serve to populate the upper laser level via rapid decay. These absorption bands are responsible for the reddish color of ruby as white light passes through it. The energy levels portrayed in Fig. 15.3-1 are designated by group-theoretical symbols rather than by term symbols for Cr^{3+} (Table 14.1-1) since the energy levels of transition-metal-ion doped materials are determined in large part by the crystal field associated with the host medium, as discussed in Sec. 14.1B. For ruby, the result is clearly a mixture of discrete energy levels and energy bands.

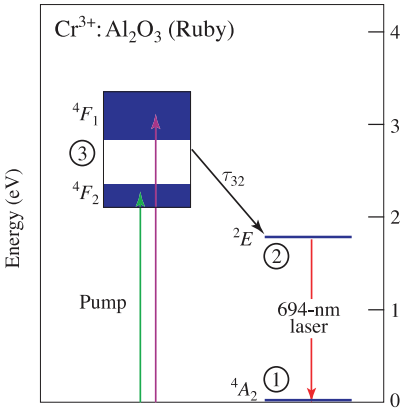


Figure 15.3-1 Relevant energy levels for operation of the ruby laser amplifier in the red. The energy levels are designated by group-theoretical symbols for reasons discussed in the text. The three interacting energy levels are indicated by encircled numbers. Level ③ comprises two broad pump bands, centered in the green and violet. These serve to populate the upper laser level ②, consisting of upper and lower sublevels R_2 and R_1, respectively (these are not resolved at the scale of the figure). Level ① is the lower laser level and ground state. Stimulated emission on the transition from the R_1 line in level ② to level ① gives rise to the well-known red laser light at $\lambda_o = 694.3$ nm.

As illustrated in Fig. 15.3-2, a ruby rod may be optically pumped from level ① to level ③ by surrounding it with a helical flashlamp. A more efficient pumping configuration places the ruby rod, along with a linear flashlamp, at the foci of a reflecting cylinder of elliptical cross section (Fig. 1.2-3). The flashlamp emits a pulse of white light, a portion of which is absorbed by level ③, which is quite broad, resulting in the excitation of a fraction of the Cr^{3+} ions to level ③. These ions rapidly decay from level ③ to level ② with a time constant τ_{32} of the order of ps. The excited ions remain in level ② for a substantial period of time since the ②→① transition lifetime $\tau_{21} \approx 3$ ms is relatively long (nonradiative decay is negligible so that $\tau_{21} \approx t_{sp}$). The ruby-laser system therefore complies with the three-level time-constant requirements dictated by Fig. 15.2-7. The transition has a homogeneously broadened linewidth $\Delta\nu \approx 330$ GHz that arises principally from elastic electron scattering from lattice vibrations (phonons). A number of key characteristics of the ruby-laser transition and the ruby-laser oscillator are provided in Tables 15.3-1 and 16.3-1, respectively.

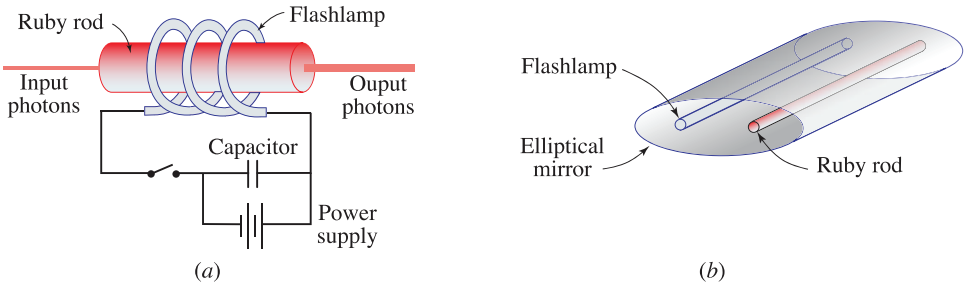


Figure 15.3-2 Ruby laser-amplifier configurations. (*a*) Geometry used for the first laser oscillator built by Maiman in 1960 (see Chapter 16). (*b*) High-efficiency pumping geometry using a linear flashlamp in a reflecting elliptical cylinder.

B. Neodymium-Doped Glass

Neodymium-doped phosphate glass, a dielectric of refractive index $n \approx 1.50$, can be manufactured in large volumes with high optical quality and excellent optical finish. Glass is isotropic and can be doped in a homogeneous fashion. Exceptionally large Nd^{3+}:glass amplifiers can therefore be fabricated and used to generate extremely powerful optical pulses. Since glass has limited thermal conductivity, however, such pulses are usually generated with low duty cycle to provide ample time for the glass to dissipate heat between pulses.

The energy levels of a neodymium-doped phosphate-glass laser amplifier are displayed in Fig. 15.3-3 (this is an expanded version of Fig. 14.1-5). Levels ⓪, ①, and ② of this *four-level* laser system represent the ground state, lower laser level, and upper laser level, respectively. Stimulated emission occurs on the ②→① laser transition, at $\lambda_o = 1.053\ \mu m$ in the near infrared. The energy levels are designated by means of term symbols for the Nd^{3+} ion (Table 14.1-1), as discussed in Sec. 14.1B. Level ③ comprises four pump bands, each about 30 nm wide and centered near 805 nm (near infrared), 745 nm (red), 585 nm (yellow), and 520 nm (green). The spectral profile of a Xe flashlamp suitable for pumping this amplifier is also shown in Fig. 15.3-3.

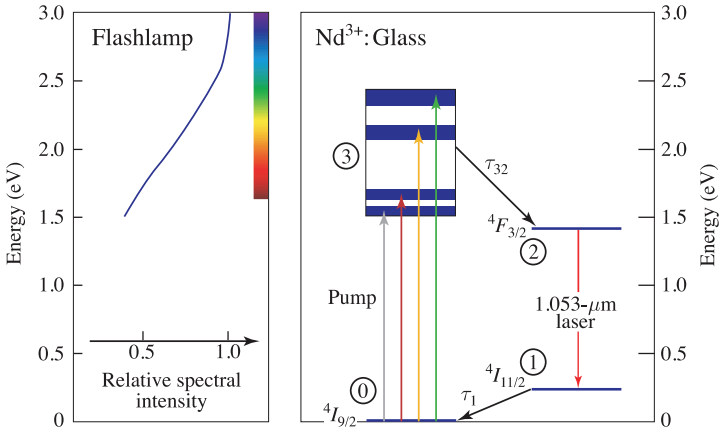


Figure 15.3-3 Right: Relevant energy levels for a neodymium-doped phosphate-glass four-level laser amplifier that operates at $1.053\ \mu m$ on the ${}^4F_{3/2} \rightarrow {}^4I_{11/2}$ transition. Level ③ comprises four pump bands (centered in the near infrared, red, yellow, and green) that serve to populate the upper laser level ②. The lower laser level and ground state are denoted ① and ⓪, respectively. The energy levels are labeled with Nd^{3+}-ion term symbols. Left: Spectral profile of the emission from a broadband Xe flashlamp suitable for pumping this amplifier.

The neodymium-doped glass laser amplifier operates in the following way. The absorption of flashlamp light by the four pump bands excites ions from level ⓪ to level ③. These excited ions decay rapidly (time constant τ_{32}) and populate the metastable upper laser level ②, which has a relatively long lifetime ($t_{sp} = 370\ \mu s$). Stimulated emission on the ②→① transition provides laser amplification at $\lambda_o = 1.053\ \mu m$. Level ① has a short lifetime ($\tau_1 \approx 300$ ps) and lies at an energy ≈ 0.24 eV above the ground state; because this is substantially larger than the thermal energy at room temperature, $kT \approx 0.026$ eV, the thermal population in the lower laser level is negligible. These features comport with the time-constant requirements specified in Fig. 15.2-6 for a four-level system. The ②→① laser transition is inhomogeneously broadened as a result of the amorphous nature of the glass, which presents a different environment at each ionic location. The sublevels in the manifolds are thus smeared into bands, as discussed in connection with Fig. 14.1-6; this gives rise to a large room-temperature linewidth of $\Delta\nu \approx 7$ THz (see Tables 15.3-1 and 16.3-1).

EXAMPLE 15.3-1. ***High-Power Neodymium-Doped Glass Laser Amplifiers at the NIF.*** The neodymium-doped glass laser amplifier plays a central role at the National Ignition Facility (NIF), located at the Lawrence Livermore National Laboratory (LLNL) in Livermore, California. Such amplifiers are widely used in experiments designed to achieve controlled thermonuclear fusion in an encapsulated fuel target since they are capable of generating optical pulses with enormous energies and peak powers. The NIF system marshals thousands of such amplifiers in a facility that occupies a ten-story building the size of a sports stadium.

The optical configuration at the NIF takes the form of a master-oscillator power-amplifier (MOPA) (Sec. 16.3B). The initial optical pulse is provided by a highly stable diode-pumped Yb^{3+}-doped fiber master-oscillator (seed laser) that generates a 1-nJ pulse with a duration of ≈ 5 ns. In a simplified description, this pulse is split and sent via optical fibers to 48 laser preamplifiers that boost the overall pulse energy to 10 J. At the same time, the spatial profiles of the beams are reshaped so they assume 18 mm $\times$ 18 mm square cross sections, which enables the individual glass amplifiers to be tightly packed into compact configurations. Each of these 48 beams is then split into four beams, resulting in 192 main beamlines. Each beamline is endowed with a power amplifier and a main amplifier, through which the pulse makes four roundtrips.

The Nd^{3+}-doped phosphate-glass laser amplifiers used in the 192 main NIF beamlines have the energy-level diagram, and use flashlamps with the spectral profile, illustrated in Fig. 15.3-3. There are four clusters of such laser amplifiers, each consisting of 6 amplifier bundles; each bundle in turn contains 8 amplifying laser-glass plates stacked inside a flashlamp-pumped cavity, as illustrated in Fig. 15.3-4. Each of the 192 beamlines contains 16 separate amplification stages, so the overall system contains 3072 phosphate-glass laser-amplifier plates. The plates are rectangular, rather than square, since they are mounted at the Brewster angle to the direction of beam propagation to minimize Fresnel reflection and maximize flashlamp coupling. The laser amplifiers boost the pulse energy in each of the 192 beams to 20 kJ, so that the pulse energy is 4 MJ for the combined beam. For a 4-ns duration pulse, this corresponds to a peak power of 1 PW. The overall amplification provided by this exceptional MOPA is thus $G \approx 4 \times 10^{15}$. The optical arrangement of the system is such that the distance traveled by each optical pulse from the seed laser to the target is 1.5 km, entailing a travel time of 5 μs. The beam can be fired several times per day without creating undue heating.

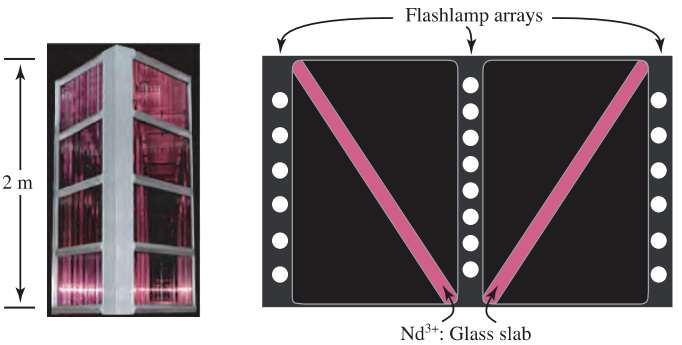


Figure 15.3-4 (a) A bundle of amplifiers comprises eight laser-glass plates stacked inside a flashlamp-pumped cavity at the National Ignition Facility (NIF) at LLNL. Each plate, which is made of specially formulated phosphate laser glass with a neodymium doping level $\approx$ 2 mol% (Schott LG-770 or Hoya LHG-8), measures 3.4 cm $\times$ 46 cm $\times$ 81 cm and weighs 42 kg. The height of the eight-amplifier bundle is $\approx$ 2 m. Six such bundles make up a cluster, and four clusters comprise the 192 individual beamlines at the NIF. Each beamline in turn contains 16 separate amplification stages, so the overall system contains 3072 glass laser-amplifier plates. The purplish color of the glass, when illuminated by white light and viewed in transmission, is a result of the red, yellow, and green absorption bands of the neodymium-doped phosphate glass portrayed in Fig. 15.3-3. (b) Top view of the linear flashlamps and amplifying Nd^{3+}:glass laser plates in a bundle. The system contains 7 680 flashlamps.

Laser-diode in-band pumping. The Nd^{3+}-doped glass laser amplifiers considered in Example 15.3-1 can be pumped with far greater efficiency by replacing the flash-lamps with laser-diode arrays, which efficiently convert electrical power to optical power. Together with the substantially reduced quantum defect associated with laser-diode pumping, this gives rise to a factor of 20 improvement in the overall efficiency of amplifier operation. The net result is a substantial decrease in heating, which in turn permits operation at a far higher repetition rate. An example of laser-diode in-band pumping is provided by the HAPLS petawatt laser system considered in Example 23.2-3. The power-amplifier portion of the pump laser for HAPLS contains a collection of neodymium-doped, phosphate-glass, laser-amplifier slabs similar to those used at the NIF. The slabs are pumped by AlGaAs laser-diode arrays that deliver high-power optical pulses at a repetition rate of 10 pulses/s. *In-band pumping* at 888 nm on the ${}^4I_{9/2} \rightarrow {}^4F_{3/2}$ Nd^{3+} transition results in the ions in the lower manifold being directly pumped into the upper manifold. In-band pumping is suitable even though the discrete sublevels within the manifolds are smeared into bands for this inhomogeneously broadened medium.

C. Erbium-Doped Silica Fiber

Rare-earth-doped fiber amplifiers. Optical fiber amplifiers (OFAs) are useful devices that offer the concomitant advantages of optical amplification and single-mode guided-wave confinement. In **rare-earth-doped fiber amplifiers** (**REFAs**), both the signal and pump are introduced into a doped glass fiber. The pump excites the ions to a higher energy level, usually via quasi-three-level or in-band pumping, thereby enabling signal amplification via downward stimulated-emission transitions. Active rare-earth ions typically incorporated in REFAs include Er^{3+}, Yb^{3+}, Tm^{3+}, Nd^{3+}, Pr^{3+}, and Ho^{3+}. Since the fiber must be transparent in the wavelength range of interest, a variety of glasses are used: e.g., silicates, phosphates, fluorides, germanates, tellurites, and chalcogenides (Sec. 10.5). Along with the active ion, the composition of the glass plays a contributing role in determining the absorption and emission transition cross sections, transition bandwidth, metastable-state lifetimes, maximum attainable dopant concentrations, fiber nonlinearities, and refractive index of the laser medium. REFAs usually operate in the near infrared and make use of silicate-based fibers because of their superior optical and mechanical properties. In particular, doped silica glasses, such as phosphosilicates, germanosilicates, and aluminosilicates, facilitate energy transfer and accommodate higher dopant concentrations than pure silica glass. Strong optical confinement and long fiber lengths endow REFAs with many salutary features.

REFA pumping configurations. Pumping may be implemented by longitudinally coupling light into the amplifying fiber, usually via dichroic couplers. As illustrated in Fig. 15.3-5, the pump light may be injected: (a) in the same direction as the signal (the forward direction); (b) in the opposite direction from the signal (the backward direction); or (c) in both directions (bidirectionally).

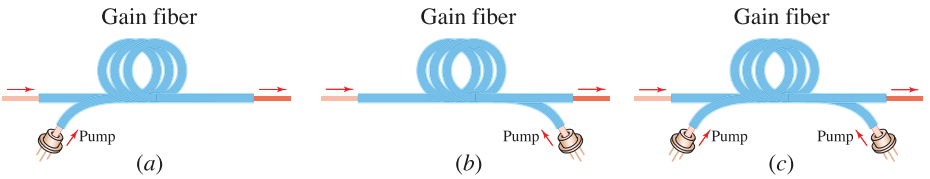


Figure 15.3-5 Longitudinal pumping of a rare-earth-doped fiber amplifier. The pumping may be (a) in the forward direction; (b) in the backward direction; or (c) bidirectional. Erbium-doped silica-fiber amplifiers are often pumped by fiber-coupled, strained-layer, InGaAs laser-diode arrays operated at $\lambda_o = 980$ nm. Similar pumping schemes are used for Raman fiber amplifiers (Sec. 15.3D).

Amplified spontaneous emission (ASE), a fundamental source of noise in optical amplifiers (Sec. 15.5), plays an important role in determining the optimal pumping configuration; if ASE is negligible, the direction of pumping is immaterial. In the presence of substantial ASE losses in quasi-three-level fiber amplifiers, analysis reveals that backward pumping can reduce these losses and increase amplifier efficiency. ASE can also be diminished by making use of a cascade of fiber amplifiers and inserting filters between successive stages to reduce the ASE. Nevertheless, the advantage of backward over forward pumping diminishes as the signal power increases to a level where gain saturation sets in. If the criterion of importance is low-noise operation, rather than high efficiency, forward pumping is often the optimal choice.

Erbium-doped fiber amplifiers. Though rare-earth-doped fiber amplifiers find application in many contexts, their most important use is in optical fiber communication systems. The Er^{3+} ion exhibits a broad laser transition near $\lambda_o = 1550$ nm that fortuitously falls in the wavelength region of minimum loss for silica optical fibers (Fig. 10.3-2). Since these fibers underlie optical fiber communications, **erbium-doped fiber amplifiers** (**EDFAs**) find extensive use in these systems (Sec. 25.1C). EDFAs exhibit many desirable properties, such as high gain, high output power, high efficiency, broad bandwidth, low insertion loss, low noise, and polarization insensitivity.

As illustrated in Fig. 15.3-6, the 980-nm-pumped Er^{3+}: silica-fiber amplifier operates on the ${}^4I_{13/2} \rightarrow {}^4I_{15/2}$ transition at a wavelength in the vicinity of $\lambda_o = 1550$ nm. It behaves as a *quasi-three-level system* at $T = 300^\circ$ K and as a *four-level system* when cooled to $T = 77^\circ$ K, in which case the thermal population in the lower manifold is reduced. Several key parameters related to this transition, and to the laser oscillator that makes use of it, are summarized in Tables 15.3-1 and 16.3-1, respectively. Appealing features of this laser transition include a long excited-state spontaneous lifetime ($t_{sp} \approx 10$ ms), the absence of intermediate energy levels between the excited and ground states, and the absence of excited-state absorption. Since Er^{3+} is a lanthanide-metal ion (Table 14.1-1), the host in which the ions are embedded plays a limited role in determining the energies of the manifolds, as explained in Sec. 14.1B. The broadening is a mixture of homogeneous (phonon-mediated) and inhomogeneous (arising from local field variations in the glass). Ytterbium is usually added to erbium as a co-dopant since the larger cross section of Yb^{3+} results in more efficient absorption of the pump photons at 980 nm; the energy is then transferred to the Er^{3+} ions, elevating them to the ${}^4I_{11/2}$ level as if they had directly absorbed the pump photons. The role of ytterbium in this context is analogous to the role of He in the He–Ne laser (Fig. 14.1-2).

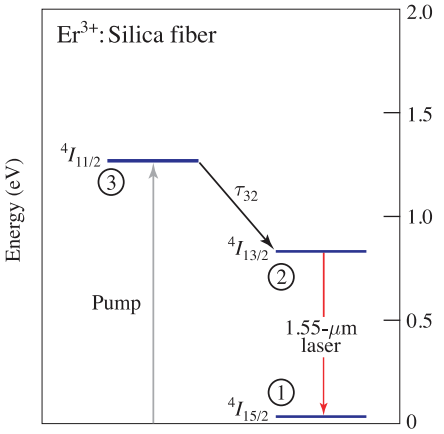


Figure 15.3-6 Schematic of Er^{3+}: silica-fiber energy-level manifolds for the ${}^4I_{13/2} \rightarrow {}^4I_{15/2}$ laser transition near 1550 nm. When pumped at 980 nm, this laser amplifier behaves as a quasi-three-level system at room temperature; the three interacting energy levels are indicated by encircled numbers. Laser amplification can also be implemented on this transition via in-band pumping at 1480 nm. Operation as a four-level system is also possible on the ${}^4I_{11/2} \rightarrow {}^4I_{13/2}$ transition, at a wavelength in the vicinity of 2.9 μm.

EDFAs can exhibit gains in excess of 50 dB with tens of mW of pump power. As a specific example, a gain of $\approx$ 30 dB is obtained by launching $\approx$ 5 mW of pump

power at 980 nm into a roughly 50-m length of silica fiber containing ≈ 300 ppm (by weight) of Er_2O_3. The highest gain efficiencies hover at about ≈ 10 dB/mW. Signal output powers in excess of 100 W can be generated since the output power increases in proportion to the pump power. The available bandwidth is $\Delta\lambda \approx 40$ nm, corresponding to $\Delta\nu \approx 5.3$ THz, which accommodates the C (conventional) telecommunications band that extends from 1530 to 1565 nm (Fig. 25.1-2). The L (long) telecommunications band, which stretches from 1565 to 1625 nm, can also be accommodated by choosing different optimization parameters for the EDFA. The large gain–bandwidth product offered by these amplifiers makes them ideal for use in wavelength-division multiplexing (WDM) systems (Sec. 25.3C). The 37% quantum defect [see (15.2-32)] associated with the quasi-three-level pumping scheme portrayed in Fig. 15.3-6 can be reduced by making use of *in-band pumping* at 1480 nm on the ${}^4I_{13/2} \rightarrow {}^4I_{15/2}$ laser transition. Pumping can then be implemented by using InGaAsP laser diodes or a Raman fiber laser (Sec. 16.3C). High-power EDFAs can be simultaneously pumped at 980 and 1480 nm.

The introduction of feedback readily converts fiber amplification into oscillation, as discussed in Sec. 16.3B. When the pump power is high, as is often the case in fiber laser oscillators, **double-clad fiber configurations** are usually used to avoid deleterious nonlinear optical effects in the fiber core (Fig. 16.3-4).

Thulium- and praseodymium-doped fiber amplifiers. Other rare-earth-doped fiber amplifiers useful for optical fiber communications include Tm^{3+}-doped multicomponent-silicate-glass REFAs operating in the 1460–1530-nm S (short) telecommunications band and Pr^{3+}-doped REFAs operating in the 1300-nm O (original) band. However, neither of these REFAs offer the exceptional gain and efficiency of Er^{3+}-doped devices. Other classes of commonly used optical amplifiers include Raman fiber amplifiers (RFAs), which are discussed and compared with EDFAs in Sec. 15.3D, and semiconductor optical amplifiers (SOAs), which are examined in Sec. 18.2 and compared with EDFAs in Sec. 18.2D.

Summary

The Er^{3+}: silica-fiber amplifier is widely used for optical fiber communications by virtue of its many salutary features:

- High gain
- High output power
- High efficiency
- Broad bandwidth
- Low insertion loss
- Low noise
- Polarization insensitivity

D. Raman Fiber Amplifiers

The class of optical fiber amplifiers (OFAs) extends beyond erbium-doped and rare-earth-doped devices. An important version of the OFA, known as the **Raman fiber amplifier** (RFA), relies on stimulated Raman scattering. The RFA thus operates on the basis of principles other than a population inversion and stimulated emission.

As discussed in Sec. 14.5C, **stimulated Raman scattering** (**SRS**) occurs when a pump photon of energy $h\nu_p$, together with a signal photon of lower energy $h\nu_s$, enter a nonlinear optical medium such as an optical fiber. The nonresonant version of the

process is illustrated in the inset in Fig. 15.3-7; the dashed horizontal line represents a virtual state. The signal photon stimulates the emission of a clone signal photon, which is obtained by Stokes-shifting the pump photon by the Raman energy $h\nu_R$ so that the energy of the clone photon precisely matches that of the incident signal photon. The surplus energy from the pump photon is transferred to the vibrational modes of the glass fiber. As is clear from (22.3-15), the strength of the effect, which is embodied in the Raman gain coefficient γ_R, depends on the nonlinear properties of the glass fiber. It is proportional to the pump intensity $I_p = P/A$, where P is the pump power and A is the interaction area. The polarization of SRS is the same as that of the exciting field.

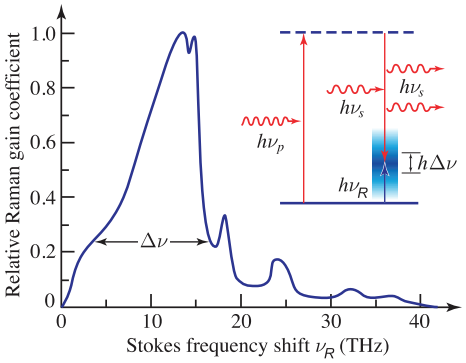


Figure 15.3-7 Stimulated Raman scattering (SRS) is schematized in the inset. Raman gain is available over a range of Stokes frequencies determined by the vibrational characteristics of the medium. In germanium-doped silica fiber, the peak Raman gain coefficient lies at a frequency below that of the pump by $\nu_R \approx 13$ THz and has a bandwidth $\Delta\nu \approx 12.5$ THz. (Gain curve adapted from R. H. Stolen, C. Lee, and R. K. Jain, Development of the Stimulated Raman Spectrum in Single-Mode Silica Fibers, *Journal of the Optical Society of America B*, vol. 1, pp. 652–657, 1984, Fig. 5.)

RFAs can be either distributed or lumped. In the **distributed Raman fiber amplifier**, the signal and pump are both sent through the transmission fiber, which serves as the gain medium. The **lumped Raman fiber amplifier**, in contrast, makes use of a short length of highly nonlinear fiber dedicated to providing gain. The core is generally made small to increase the pump intensity I_p and thereby to reduce the length of fiber required, which can be considerable.

Raman fiber amplifiers offer substantially broader bandwidths than EDFAs. The bandwidth over which Raman amplification obtains is governed by the vibrational spectrum of the glass host rather than by a transition linewidth, as in stimulated-emission lasers. Silicate, germanate, phosphate, and borate glasses exhibit very different SRS spectra and magnitudes. As is evident in Fig. 15.3-7, the dominant peak in the Raman gain coefficient for germanium-doped silica fiber is Stokes shifted from the pump frequency by approximately $\nu_R = 13$ THz, corresponding to about 100 nm at $\lambda_o = 1550$ nm. The bandwidth over which substantial Raman gain is available is of about the same magnitude as the shift of the peak, namely $\Delta\nu \approx 12.5$ THz, again corresponding to about $\Delta\lambda \approx 100$ nm at 1550 nm. Phosphosilicate glass fibers offer even greater Stokes shifts (Example 16.3-3).

RFA pumping can be achieved by making use of polarization-diverse laser diodes, fiber lasers, or Raman fiber lasers, operated at a wavelength about 100 nm below that desired for amplification (if the medium is germanium-doped silica fiber). As with the EDFA, the pump may be injected in the forward direction, in the backward direction, or bidirectionally (Fig. 15.3-5); backward pumping is generally employed since it reduces the noise transferred from the pump to the signal. Combining multiple pumps at different frequencies can substantially broaden the available bandwidth since the Stokes shift is linked to the pump frequency. In principle, Raman amplification can be employed over the entire region of fiber transparency.

Raman fiber amplifiers offer gains reaching 20 dB. The RFA gain efficiency in germanium-doped silica fiber is ≈ 0.02 dB/mW, which is to be compared with a gain efficiency ≈ 10 dB/mW for an EDFA. Hence, the pump power required for achieving

useful levels of Raman gain in such a distributed amplifier is typically hundreds of mW, far greater than that required for an EDFA. In lumped Raman amplifiers, pump powers in excess of 1 W can be used. Unlike EDFAs, polarization-diverse pumping is required since the Raman gain is maximized when the signal and pump beams have the same polarization.

Though RFA efficiencies are substantially lower than those offered by EDFAs, they can be measurably enhanced by combining gain and dispersion compensation to accommodate the different frequencies of the signal and pump pulses in a single fiber. RFAs are also sometimes used in conjunction with EDFAs; hybrid amplifiers that make use of low-noise Raman amplification together with high-power erbium-doped fiber amplification can accommodate larger repeater spacings and increased system capacity, but this comes at the expense of increased electrical power consumption. Raman fiber amplifiers are particularly useful in those telecommunications windows where EDFAs and other REFAs are unavailable or inefficient.

Brillouin fiber amplifiers. These devices make use of **stimulated Brillouin scattering** (**SBS**) [Fig. 14.5-5(*d*)] and behave in a manner analogous to that of Raman fiber amplifiers, but the interaction is with acoustic rather than optical phonons. Hence, the Brillouin frequency shift and bandwidth are orders of magnitude smaller than the Raman frequency shift and bandwidth in the same material. In silica fibers, for example the Brillouin shift and bandwidth are roughly 10 GHz and 100 MHz, respectively, whereas the Raman shift and bandwidth are both of the order of 13 THz.

Summary

The Raman fiber amplifier enjoys both advantages and disadvantages in comparison with the erbium-doped fiber amplifier:

Advantages of the RFA relative to the EDFA:

- Wider bandwidth
- Bandwidth extendable by use of multiple pumps
- Operation over a broad range of wavelengths
- Arbitrary fiber host
- Compatible with existing fiber links
- Lower noise
- Higher saturation power

Disadvantages of the RFA relative to the EDFA:

- Lower gain
- Greater pump power required
- Lower efficiency
- Longer fiber lengths required
- Sensitivity to signal polarization
- Stringent requirements for fiber and splice maintenance

E. Tabulation of Selected Laser Transitions

In practice, the most widely used laser amplifiers are solid-state, rare-earth-doped fiber, and Raman fiber devices, as exemplified by the examples considered in Secs. 15.3B, 15.3C, and 15.3D, respectively. However, laser amplification can also be provided by gases, dyes, exciplexes, ionized atoms, free-electron systems, and semiconductors. Table 15.3-1 provides a list of the wavelengths, cross sections, spontaneous lifetimes,

transition linewidths, and refractive indices for a number of representative laser transitions that operate in spectral regions stretching from the infrared to the X-ray domains. It is evident that the values of λ_o, σ_0, $t_{\rm sp}$, and $\Delta\nu$ vary over a broad range.

Table 15.3-1 Characteristics of some familiar laser transitions.

Laser Medium	Transition Wavelength[a] λ_o (nm)	Transition Cross Section[b] σ_0 ($\rm cm^2$)	Spontaneous Lifetime $t_{\rm sp}$	Transition Linewidth[c] $\Delta\nu$		Refractive Index n
Cu^+ (Kα)	0.154	7×10^{-18}	2 fs	500 THz	H	≈ 1
Ne^+ (Kα)	1.46	1×10^{-16}	130 fs	65 THz	H	≈ 1
C^{5+}	18.2	5×10^{-16}	12 ps	1 THz	I	≈ 1
ArF Exciplex	193	3×10^{-16}	10 ns	10 THz	I	≈ 1
Ar^+	515	3×10^{-12}	10 ns	3.5 GHz	I	≈ 1
Rhodamine-6G dye	560–640	2×10^{-16}	5 ns	40 THz	H/I	1.40
Ne (He–Ne)	633	3×10^{-13}	250 ns	1.5 GHz	I	≈ 1
Cr^{3+}:Al_2O_3 (ruby)	694	2×10^{-20}	3 ms	330 GHz	H	1.76
Cr^{3+}:$BeAl_2O_4$ (alexandrite)	700–820	7×10^{-21}	260 μs	25 THz	H	1.75
Ti^{3+}:Al_2O_3	700–1050	3×10^{-19}	3.9 μs	100 THz	H	1.76
Yb^{3+}:YAG	1030	2×10^{-20}	1 ms	5 THz	H	1.82
Nd^{3+}:Glass (phosphate)	1053	4×10^{-20}	370 μs	7 THz	I	1.50
Nd^{3+}:YAG	1064	3×10^{-19}	230 μs	150 GHz	H	1.82
Nd^{3+}:YVO_4	1064	3×10^{-18}	90 μs	260 GHz	H	2.0
Cr^{4+}:Mg_2SiO_4 (forsterite)	1100–1400	1×10^{-19}	3 μs	50 THz	H	1.65
InGaAsP[d]	1300–1600	2×10^{-16}	2.5 ns	10 THz	H	3.54
Er^{3+}:Silica fiber	1550	6×10^{-21}	10 ms	5 THz	H/I	1.46
Cr^{2+}:ZnS	1900–3000	1×10^{-18}	5 μs	40 THz	H	2.27
CO_2	10 600	3×10^{-18}	3 s	60 MHz	I	≈ 1

[a] The free-space wavelength shown in the table represents the most commonly used transition in each laser medium. The He–Ne gas laser system, for example, is most often used on the red-orange line at 0.633 μm, but it is also extensively used at 0.543, 1.15, and 3.39 μm (it also has laser transitions at hundreds of other wavelengths stretching to ≈ 100 μm).

[b] The value reported is the peak of the effective emission cross section $\sigma_{\rm em}(\nu)$.

[c] Values reported for gases such as CO_2 are typical for low-pressure operation (the atomic linewidth in a gas depends on its pressure because of collision broadening, which is homogeneous). H and I indicate line broadening dominated by homogeneous and inhomogeneous mechanisms, respectively.

[d] Values are for $In_{0.72}Ga_{0.28}As_{0.6}P_{0.4}$ assuming an injected carrier concentration of $\Delta n = 1.8\times10^{18}$ $\rm cm^{-3}$ (see Examples 18.2-1–18.2-3).

15.4 AMPLIFIER NONLINEARITY

A. Saturated Gain in Homogeneously Broadened Media

Gain Coefficient

The following relationships have been established in earlier sections of this chapter: 1) the gain coefficient $\gamma(\nu)$ of a laser medium depends on the population difference N, in accordance with (15.1-4); 2) N is in turn governed by the pumping rate R, in accordance with (15.2-15); 3) N also depends on the transition rate W_i, in accordance with (15.2-10); and 4) the probability density W_i in turn depends on the radiation photon-flux density ϕ, in accordance with (15.1-1). It follows that the gain coefficient of a laser medium is dependent on the photon-flux density to be amplified. This is the origin of amplifier nonlinearity and gain saturation, as we now demonstrate.

Substituting (15.1-1) into (15.2-10) yields

$$N = \frac{N_0}{1 + \phi/\phi_s(\nu)} \tag{15.4-1}$$

where

$$\boxed{\frac{1}{\phi_s(\nu)} = \tau_s\,\sigma(\nu) = \frac{\lambda^2}{8\pi}\frac{\tau_s}{t_{\rm sp}}\,g(\nu)\,,} \tag{15.4-2}$$

(15.4-2) Saturation Photon-Flux Density

thereby providing an expression for the dependence of the population difference N on the photon-flux density ϕ. Now, substituting (15.4-1) into the expression for the gain coefficient (15.1-4) leads directly to the **saturated gain coefficient** for homogeneously broadened media:

$$\boxed{\gamma(\nu) = \frac{\gamma_0(\nu)}{1 + \phi/\phi_s(\nu)}\,,} \tag{15.4-3}$$

(15.4-3) Saturated Gain Coefficient

where

$$\boxed{\gamma_0(\nu) = N_0\,\sigma(\nu) = N_0\frac{\lambda^2}{8\pi t_{\rm sp}}\,g(\nu)\,.} \tag{15.4-4}$$

(15.4-4) Small-Signal Gain Coefficient

The gain coefficient thus decreases as the photon-flux density ϕ increases, as illustrated in Fig. 15.4-1. Since the quantity $\phi_s(\nu) = 1/\tau_s\sigma(\nu)$ represents the photon-flux density at which the gain coefficient decreases to half its maximum value, it is called the **saturation photon-flux density**. When $\tau_s \approx t_{\rm sp}$, the interpretation of $\phi_s(\nu)$ is straightforward: (15.4-2) provides that $\phi_s(\nu)\,t_{\rm sp}\,\sigma(\nu) = 1$ so that roughly one photon is emitted during each spontaneous emission lifetime into each transition cross-sectional area.

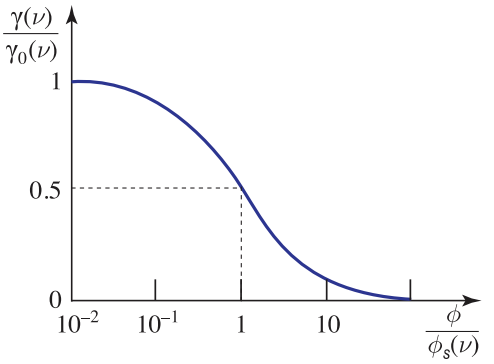


Figure 15.4-1 Dependence of the normalized saturated gain coefficient $\gamma(\nu)/\gamma_0(\nu)$ on the normalized photon-flux density $\phi/\phi_s(\nu)$. When ϕ is equal to its saturation value $\phi_s(\nu)$, the gain coefficient is reduced to half its unsaturated value.

EXERCISE 15.4-1

Saturation Photon-Flux Density for Ruby. Determine the saturation photon-flux density, and the corresponding saturation intensity, for the $\lambda_o = 694.3$-nm ruby laser transition at $\nu = \nu_0$. Use the parameters provided in Table 15.3-1. Assume that $\tau_s \approx 2t_{\rm sp}$, in accordance with (15.2-28).

EXERCISE 15.4-2

Spectral Broadening of a Saturated Amplifier. Consider a homogeneously broadened amplifying medium with a Lorentzian lineshape function of width $\Delta\nu$ [see (15.1-8)]. Show that for a photon-flux density ϕ, the amplifier gain coefficient $\gamma(\nu)$ assumes a Lorentzian lineshape function of width

$$\Delta\nu_s = \Delta\nu\sqrt{1+\frac{\phi}{\phi_s(\nu_0)}}\,. \tag{15.4-5}$$

Linewidth of Saturated Amplifier

This demonstrates that gain saturation is accompanied by an increase in bandwidth, corresponding to reduced frequency selectivity, as illustrated in Fig. 15.4-2.

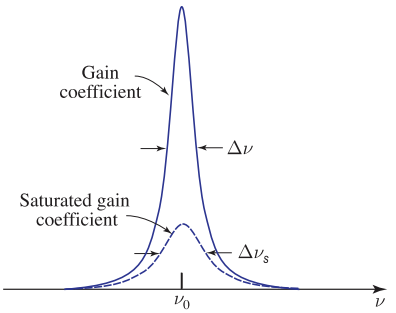


Figure 15.4-2 Gain coefficient reduction and bandwidth increase resulting from saturation when $\phi = 2\phi_s(\nu_0)$.

Gain

Having determined the effect of saturation on the gain coefficient (gain per unit length), we embark on a determination of the behavior of the **saturated gain** for a homogeneously broadened laser amplifier of length d [Fig. 15.4-3(a)]. For simplicity, we suppress the frequency dependencies of $\gamma(\nu)$ and $\phi_s(\nu)$ and write γ and ϕ_s instead.

If the photon-flux density at position z is $\phi(z)$, then in accordance with (15.4-3) the gain coefficient at that position is also a function of z. We know from (15.1-3) that the incremental increase in photon-flux density at the position z is $d\phi = \gamma\phi\,dz$, which leads to the differential equation

$$\frac{d\phi}{dz} = \frac{\gamma_0\phi}{1+\phi/\phi_s}\,. \tag{15.4-6}$$

Rewriting this equation as $(1/\phi + 1/\phi_s)\,d\phi = \gamma_0\,dz$, and integrating, we obtain

$$\ln\frac{\phi(z)}{\phi(0)} + \frac{\phi(z)-\phi(0)}{\phi_s} = \gamma_0 z\,. \tag{15.4-7}$$

The relation between the photon-flux densities at the input and output, $\phi(0)$ and $\phi(d)$, respectively, is therefore

$$[\ln(Y)+Y] = [\ln(X)+X] + \gamma_0 d, \tag{15.4-8}$$

where $X = \phi(0)/\phi_s$ and $Y = \phi(d)/\phi_s$ are the input and output photon-flux densities normalized to the saturation photon-flux density, respectively.

It is useful to examine the solution for the gain $G = \phi(d)/\phi(0) = Y/X$ in two limiting cases:

1. If both X and Y are much smaller than unity (i.e., the photon-flux densities are much smaller than the saturation photon-flux density), then X and Y are negligible in comparison with $\ln(X)$ and $\ln(Y)$, whereupon we obtain the approximate solution $\ln(Y) \approx \ln(X) + \gamma_0 d$, from which

$$Y \approx X \exp(\gamma_0 d). \tag{15.4-9}$$

   In this case the relation between Y and X is linear, and the gain $G = Y/X \approx \exp(\gamma_0 d)$ [leftmost dashed curve in Fig. 15.4-3(*b*)]. This accords with (15.1-7), which was obtained under the small-signal approximation, valid when the gain coefficient is independent of the photon-flux density, i.e., $\gamma \approx \gamma_0$.
2. When $X \gg 1$, we can neglect $\ln(X)$ in comparison with X and $\ln(Y)$ in comparison with Y in (15.4-8), whereupon

$$Y \approx X + \gamma_0 d \tag{15.4-10}$$

   which, with the help of (15.4-2) and (15.4-4), yields

$$\begin{aligned} \phi(d) &\approx \phi(0) + \gamma_0 \phi_s d \\ &\approx \phi(0) + \frac{N_0 d}{\tau_s}. \end{aligned} \tag{15.4-11}$$

   Under these heavily saturated conditions, the atoms of the medium are "busy" emitting a constant photon-flux density $N_0 d/\tau_s$. Incoming input photons therefore simply leak through to the output, augmented by a constant photon-flux density that is independent of the amplifier input [short dashed curve at right in Fig. 15.4-3(*b*)].

For intermediate values of X and Y, (15.4-8) must be solved numerically. A plot of the solution is shown as the solid curve in Fig. 15.4-3(*b*). The linear input–output relationship obtained for $X \ll 1$, and the saturated relationship for $X \gg 1$, are evident as limiting cases of the numerical solution. The gain $G = Y/X$ for $\gamma_0 d = 2$ is plotted in Fig. 15.4-3(*c*). It achieves its maximum value $\exp(\gamma_0 d)$ for small values of the input photon-flux density ($X \ll 1$), and decreases toward unity as $X \to \infty$.

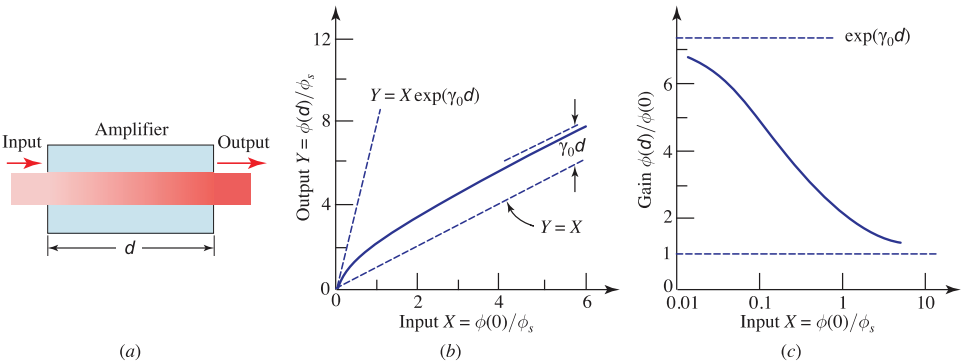


Figure 15.4-3 (*a*) A nonlinear (saturated) amplifier. (*b*) Relation between the normalized output photon-flux density $Y = \phi(d)/\phi_s$ and the normalized input photon-flux density $X = \phi(0)/\phi_s$. For $X \ll 1$, the gain $G = Y/X \approx \exp(\gamma_0 d)$. For $X \gg 1$, we obtain $Y \approx X + \gamma_0 d$. The numerical solution for (15.4-8) is indicated by the solid curve. (*c*) Gain as a function of the input normalized photon-flux density X for a saturated amplifier of length d with $\gamma_0 d = 2$.

Saturable Absorbers

If the gain coefficient γ_0 is negative, i.e., if the population difference is normal rather than inverted ($N_0 < 0$), the medium provides attenuation rather than amplification. The attenuation coefficient $\alpha(\nu) = -\gamma(\nu)$ then also suffers from saturation, in accordance with the relation $\alpha(\nu) = \alpha_0(\nu)/[1 + \phi/\phi_s(\nu)]$, which is analogous to (15.4-3). This indicates that there is less absorption for large values of the photon-flux density. A material that exhibits this property is called a **saturable absorber**.

The relation between the output and input photon-flux densities, $\phi(d)$ and $\phi(0)$, respectively, for an absorber of length d is governed by (15.4-8) with negative γ_0. The overall transmittance of the absorber $Y/X = \phi(d)/\phi(0)$ is presented as a function of $X = \phi(0)/\phi_s$ as the solid curve in Fig. 15.4-4. The transmittance increases with increasing $\phi(0)$, ultimately reaching a limiting value of unity. This effect occurs because the population difference $N \to 0$, so that there is no net absorption.

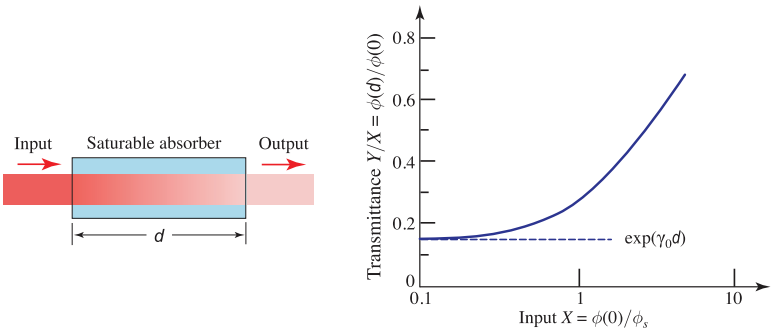


Figure 15.4-4 The transmittance of a saturable absorber $Y/X = \phi(d)/\phi(0)$ versus the normalized photon-flux density $X = \phi(0)/\phi_s$, for $\gamma_0 d = -2$. The transmittance increases with increasing input photon-flux density.

*B. Saturated Gain in Inhomogeneously Broadened Media

Gain Coefficient

An inhomogeneously broadened medium comprises a collection of atoms containing subsets thereof with different properties. As discussed in Sec. 14.3D, the subset of atoms labeled β has a homogeneously broadened lineshape function $g_\beta(\nu)$. The overall inhomogeneous average lineshape function of the medium is described by $\overline{g}(\nu) = \langle g_\beta(\nu) \rangle$, where $\langle \cdot \rangle$ represents an average with respect to β. Because the small-signal gain coefficient $\gamma_0(\nu)$ is proportional to $g(\nu)$, as provided in (15.4-4), different subsets β of atoms have different gain coefficients $\gamma_{0\beta}(\nu)$. The average *small-signal gain coefficient* is therefore

$$\overline{\gamma}_0(\nu) = N_0 \frac{\lambda^2}{8\pi t_{\text{sp}}} \overline{g}(\nu). \tag{15.4-12}$$

Solving for the *saturated gain coefficient* is more subtle, however. This is because the saturation photon-flux density $\phi_s(\nu)$, which is inversely proportional to $g(\nu)$ as provided in (15.4-2), is itself dependent on the subset of atoms β. An average gain coefficient may be defined by using (15.4-2) and (15.4-3),

$$\overline{\gamma}(\nu) = \langle \gamma_\beta(\nu) \rangle, \tag{15.4-13}$$

where

$$\gamma_\beta = \frac{\gamma_{0\beta}(\nu)}{1 + \phi/\phi_{s\beta}(\nu)} = b\,\frac{g_\beta(\nu)}{1 + \phi\, a^2 g_\beta(\nu)}\,, \tag{15.4-14}$$

with $b = N_0(\lambda^2/8\pi t_{\rm sp})$ and $a^2 = (\lambda^2/8\pi)(\tau_s/t_{\rm sp})$. Evaluating the average of (15.4-14) requires care because the average of a ratio is not equal to the ratio of the averages.

Doppler-Broadened Medium

Though all of the atoms in a **Doppler-broadened medium** share a lineshape function $g(\nu)$ of identical shape, the center frequency of the subset β is shifted by an amount ν_β proportional to the velocity v_β of the subset. If $g(\nu)$ is Lorentzian with width $\Delta\nu$, (15.1-8) provides $g(\nu) = (\Delta\nu/2\pi)/[(\nu - \nu_0)^2 + (\Delta\nu/2)^2]$ and $g_\beta(\nu) = g(\nu - \nu_\beta)$. Substituting $g_\beta(\nu)$ into (15.4-14) provides

$$\gamma_\beta(\nu) = \frac{b\,(\Delta\nu/2\pi)}{(\nu - \nu_\beta - \nu_0)^2 + (\Delta\nu_s/2)^2}\,, \tag{15.4-15}$$

where

$$\Delta\nu_s = \Delta\nu\sqrt{1 + \frac{\phi}{\phi_s(\nu_0)}} \tag{15.4-16}$$

and

$$\phi_s^{-1}(\nu_0) = \frac{2a^2}{\pi\Delta\nu} = \frac{\lambda^2}{8\pi}\frac{\tau_s}{t_{\rm sp}}\frac{2}{\pi\Delta\nu} = \frac{\lambda^2}{8\pi}\frac{\tau_s}{t_{\rm sp}}\,g(\nu_0). \tag{15.4-17}$$

Equation (15.4-16) was obtained for the homogeneously broadened saturated amplifier considered in Exercise 15.4-2 [see (15.4-5)]. It is evident that the subset of atoms with velocity v_β has a saturated gain coefficient $\gamma_\beta(\nu)$ with a Lorentzian shape of width $\Delta\nu_s$ that increases as the photon-flux density becomes larger.

The average of $\gamma_\beta(\nu)$ in (15.4-13) is readily calculated for a Doppler-broadened medium since the shifts ν_β follow a zero-mean Gaussian probability density function $p\,(\nu_\beta) = (2\pi\sigma_D^2)^{-1/2}\,\exp(-\nu_\beta^2/2\sigma_D^2)$, with standard deviation σ_D (Exercise 14.3-2). Thus, $\overline{\gamma}(\nu) = \langle\gamma_\beta(\nu)\rangle$ is given by

$$\overline{\gamma}(\nu) = \int_{-\infty}^{\infty} \gamma_\beta(\nu)\, p(\nu_\beta)\, d\nu_\beta. \tag{15.4-18}$$

If $p\,(\nu_\beta)$ is much broader than $\gamma_\beta(\nu)$ (i.e., the Doppler broadening is much wider than $\Delta\nu_s$), we may regard the broad function $p(\nu_\beta)$ as constant and remove it from the integral when evaluating $\overline{\gamma}(\nu_0)$. Setting $\nu = \nu_0$ and $\nu_\beta = 0$ in the exponential provides

$$\overline{\gamma}(\nu_0) = \frac{bp(0)}{\sqrt{1 + 2\phi\, a^2/\pi\Delta\nu}} = \frac{\overline{\gamma}_0}{\sqrt{1 + \phi/\phi_s(\nu_0)}}\,, \tag{15.4-19}$$

where the average small-signal gain coefficient $\overline{\gamma}_0$ is

$$\overline{\gamma}_0 = N_0 \frac{\lambda^2}{8\pi t_{\text{sp}}} \frac{1}{\sqrt{2\pi\sigma_D^2}}. \tag{15.4-20}$$

Equation (15.4-19) provides an expression for the average saturated gain coefficient of a Doppler broadened medium at the central frequency ν_0, as a function of the photon-flux density ϕ at $\nu = \nu_0$. The gain coefficient saturates as ϕ increases in accordance with a square-root law. The gain coefficient in an inhomogeneously broadened medium therefore saturates more slowly than the gain coefficient in a homogeneously broadened medium [see (15.4-3)], as illustrated in Fig. 15.4-5.

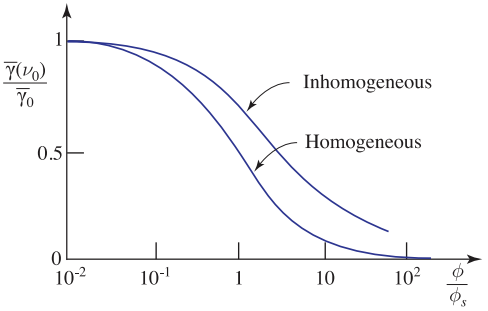


Figure 15.4-5 Comparison of gain saturation in homogeneously and inhomogeneously broadened media.

Hole Burning

When a large flux density of monochromatic photons at frequency ν_1 is applied to an inhomogeneously broadened medium, the gain saturates only for those atoms whose lineshape functions overlap ν_1. Other atoms simply do not interact with the photons and remain unsaturated. When the saturated medium is probed by a weak monochromatic light source of varying frequency ν, the profile of the gain coefficient therefore exhibits a hole centered about ν_1, as illustrated in Fig. 15.4-6. This phenomenon is known as **hole burning**. Since the gain coefficient $\gamma_\beta(\nu)$ of the subset of atoms with velocity v_β has a Lorentzian shape with width $\Delta\nu_s$ given by (15.4-16), it follows that the width of the hole is $\Delta\nu_s$. As the flux density of saturating photons at ν_1 increases, both the depth and the width of the hole increase.

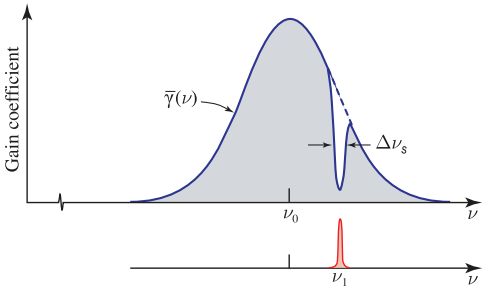


Figure 15.4-6 The gain coefficient of an inhomogeneously broadened medium is locally saturated by a large flux density of monochromatic photons at frequency ν_1.

*15.5 AMPLIFIER NOISE

The resonant medium that provides amplification via stimulated emission also generates spontaneous emission. The light arising from the latter process, which is independent of the input to the amplifier, represents a fundamental source of laser amplifier noise. Whereas the amplified signal has a specific frequency, direction, and

polarization, the noise associated with **amplified spontaneous emission** (**ASE**) is broadband, multidirectional, and unpolarized. As a consequence, it is possible to filter out some of this noise by following the amplifier with a narrowband optical filter, a collection aperture, and a polarizer.

The probability density (per second) that an atom in the upper laser level spontaneously emits a photon of frequency between ν and $\nu + d\nu$ is (Exercise 14.3-1):

$$P_{\text{sp}}(\nu)\, d\nu = \frac{1}{t_{\text{sp}}}\, g(\nu)\, d\nu. \tag{15.5-1}$$

The probability density of the spontaneous emission of a photon of any frequency is, of course, $P_{\text{sp}} = 1/t_{\text{sp}}$. If N_2 is the atomic density in the upper energy level, the average spontaneously emitted photon density is $N_2 P_{\text{sp}}(\nu)$. (The average spontaneously emitted power per unit volume per unit frequency is therefore $h\nu N_2 P_{\text{sp}}(\nu)$.) The spontaneously emitted photon density is emitted uniformly in all directions and is equally divided between the two polarizations. If the amplifier output is collected from a solid angle $d\Omega$, as illustrated in Fig. 15.5-1, and from only one of the polarizations, it contains only a fraction $\frac{1}{2}d\Omega/4\pi$ of the spontaneously emitted photon density. Furthermore, if a filter is used to limit the collected photons to a narrow frequency band of width B centered about the amplified signal frequency ν, the number of photons added per second by spontaneous emission from an incremental volume of unit area and length dz is $\xi_{\text{sp}}(\nu)\, dz$, where

$$\xi_{\text{sp}}(\nu) = N_2 \frac{1}{t_{\text{sp}}}\, g(\nu) B \frac{d\Omega}{8\pi} \tag{15.5-2}$$

is the noise photon-flux density per unit length.

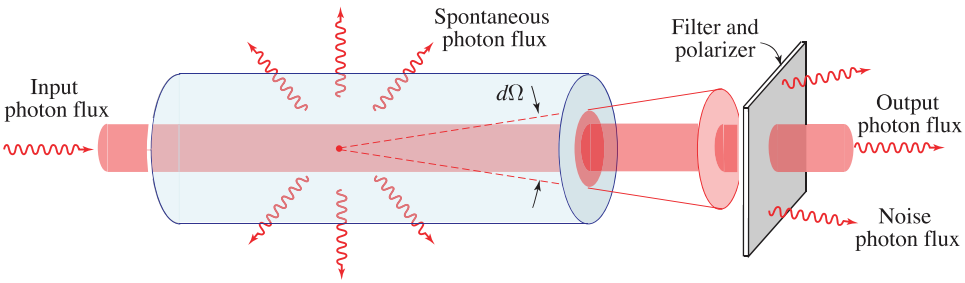


Figure 15.5-1 Spontaneous emission is a source of amplifier noise. It is broadband, radiated in all directions, and unpolarized. Optics can be used at the output of the amplifier to limit the spontaneous emission noise to a narrow optical band, an incremental solid angle $d\Omega$, and a single polarization.

In determining the noise photon-flux density contributed by the amplifier, the photon-flux density per unit length should not simply be multiplied by the length of the amplifier. This is because the spontaneous-emission noise is itself amplified by the medium; spontaneous-emission noise generated near the input end of the amplifier provides a greater contribution than noise generated near the output end. ASE noise can be accommodated by modifying the differential equation governing the growth of the photon-flux density provided in (15.1-3), so that it becomes

$$\frac{d\phi}{dz} = \gamma(\nu)\phi + \xi_{\text{sp}}(\nu). \tag{15.5-3}$$

Equation (15.5-3) incorporates the photon-flux density arising from both the amplified signal and the amplified spontaneous emission noise.

EXERCISE 15.5-1

Amplified Spontaneous Emission (ASE).

(a) Use (15.5-3) to show that, in the absence of any input signal, spontaneous emission produces a photon-flux density at the output of an unsaturated amplifier $[\gamma(\nu) = \gamma_0(\nu)]$ of length d that can be expressed as $\phi(d) = \phi_{\text{sp}}\{\exp[\gamma_0(\nu)d] - 1\}$, where $\phi_{\text{sp}} = \xi_{\text{sp}}(\nu)/\gamma_0(\nu)$.

(b) Since both $\xi_{\text{sp}}(\nu)$ and $\gamma_0(\nu)$ are proportional to $g(\nu)$, ϕ_{sp} is independent of $g(\nu)$ so that the frequency dependence of $\phi(d)$ is governed solely by the factor $\{\exp[\gamma_0(\nu)d] - 1\}$. If $\gamma_0(\nu)$ is Lorentzian with width $\Delta\nu$, i.e., $\gamma_0(\nu) = \gamma_0(\nu_0)(\Delta\nu/2)^2/[(\nu - \nu_0)^2 + (\Delta\nu/2)^2]$, show that the width of the factor $\{\exp[\gamma_0(\nu)d] - 1\}$ is smaller than $\Delta\nu$, i.e., that the amplification of spontaneous emission is accompanied by spectral narrowing.

Photon statistics after amplification. In the course of amplification, the photon-number statistics (Sec. 13.2C) of the incoming light are altered. A coherent signal presented to the input of the amplifier exhibits Poisson photon-number statistics, with a variance σ_S^2 equal to the mean signal photon number $\overline{n}_S$. The ASE photons, on the other hand, exhibit Bose–Einstein photon-number statistics with $\sigma_{\text{ASE}}^2 = \overline{n}_{\text{ASE}} + \overline{n}_{\text{ASE}}^2$, which has a substantially larger variance. The photon-number statistics of the light after amplification, comprising both signal and spontaneous-emission contributions, exhibits a photon-number distribution intermediate between the Poisson and Bose–Einstein distributions. If the detector counting time and area are small, and the emerging light is linearly polarized, the photon statistics at the amplifier output obey a special case of the **noncentral-negative-binomial photon-number distribution** (Prob. 15.5-3), which has a variance given by

$$\sigma_n^2 = \overline{n}_S + \left(\overline{n}_{\text{ASE}} + \overline{n}_{\text{ASE}}^2\right) + 2\overline{n}_S\overline{n}_{\text{ASE}}. \tag{15.5-4}$$

This expression contains contributions from the signal and spontaneous emission individually, as well as a cross term that involves both.

READING LIST

Rare-Earth, Raman, and Brillouin Fiber Amplifiers

E. Garmire, Perspectives on Stimulated Brillouin Scattering, *New Journal of Physics*, vol. 19, 011003, 2017.

N. K. Dutta, *Fiber Amplifiers and Fiber Lasers*, World Scientific, 2015.

V. V. Ter-Mikirtychev, *Fundamentals of Fiber Lasers and Fiber Amplifiers*, Springer-Verlag, 2014.

M. Premaratne and G. P. Agrawal, *Light Propagation in Gain Media: Optical Amplifiers*, Cambridge University Press, 2011.

A. Kobyakov, M. Sauer, and D. Chowdhury, Stimulated Brillouin Scattering in Optical Fibers, *Advances in Optics and Photonics*, vol. 2, pp. 1–59, 2010.

C. Headley and G. P. Agrawal, eds., *Raman Amplification in Fiber Optical Communication Systems*, Elsevier, 2005.

M. N. Islam, ed., *Raman Amplifiers for Telecommunications, Volume 1: Physical Principles*, Springer-Verlag, 2004.

M. N. Islam, ed., *Raman Amplifiers for Telecommunications, Volume 2: Sub-Systems and Systems*, Springer-Verlag, 2004.

E. Desurvire, D. Bayart, B. Desthieux, and S. Bigo, *Erbium-Doped Fiber Amplifiers: Device and System Developments*, Wiley, 2002.

M. J. F. Digonnet, ed., *Rare-Earth-Doped Fiber Lasers and Amplifiers*, CRC Press, 2nd ed. 2001.

P. C. Becker, N. A. Olsson, and J. R. Simpson, *Erbium-Doped Fiber Amplifiers: Fundamentals and Technology*, Academic Press, 1999.

E. Desurvire, *Erbium-Doped Fiber Amplifiers: Principles and Applications* Wiley, 1994.

M. J. Digonnet, ed., *Selected Papers on Rare Earth-Doped Fiber Laser Sources and Amplifiers*, SPIE Optical Engineering Press (Milestone Series Volume 37), 1992.

E. Desurvire, J. R. Simpson, and P. C. Becker, High-Gain Erbium-Doped Traveling-Wave Fiber Amplifier, *Optics Letters*, vol. 12, pp. 888–890, 1987.

R. J. Mears, L. Reekie, I. M. Jauncey, and D. N. Payne, Low-Noise Erbium-Doped Fibre Amplifier Operating at 1.54 μm, *Electronics Letters*, vol. 23, pp. 1026–1028, 1987.

S. B. Poole, D. N. Payne, and M. E. Fermann, Fabrication of Low Loss Optical Fibers Containing Rare Earth Ions, *Electronics Letters*, vol. 21, pp. 737–738, 1985.

E. P. Ippen and R. H. Stolen, Stimulated Brillouin Scattering in Optical Fibers, *Applied Physics Letters*, vol. 21, pp. 539–541, 1972.

R. H. Stolen, E. P. Ippen, and A. R. Tynes, Raman Oscillation in Glass Optical Waveguide, *Applied Physics Letters*, vol. 20, pp. 62–64, 1972.

C. J. Koester and E. Snitzer, Amplification in a Fiber Laser, *Applied Optics*, vol. 3, pp. 1182–1186, 1964.

H. W. Etzel, H. W. Gandy, and R. J. Ginther, Stimulated Emission of Infrared Radiation from Ytterbium-Activated Silicate Glass, *Applied Optics*, vol. 1, pp. 534–536, 1962.

C. V. Raman and K. S. Krishnan, A New Type of Secondary Radiation, *Nature*, vol. 121, pp. 501–502, 1928.

L. Brillouin, Diffusion de la lumière et des rayons X par un corps transparent homogène: Influence de l'agitation thermique [Diffusion of Light and X-rays by a Transparent Homogeneous Body: The Influence of Thermal Excitations], *Annales de Physique* (*Paris*), vol. 9, no. 17, pp. 88–122, 1922.

High-Power Pulsed Laser Amplifiers

L. Hu, D. He, H. Chen, X. Wang, T. Meng, L. Wen, J. Hu, Y. Xu, S. Li, Y. Chen, W. Chen, S. Chen, J. Tang, and B. Wang, Research and Development of Neodymium Phosphate Laser Glass for High Power Laser Application, *Optical Materials*, vol. 63, pp. 213–220, 2017.

M. L. Spaeth *et al.*, Description of the NIF Laser, *Fusion Science and Technology*, vol. 69, pp. 25–145, 2016.

R. Betti and O. A. Hurricane, Inertial-Confinement Fusion with Lasers, *Nature Physics*, vol. 12, pp. 435–448, 2016.

M. Rose, Laser Fusion, *Optics & Photonics News*, vol. 25, no. 9, pp. 34–41, 2014.

R. Courtland, Star Power, *IEEE Spectrum*, vol. 50 (4NA), pp. 26–32 & 57–58, 2013.

J. H. Campbell and T. I. Suratwala, Nd-Doped Phosphate Glasses for High-Energy/High-Peak-Power Lasers, *Journal of Non-Crystalline Solids*, vol. 263/264, pp. 318–341, 2000.

Optical Amplifier Noise and Photon Statistics

A. R. Karthikeyan and H. S. Ramachandran, Convergence of Quantum and Semiclassical Models of Erbium-Doped Fiber Amplifiers, *Journal of the Optical Society of America B*, vol. 28, pp. 533–542, 2011.

G. Kahraman and B. E. A. Saleh, Quantum-Statistical Properties of Pulse Amplification in Optical Fibers with Gain Saturation, *Journal of Lightwave Technology*, vol. 13, pp. 1127–1134, 1995.

T. Li and M. C. Teich, Photon Point Process for Traveling-Wave Laser Amplifiers, *IEEE Journal of Quantum Electronics*, vol. 29, pp. 2568–2578, 1993.

P. Diament and M. C. Teich, Evolution of the Statistical Properties of Photons Passed Through a Traveling-Wave Laser Amplifier, *IEEE Journal of Quantum Electronics*, vol. 28, pp. 1324–1334, 1992.

E. Desurvire, Analysis of Noise Figure Spectral Distribution in Erbium Doped Fiber Amplifiers Pumped Near 980 and 1480 nm, *Applied Optics*, vol. 29, pp. 3118–3125, 1990.

K. Shimoda, H. Takahasi, and C. H. Townes, Fluctuations in Amplification of Quanta with Application to Maser Amplifiers, *Journal of the Physical Society of Japan*, vol. 12, pp. 686–700, 1957.

PROBLEMS

15.1-2 **Amplifier Gain and Rod Length.** Consider a ruby laser amplifier that makes use of a 15-cm-long rod and has a small-signal gain of 12. Neglecting the effects of gain saturation, determine the small-signal gain of a 20-cm-long rod?

15.1-3 **Laser Amplifier Gain and Population Difference.** A 15-cm-long Nd^{3+}:glass rod used as a laser amplifier has a total small-signal gain of 10 at $\lambda_o = 1.06\ \mu m$. Use the data provided in Table 15.3-1 to determine the population difference N (Nd^{3+} ions per cm^3) required to achieve this gain.

15.1-4 **Amplification of a Broadband Signal.** The transition between two energy levels exhibits a Lorentzian lineshape of central frequency $\nu_0 = 5 \times 10^{14}$ with a linewidth $\Delta\nu = 1$ THz. The population is inverted so that the maximum gain coefficient $\gamma(\nu_0) = 0.1\ cm^{-1}$. The medium has an additional loss coefficient $\alpha_s = 0.05\ cm^{-1}$, which is independent of ν. Estimate the loss or gain encountered by a light wave traversing 1 cm if it has a uniform power spectral density centered about ν_0 with a bandwidth $2\Delta\nu$.

15.2-4 **Two-Level Pumping System.** Write the rate equations for a two-level system and demonstrate that a steady-state population inversion cannot be achieved by using direct optical pumping between levels ① and ②.

15.2-5 **Two Laser Lines.** Consider an atomic system with four levels: the ground state ⓪, ①, ②, and ③. Two pumps are applied: one between the ground state and level ③ at a rate R_3 and the other between the ground state and level ② at a rate R_2. Population inversion can occur between levels ③ and ① and/or between levels ② and ① (as in a four-level system). Assuming that decay from level ③ to level ② is not possible, and that decay from levels ③ and ② to the ground state are negligible, write the rate equations for levels ①, ②, and ③ in terms of the lifetimes τ_1, τ_{31}, and τ_{21}. Determine the steady-state population densities N_1, N_2, and N_3 and examine the possibility of simultaneous population inversions between levels ③ and ①, and between levels ② and ①. Show that the presence of radiation at the ②→① transition reduces the population difference for the ③→① transition.

15.4-3 **Significance of the Saturation Photon-Flux Density.** In the general two-level atomic system of Fig. 15.2-3, τ_2 represents the lifetime of level ② in the absence of stimulated emission. In the presence of stimulated emission, the rate of decay from level ② increases and the effective lifetime decreases. Find the photon-flux density ϕ at which the lifetime decreases to half its value. How is that photon-flux density related to the saturation photon-flux density ϕ_s?

15.4-4 **Saturation Optical Intensity.** Determine the saturation photon-flux density $\phi_s(\nu_0)$ and the corresponding saturation optical intensity $I_s(\nu_0)$, for the homogeneously broadened ruby and Nd^{3+}:YAG laser transitions by making use of the parameters provided in Table 15.3-1.

15.4-5 **Growth of the Photon-Flux Density in a Saturated Laser Amplifier.** The growth of the photon-flux density $\phi(z)$ in a saturated laser amplifier is described by (15.4-7). Plot $\phi(z)/\phi_s$ versus $\gamma_0 z$ for $\phi(0)/\phi_s = 0.05$. Identify the onset of saturation in this amplifier.

15.4-6 **Resonant Absorption of a Medium in Thermal Equilibrium.** A unity refractive-index medium of volume $1\ cm^3$ contains $N_a = 10^{23}$ atoms in thermal equilibrium. The ground state is energy level ①; level ② has energy 2.48 eV above the ground state ($\lambda_o = 0.5\ \mu m$). The transition between these two levels is characterized by a spontaneous lifetime $t_{sp} = 1$ ms, and a Lorentzian lineshape of width $\Delta\nu = 1$ GHz. Consider two temperatures, T_1 and T_2, such that $kT_1 = 0.026$ eV and $kT_2 = 0.26$ eV.

(a) Determine the populations N_1 and N_2.

(b) Determine the number of photons emitted spontaneously every second.

(c) Determine the attenuation coefficient of this medium at $\lambda_o = 0.5\ \mu m$ assuming that the incident photon flux is small.

(d) Sketch the dependence of the attenuation coefficient on frequency, indicating the important parameters on the sketch.

(e) Find the value of photon-flux density at which the attenuation coefficient decreases by a factor of 2 (i.e., the saturation photon-flux density).

(f) Sketch the dependence of the transmitted photon-flux density $\phi(d)$ on the incident photon-flux density $\phi(0)$ for $\nu = \nu_0$ and $\nu = \nu_0 + \Delta\nu$ when $\phi(0)/\phi_s \ll 1$.

15.4-7 **Gain in a Saturated Amplifying Medium.** Consider a homogeneously broadened laser amplifying medium of length $d = 10$ cm and assume that the saturation photon-flux density $\phi_s = 4 \times 10^{18}$ photons/cm^2-s. It is known that a photon-flux density at the input $\phi(0) = 4 \times 10^{15}$ photons/cm^2-s produces a photon-flux density at the output $\phi(d) = 4 \times 10^{16}$ photons/cm^2-s.

(a) Determine the small-signal gain of the system G_0.

(b) Determine the small-signal gain coefficient γ_0.

(c) What is the photon-flux density at which the gain coefficient decreases by a factor of 5?

(d) Determine the gain coefficient for an input photon-flux density given by $\phi(0) = 4\times10^{19}$ photons/cm^2-s. Under these conditions, is the gain of the system greater than, less than, or the same as the small-signal gain determined in (a)?

*15.5-2 **Ratio of Signal Power to ASE Power.** An unsaturated laser amplifier of length d and gain coefficient $\gamma_0(\nu)$ amplifies an input signal $\phi_S(0)$ of frequency ν and introduces amplified spontaneous emission (ASE) at a rate ξ_{sp} (per unit length). The amplified signal photon-flux density is $\phi_S(d)$ and the ASE at the output is ϕ_{ASE}. Sketch the dependence of the ratio $\phi_S(d)/\phi_{\text{ASE}}$ on the product of the amplifier gain coefficient and length, $\gamma_0(\nu)d$.

*15.5-3 **Photon-Number Distribution for Amplified Coherent Light.** A linearly polarized superposition of interfering pulses of coherent and narrowband thermal light serves as a suitable model for the light emerging from a laser amplifier. The resulting light pulses are known to have random energy fluctuations w that obey the noncentral-chi-square probability density function,[†]

$$p(w) = \frac{1}{\overline{w}_{\text{ASE}}} \exp\left(-\frac{w + w_S}{\overline{w}_{\text{ASE}}}\right) I_0\left[\frac{2\sqrt{w_S w}}{\overline{w}_{\text{ASE}}}\right],$$

provided that the measurement time and detector area are sufficiently small. Here I_0 denotes the modified Bessel function of order zero, $\overline{w}_{\text{ASE}}$ is the mean energy of the ASE, and w_S is the (constant) energy of the amplified coherent signal.

(a) Calculate the mean and variance of w.

(b) Use (13.2-27) and (13.2-28) to determine the photon-number mean $\overline{n}$ and variance σ_n^2, thereby confirming the validity of (15.5-4).

(c) Use (13.2-26) to show that the photon-number distribution is given by

$$p(n) = \frac{\overline{n}_{\text{ASE}}^{\,n}}{(1 + \overline{n}_{\text{ASE}})^{n+1}} \exp\left(-\frac{\overline{n}_S}{1 + \overline{n}_{\text{ASE}}}\right) \mathbb{L}_n\left(-\frac{\overline{n}_S/\overline{n}_{\text{ASE}}}{1 + \overline{n}_{\text{ASE}}}\right),$$

where $\mathbb{L}_n$ represents the Laguerre polynomial (see footnote on page 103)

$$\mathbb{L}_n(-x) = n! \sum_{i=0}^{n} \frac{x^i}{(n-i)!\,[i!]^2},$$

and $\overline{n}_S$ and $\overline{n}_{\text{ASE}}$ are the mean signal and amplified-spontaneous-emission photon numbers, respectively. This is a special case of the noncentral-negative-binomial (NNB) distribution.

(d) Plot $p(n)$ for $\overline{n}_S/\overline{n} = 0$, 0.5, 0.8, and 1, when $\overline{n} = 5$, demonstrating that the NNB photon-number distribution provided above reduces to the Bose–Einstein distribution for $\overline{n}_S/\overline{n} = 0$ and to the Poisson distribution for $\overline{n}_S/\overline{n} = 1$.

[†] See, e.g., T. Li and M. C. Teich, Photon Point Process for Traveling-Wave Laser Amplifiers, *IEEE Journal of Quantum Electronics*, vol. 29, pp. 2568–2578, 1993.

CHAPTER
16

LASERS

16.1 THEORY OF LASER OSCILLATION 659
- A. Optical Amplification and Feedback
- B. Conditions for Laser Oscillation

16.2 CHARACTERISTICS OF THE LASER OUTPUT 666
- A. Power
- B. Spectral Distribution
- C. Spatial Distribution and Polarization
- D. Mode Selection

16.3 TYPES OF LASERS 680
- A. Solid-State Lasers
- B. Fiber Lasers
- C. Raman Lasers
- D. Random Lasers
- E. Gas and Dye Lasers
- F. X-Ray and Free-Electron Lasers
- G. Tabulation of Selected Laser Characteristics

16.4 PULSED LASERS 707
- A. Methods of Pulsing Lasers
- *B. Analysis of Transient Effects
- *C. Q-Switching
- D. Mode Locking
- *E. Optical Frequency Combs

Arthur L. Schawlow
(1921–1999)

Theodore H. Maiman
(1927–2007)

In 1958 Arthur Schawlow and Charles Townes suggested a method for extending the principle of the maser to the optical region of the spectrum. Schawlow shared the 1981 Nobel Prize with Nicolaas Bloembergen (pictured on p. 1015). Theodore Maiman achieved the first successful operation of a (ruby) laser on 16 May 1960, a date commemorated as the *International Day of Light*.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

The laser is an optical oscillator. It comprises a resonant optical amplifier whose output is fed back to the input with matching phase (Fig. 16.0-1). The oscillation process can be initiated by the presence at the amplifier input of even a small amount of noise that contains frequency components lying within the bandwidth of the amplifier. This input is amplified and the output is fed back to the input, where it undergoes further amplification. The process continues indefinitely until a large output is produced. The increase of the signal is ultimately limited by saturation of the amplifier gain, and the system reaches a steady state in which an output signal is created at the frequency of the resonant amplifier.

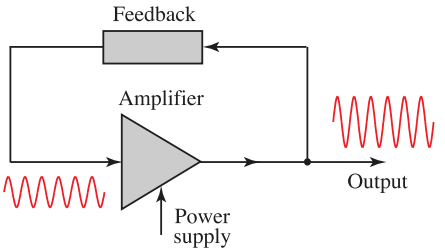


Figure 16.0-1 An oscillator is an amplifier with positive feedback.

Two conditions must be satisfied for oscillation to occur:

- The amplifier gain must be greater than the loss in the feedback system so that net gain is incurred in a round trip through the feedback loop.
- The total phase shift in a single round trip must be a multiple of 2π so that the feedback input phase matches the phase of the original input.

If these conditions are satisfied, the system becomes unstable and oscillation begins. As the power in the oscillator grows, the amplifier gain saturates and thus falls below its initial value. A stable condition is reached when the reduced gain is equal to the loss (Fig. 16.0-2). The gain then just compensates the loss so that the cycle of amplification and feedback is repeated without change and steady-state oscillation prevails.

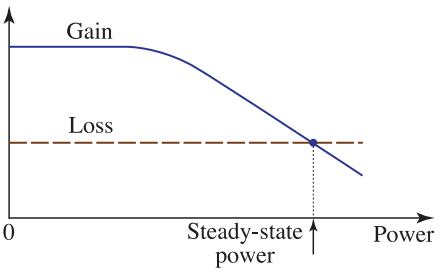


Figure 16.0-2 If the initial amplifier gain is greater than the loss, oscillation may begin. As the oscillator power increases, the amplifier saturates, causing its gain to decrease. A steady-state condition is reached when the gain just equals the loss.

Because the gain and phase shift are functions of frequency, the two oscillation conditions are satisfied only at one or more particular frequencies, namely the resonance frequencies of the oscillator. The useful output is extracted by coupling a portion of the power out of the oscillator.

In summary, an oscillator comprises:

- An amplifier with a gain-saturation mechanism.
- A feedback system.
- A frequency-selection mechanism.
- An output coupling scheme.

The laser is an optical oscillator (Fig. 16.0-3) in which the amplifier is the pumped active medium considered in Secs. 15.1 and 15.2. Gain saturation is a basic property of laser amplifiers, as discussed in Sec. 15.4. Feedback is engendered by placing the active medium in an optical resonator that reflects the light back and forth between its mirrors, as discussed in Chapter 11. Frequency selection is jointly achieved by the resonant amplifier and the resonator, which admits only certain modes. Output coupling is accomplished by making one of the resonator mirrors partially transmitting.

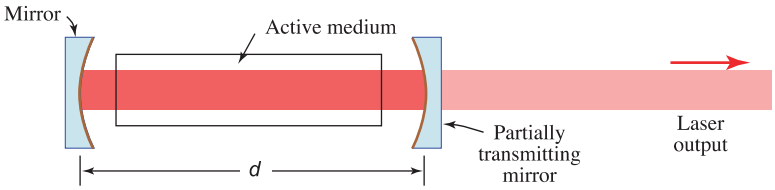


Figure 16.0-3 A laser consists of an optical amplifier (comprising an active medium) placed within an optical resonator. The output is extracted through a partially transmitting mirror.

Lasers have an enormous variety of forms and are used in myriad scientific and technical applications such as interferometry, spectroscopy, imaging, lithography, metrology, communications, lidar, atomic cooling, materials processing, biology, and neuroscience, among others. Needless to say, they are invaluable for fundamental studies and applications in photonics, as well as in all branches of science, engineering, and medicine. The precursor to the laser was the **maser**, an acronym for *M*icrowave *A*mplification by *S*timulated *E*mission of *R*adiation. The maser/laser principle also holds promise for waves other than electromagnetic radiation. The **saser**, for example, is an acoustic version of the laser that emits a beam of phonons, offering *S*ound *A*mplification by *S*timulated *E*mission of *R*adiation.

This Chapter

This chapter provides an introduction to the operation of lasers. In Sec. 16.1 the behavior of the laser amplifier and the laser resonator are summarized, and oscillation conditions are derived. The properties of the light emitted by lasers, including power, spectral distribution, spatial distribution, and polarization, are considered in Sec. 16.2. Various types of lasers are discussed in Sec. 16.3, while Sec. 16.4 is devoted to the operation of pulsed lasers. Laser diodes, microlasers, and nanolasers are considered in Chapter 18 while high-power (petawatt) lasers are briefly examined in Chapter 23.

16.1 THEORY OF LASER OSCILLATION

We begin this section with a summary of the properties of the two basic components of the laser — the resonator and the amplifier. These topics have been discussed in detail in Chapters 11 and 15, respectively, but are reviewed here for convenience.

A. Optical Amplification and Feedback

Laser Amplification

The laser amplifier is a narrowband coherent amplifier of light. Amplification is achieved by stimulated emission from an atomic or molecular system with a transition

whose population is inverted (i.e., the upper energy level is more populated than the lower). The amplifier bandwidth is determined by the linewidth of the atomic transition, or by an inhomogeneous broadening mechanism such as Doppler broadening in gas lasers.

The laser amplifier is a distributed-gain device characterized by its gain coefficient (gain per unit length) $\gamma(\nu)$, which governs the rate at which the photon-flux density ϕ (or the optical intensity $I = h\nu\,\phi$) increases. When the photon-flux density ϕ is small, the **gain coefficient** is given by

$$\gamma_0(\nu) = N_0\,\sigma(\nu) = N_0\,\frac{\lambda^2}{8\pi t_{\rm sp}}\,g(\nu), \tag{16.1-1}$$

Small-Signal Gain Coefficient

where

N_0 = equilibrium population density difference (density of atoms in the upper energy state minus that in the lower state); N_0 increases with increasing pumping rate
$\sigma(\nu)$ = $(\lambda^2/8\pi t_{\rm sp})g(\nu)$ = transition cross section
$t_{\rm sp}$ = effective spontaneous lifetime for stimulated emission
$g(\nu)$ = transition lineshape function
λ = λ_o/n = wavelength in the medium, where n = refractive index

As the photon-flux density increases, the amplifier ultimately enters a region of nonlinear operation in which the gain saturates and decreases. The amplification process then depletes the initial population difference N_0, reducing it to $N = N_0/[1+\phi/\phi_s(\nu)]$ for a homogeneously broadened medium, where

$\phi_s(\nu)$ = $[\tau_s\,\sigma(\nu)]^{-1}$ = saturation photon-flux density
τ_s = saturation time constant, which depends on the decay times of the energy levels involved; in an ideal four-level pumping scheme, $\tau_s \approx t_{\rm sp}$, whereas in an ideal three-level pumping scheme, $\tau_s = 2t_{\rm sp}$

The gain coefficient of the saturated amplifier is therefore reduced to $\gamma(\nu) = N\sigma(\nu)$, so that for homogeneous broadening we have

$$\gamma(\nu) = \frac{\gamma_0(\nu)}{1 + \phi/\phi_s(\nu)}\,. \tag{16.1-2}$$

Saturated Gain Coefficient

The laser amplification process also introduces a phase shift. When the lineshape is Lorentzian with linewidth $\Delta\nu$, so that $g(\nu) = (\Delta\nu/2\pi)/[(\nu-\nu_0)^2 + (\Delta\nu/2)^2]$, the amplifier phase shift per unit length becomes

$$\varphi(\nu) = \frac{\nu - \nu_0}{\Delta\nu}\,\gamma(\nu). \tag{16.1-3}$$

Phase-Shift Coefficient (Lorentzian Lineshape)

This phase shift is above and beyond that introduced by the medium hosting the laser atoms. The gain and phase-shift coefficients for an amplifier with Lorentzian lineshape function are illustrated in Fig. 16.1-1.

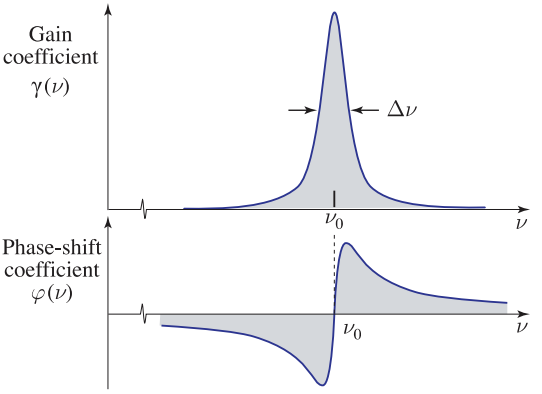


Figure 16.1-1 Spectral dependence of the gain and phase-shift coefficients for an optical amplifier with a Lorentzian lineshape function.

Feedback and Loss: The Optical Resonator

Optical feedback is achieved by placing the active medium in an optical resonator. A Fabry–Perot resonator, comprising two mirrors separated by a distance d, contains the medium (refractive index n) in which the active atoms of the amplifier reside. Travel through the medium introduces a phase shift per unit length equal to the wavenumber

$$k = \frac{2\pi\nu}{c}. \tag{16.1-4}$$

Phase-Shift Coefficient

The resonator also contributes to losses in the system. Absorption and scattering of light in the medium introduces a distributed loss characterized by the attenuation coefficient α_s (loss per unit length). In traveling a round trip through a resonator of length d, the photon-flux density is reduced by the factor $\mathcal{R}_1\mathcal{R}_2 \exp(-2\alpha_s d)$, where $\mathcal{R}_1$ and $\mathcal{R}_2$ are the reflectances of the two mirrors. The overall loss in one round trip can therefore be described by a total effective **distributed loss coefficient** α_r, where

$$\exp(-2\alpha_r d) = \mathcal{R}_1\mathcal{R}_2 \exp(-2\alpha_s d), \tag{16.1-5}$$

so that

$$\begin{aligned} \alpha_r &= \alpha_s + \alpha_{m1} + \alpha_{m2} \\ \alpha_{m1} &= \frac{1}{2d} \ln\frac{1}{\mathcal{R}_1} \\ \alpha_{m2} &= \frac{1}{2d} \ln\frac{1}{\mathcal{R}_2}, \end{aligned} \tag{16.1-6}$$

Loss Coefficient

where α_{m1} and α_{m2} represent the loss contributions of mirrors 1 and 2, respectively. The contribution from both mirrors is therefore

$$\alpha_m = \alpha_{m1} + \alpha_{m2} = \frac{1}{2d} \ln\frac{1}{\mathcal{R}_1\mathcal{R}_2}. \tag{16.1-7}$$

Since α_r represents the total loss of energy (or number of photons) per unit length, $\alpha_r c$ represents the loss of photons per second. Thus,

$$\tau_p = \frac{1}{\alpha_r c} \tag{16.1-8}$$

represents the **photon lifetime**, which decreases with increasing loss.

The resonator sustains only frequencies that correspond to a round-trip phase shift that is a multiple of 2π. For a resonator devoid of active atoms (a so-called "cold resonator"), the round-trip phase shift is simply $k2d = 2\pi\nu d/c = q2\pi$, corresponding to modes of frequencies

$$\nu_q = q\nu_F, \qquad q = 1, 2, \ldots, \tag{16.1-9}$$

where $\nu_F = c/2d$ is the resonator mode spacing and $c = c_o/n$ is the speed of light in the medium (Fig. 16.1-2). The (full width at half maximum) spectral width of these resonator modes is

$$\delta\nu \approx \frac{\nu_F}{\mathcal{F}}, \tag{16.1-10}$$

where $\mathcal{F}$ is the finesse of the resonator (Sec. 11.1A). When the resonator losses are small and the finesse is large,

$$\mathcal{F} \approx \frac{\pi}{\alpha_r d} = 2\pi\tau_p\,\nu_F. \tag{16.1-11}$$

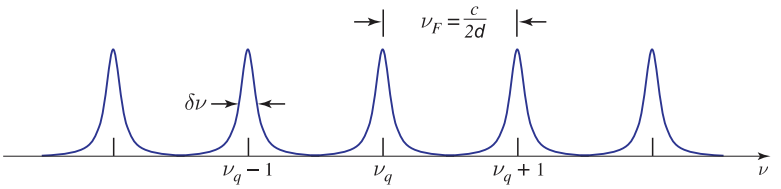


Figure 16.1-2 Resonator modes are separated by the frequency $\nu_F = c/2d$ and have linewidths $\delta\nu = \nu_F/\mathcal{F} = 1/2\pi\tau_p$.

B. Conditions for Laser Oscillation

Two conditions must be satisfied for the laser to oscillate (lase). The *gain condition* determines the minimum population difference, and therefore the pumping threshold, required for lasing. The *phase condition* determines the frequency (or frequencies) at which oscillation takes place.

Gain Condition: Laser Threshold

The initiation of laser oscillation requires that the small-signal gain coefficient be greater than the loss coefficient,

$$\boxed{\gamma_0(\nu) > \alpha_r\,,} \tag{16.1-12}$$

Threshold Gain Condition

or, equivalently, that the gain be greater than the loss. In accordance with (16.1-1), the small-signal gain coefficient $\gamma_0(\nu)$ is proportional to the equilibrium population

density difference N_0, which in turn is known from Chapter 15 to increase with the pumping rate R. Indeed, (16.1-1) may be used to translate (16.1-12) into a condition on the population difference, i.e., $N_0 = \gamma_0(\nu)/\sigma(\nu) > \alpha_r/\sigma(\nu)$. Thus,

$$N_0 > N_\mathrm{t}\,, \tag{16.1-13}$$

where the quantity

$$N_\mathrm{t} = \frac{\alpha_r}{\sigma(\nu)} \tag{16.1-14}$$

is called the **threshold population difference**. This quantity, which is proportional to α_r, determines the minimum pumping rate R_t for the initiation of laser oscillation.[†]

Using (16.1-8), α_r may alternatively be written in terms of the photon lifetime, $\alpha_r = 1/c\tau_p$, whereupon (16.1-14) takes the form

$$N_\mathrm{t} = \frac{1}{c\tau_p\,\sigma(\nu)}\,. \tag{16.1-15}$$

The threshold population density difference is therefore directly proportional to α_r and inversely proportional to τ_p. Higher loss (shorter photon lifetime) requires more vigorous pumping to achieve lasing.

Finally, use of the standard formula for the transition cross section, $\sigma(\nu) = (\lambda^2/8\pi t_\mathrm{sp})g(\nu)$, leads to yet another expression for the threshold population difference,

$$\boxed{N_\mathrm{t} = \frac{8\pi}{\lambda^2 c}\frac{t_\mathrm{sp}}{\tau_p}\frac{1}{g(\nu)}\,,} \tag{16.1-16}$$

Threshold Population Difference

from which it is clear that N_t is a function of the frequency ν. The threshold is lowest, and lasing is therefore most readily achieved, at the frequency where the lineshape function is greatest, i.e., at its central frequency $\nu = \nu_0$. For a Lorentzian lineshape function, (14.3-35) provides that $g(\nu_0) = 2/\pi\Delta\nu$, so that the minimum population difference for oscillation at the central frequency ν_0 turns out to be

$$N_\mathrm{t} = \frac{2\pi}{\lambda^2 c}\frac{2\pi\Delta\nu\, t_\mathrm{sp}}{\tau_p}\,. \tag{16.1-17}$$

It is directly proportional to the linewidth $\Delta\nu$.

If, furthermore, the transition is limited by lifetime broadening with a decay time given by t_sp, $\Delta\nu$ assumes the value $1/2\pi t_\mathrm{sp}$ (Sec. 14.3D), whereupon (16.1-17) simplifies to

$$N_\mathrm{t} = \frac{2\pi}{\lambda^2 c\tau_p} = \frac{2\pi\alpha_r}{\lambda^2}\,. \tag{16.1-18}$$

This formula reveals that the minimum threshold population difference required to achieve oscillation is a simple function of the wavelength λ and the photon lifetime τ_p. It is clear that laser oscillation becomes more difficult to achieve as the wavelength decreases. As a numerical example, if $\lambda_o = 1\ \mu\mathrm{m}$, $\tau_p = 1$ ns, and the refractive index $n = 1$, we obtain $N_\mathrm{t} \approx 2.1 \times 10^7\ \mathrm{cm}^{-3}$.

[†] It is possible, however, to achieve **lasing without inversion** (**LWI**) within the energy-level structure of a conventional laser medium by making use of an external optical field that creates an additional path from the lower to the upper energy level via an auxiliary energy level. Under appropriate circumstances, the presence of the two paths can result in destructive quantum interference that reduces or eliminates absorption.

EXERCISE 16.1-1

Threshold of a Ruby Laser.

(a) At the line center of the $\lambda_o = 694.3$-nm transition (see Table 15.3-1), the absorption coefficient of ruby in thermal equilibrium (i.e., without pumping) at $T = 300°$ K is $\alpha(\nu_0) \equiv -\gamma_0(\nu_0) \approx 0.2\ \text{cm}^{-1}$. If the concentration of Cr^{3+} ions responsible for the transition is $N_a = 1.58 \times 10^{19}\ \text{cm}^{-3}$, determine the transition cross section $\sigma_0 = \sigma(\nu_0)$.

(b) A ruby laser makes use of a 10-cm-long ruby rod (refractive index $n = 1.76$) of cross-sectional area $1\ \text{cm}^2$ and operates on this transition at $\lambda_o = 694.3\ \text{nm}$. Both of its ends are polished and coated so that each has a reflectance of 80%. Assuming that there are no scattering or other extraneous losses, determine the resonator loss coefficient α_r and the resonator photon lifetime τ_p.

(c) As the laser pumping increases, $\gamma(\nu_0)$ increases from its initial thermal equilibrium value of $-0.2\ \text{cm}^{-1}$ and changes sign, thereby providing gain. Determine the threshold population difference N_t for laser oscillation.

Phase Condition: Laser Frequencies

The second condition of oscillation requires that the phase shift imparted to a light wave completing a round trip within the resonator must be a multiple of 2π, i.e.,

$$2kd + 2\varphi(\nu)d = 2\pi q, \qquad q = 1, 2, \ldots . \tag{16.1-19}$$

If the contribution arising from the active laser atoms $[2\varphi(\nu)d]$ is small, then dividing (16.1-19) by $2d$ reduces to the cold-resonator result obtained earlier, $\nu = \nu_q = q(c/2d)$. In the presence of the active medium, when $2\varphi(\nu)d$ does contribute, the solution of (16.1-19) gives rise to a set of oscillation frequencies ν_q' that are slightly displaced from the cold-resonator frequencies ν_q. It turns out that the cold-resonator modal frequencies are all pulled slightly toward the central frequency of the atomic transition, as shown below.

*Frequency Pulling

Using the relation $k = 2\pi\nu/c$, and the phase-shift coefficient for the Lorentzian lineshape function provided in (16.1-3), the phase-shift condition (16.1-19) provides

$$\nu + \frac{c}{2\pi}\frac{\nu - \nu_0}{\Delta\nu}\gamma(\nu) = \nu_q\,. \tag{16.1-20}$$

This equation can be solved for the oscillation frequency $\nu = \nu_q'$ that corresponds to each cold-resonator mode ν_q. Because the equation is nonlinear, a graphical solution is useful. The left-hand side of (16.1-20), designated $\psi(\nu)$, is plotted in Fig. 16.1-3 as the sum of a straight line, representing ν, plus the Lorentzian phase-shift coefficient shown schematically in Fig. 16.1-1. The value of $\nu = \nu_q'$ that renders $\psi(\nu) = \nu_q$ is graphically determined. It is apparent from the figure that the cold-resonator modes ν_q are always frequency-pulled toward the central frequency of the resonant medium ν_0.

An approximate analytical solution of (16.1-20) can also be obtained. We write (16.1-20) in the form

$$\nu = \nu_q - \frac{c}{2\pi}\frac{\nu - \nu_0}{\Delta\nu}\gamma(\nu)\,. \tag{16.1-21}$$

When $\nu = \nu_q' \approx \nu_q$, the second term on the right-hand side of (16.1-21) is small so that ν may be replaced with ν_q without much loss of accuracy. This leads to

$$\nu_q' = \nu_q - \frac{c}{2\pi}\frac{\nu_q - \nu_0}{\Delta\nu}\gamma(\nu_q)\,, \tag{16.1-22}$$

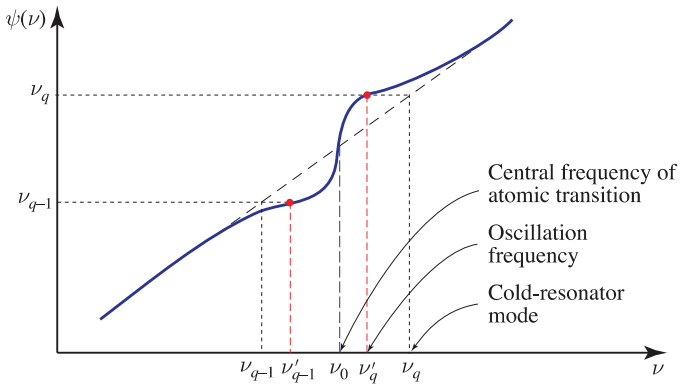


Figure 16.1-3 The left-hand side of (16.1-20), $\psi(\nu)$, plotted as a function of ν. The frequency ν for which $\psi(\nu) = \nu_q$ is the solution of (16.1-20). Each "cold" resonator frequency ν_q corresponds to a "hot" resonator frequency ν'_q, which is shifted in the direction of the atomic-resonance central frequency ν_0.

which is an explicit expression for the oscillation frequency ν'_q as a function of the cold-resonator frequency ν_q. Furthermore, under steady-state conditions, the gain equals the loss so that $\gamma(\nu_q) = \alpha_r \approx \pi/\mathcal{F}d = (2\pi/c)\delta\nu$, where $\delta\nu$ is the spectral width of the cold resonator modes. Substituting this relation into (16.1-22) then leads to

$$\nu'_q \approx \nu_q - (\nu_q - \nu_0)\frac{\delta\nu}{\Delta\nu}. \tag{16.1-23}$$

Laser Frequencies

The cold-resonator frequency ν_q is thus seen to be pulled toward the atomic resonance frequency ν_0 by a fraction $\delta\nu/\Delta\nu$ of its original distance from the central frequency $(\nu_q - \nu_0)$, as illustrated in Fig. 16.1-4. The sharper the resonator mode (the smaller the value of $\delta\nu$), the less significant is the pulling effect. In contrast, the narrower the atomic resonance linewidth (the smaller the value of $\Delta\nu$), the more effective is the pulling.

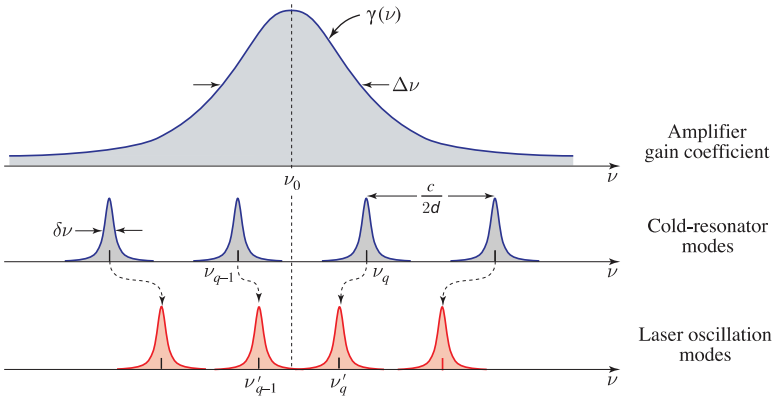


Figure 16.1-4 The laser oscillation modes fall near the cold-resonator modes; they are pulled slightly toward the atomic-resonance central frequency ν_0. The illustration is not to scale and the degree of pulling is exaggerated for clarity.

16.2 CHARACTERISTICS OF THE LASER OUTPUT

A. Power

Internal Photon-Flux Density

A laser pumped above the threshold ($N_0 > N_t$) exhibits a small-signal gain coefficient $\gamma_0(\nu)$ that is greater than the loss coefficient α_r, as indicated in (16.1-12). Laser oscillation may then begin, provided that the phase condition (16.1-19) is satisfied. As the photon-flux density ϕ inside the resonator increases (Fig. 16.2-1), the gain coefficient $\gamma(\nu)$ begins to decrease in accordance with (16.1-2) for a homogeneously broadened medium. As long as the gain coefficient remains larger than the loss coefficient, the photon flux continues to grow.

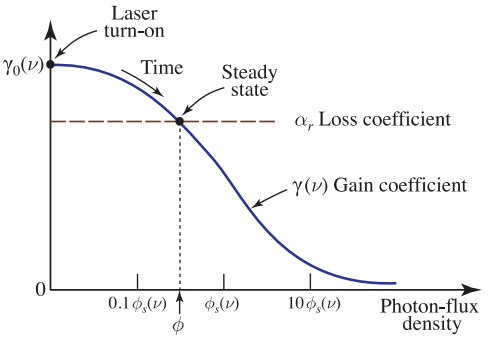


Figure 16.2-1 Determination of the steady-state laser photon-flux density ϕ. At the time of laser turn-on, $\phi = 0$ so that $\gamma(\nu) = \gamma_0(\nu)$. As the oscillation builds up in time, the increase in ϕ causes $\gamma(\nu)$ to decrease through gain saturation. When γ reaches α_r, the photon-flux density ceases its growth and steady-state conditions are achieved. The smaller the loss, the greater the value of ϕ.

Finally, when the saturated gain coefficient becomes equal to the loss coefficient (or equivalently $N = N_t$), the photon flux ceases its growth and the oscillation reaches steady-state conditions. The result is **gain clamping** at the value of the loss. The steady-state laser internal photon-flux density is therefore determined by equating the large-signal (saturated) gain coefficient to the loss coefficient, i.e., $\gamma_0(\nu)/[1 + \phi/\phi_s(\nu)] = \alpha_r$, which provides

$$\phi = \begin{cases} \phi_s(\nu)\left(\dfrac{\gamma_0(\nu)}{\alpha_r} - 1\right), & \gamma_0(\nu) > \alpha_r \\ 0, & \gamma_0(\nu) \le \alpha_r. \end{cases} \tag{16.2-1}$$

Equation (16.2-1) represents the steady-state photon-flux density arising from laser action. This is the mean number of photons per second crossing a unit area in both directions, since photons traveling in both directions contribute to the saturation process. The photon-flux density for photons traveling in a single direction is therefore $\phi/2$. Spontaneous emission has been neglected in this simplified treatment. Of course, (16.2-1) represents the mean photon-flux density; there are random fluctuations about this mean as discussed in Sec. 13.2.

Since $\gamma_0(\nu) = N_0\,\sigma(\nu)$ and $\alpha_r = N_t\,\sigma(\nu)$, (16.2-1) may be written in the form

$$\phi = \begin{cases} \phi_s(\nu)\left(\dfrac{N_0}{N_t} - 1\right), & N_0 > N_t \\ 0, & N_0 \le N_t. \end{cases} \tag{16.2-2}$$

Steady-State Internal Photon-Flux Density

Below threshold, the laser photon-flux density is zero; any increase in the pumping rate is manifested as an increase in the spontaneous-emission photon flux, but there is no sustained oscillation. Above threshold, the steady-state internal laser photon-flux density is directly proportional to the initial population difference N_0, and therefore increases with the pumping rate R [see (15.2-13) and (15.2-27)]. If N_0 is twice the threshold value N_t, the photon-flux density is precisely equal to the saturation value $\phi_s(\nu)$, which is the photon-flux density at which the gain coefficient decreases to half its maximum value. Both the population difference N and the photon-flux density ϕ are shown as functions N_0 in Fig. 16.2-2.

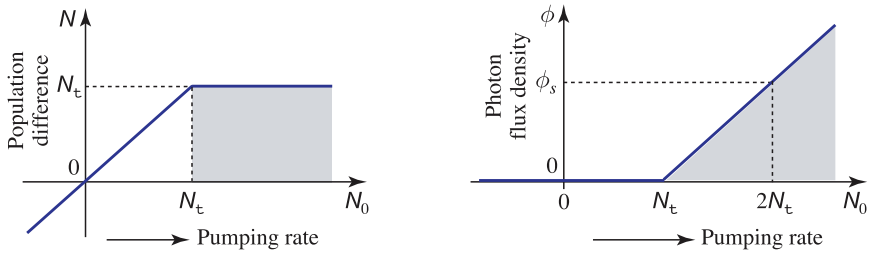


Figure 16.2-2 Steady-state values of the population difference N, and the laser internal photon-flux density ϕ, as functions of N_0 (the population difference in the absence of radiation; N_0 increases with the pumping rate R). Laser oscillation occurs when N_0 exceeds N_t; the steady-state value of N then saturates, clamping at the value N_t [just as $\gamma_0(\nu)$ is clamped at α_r]. Above threshold, ϕ is proportional to $N_0 - N_t$.

Output Photon-Flux Density

Only a portion of the steady-state internal photon-flux density, as determined by (16.2-2), leaves the resonator in the form of useful light. The output photon-flux density ϕ_o is that part of the internal photon-flux density that propagates toward, and is transmitted by, mirror 1, i.e., $\phi/2$. If the transmittance of this mirror is $\mathcal{T}$, the output photon-flux density is then

$$\phi_o = \tfrac{1}{2}\mathcal{T}\phi. \tag{16.2-3}$$

The corresponding optical intensity of the laser output I_o is

$$I_o = \tfrac{1}{2}h\nu\,\mathcal{T}\phi, \tag{16.2-4}$$

and the laser output power is $P_o = I_o A$, where A is the cross-sectional area of the laser beam. These equations, together with (16.2-2), permit the output power of the laser to be explicitly calculated in terms of $\phi_s(\nu)$, N_0, N_t, $\mathcal{T}$, and A.

Optimization of the Output Photon-Flux Density

The useful photon-flux density at the laser output diminishes the internal photon-flux density and therefore contributes to the losses of the laser oscillator. Any attempt to increase the fraction of photons allowed to escape from the resonator (in the expectation of increasing the useful light output) results in increased losses so that the steady-state photon-flux density inside the resonator decreases. The net result may therefore be a decrease, rather than an increase, in the useful light output.

We proceed to demonstrate that there is an optical transmittance $\mathcal{T}$ $(0 < \mathcal{T} < 1)$ that maximizes the laser output intensity. The output photon-flux density $\phi_o = \mathcal{T}\phi/2$

is a product of the mirror's transmittance $\mathcal{T}$ and the internal photon-flux density $\phi/2$. As $\mathcal{T}$ is increased, ϕ decreases as a result of the greater losses. At one extreme, when $\mathcal{T} = 0$, the oscillator has the least loss (ϕ is maximum), but there is no laser output whatever ($\phi_o = 0$). At the other extreme, when the mirror is removed so that $\mathcal{T} = 1$, the increased losses make $\alpha_r > \gamma_0(\nu)$ ($N_t > N_0$), thereby preventing laser oscillation. In this case $\phi = 0$, so that again $\phi_o = 0$. The optimal value of $\mathcal{T}$ lies somewhere between these two extremes.

To determine this value, we must obtain an explicit relation between ϕ_o and $\mathcal{T}$. Let us assume that mirror 1, with reflectance $\mathcal{R}_1$ and transmittance $\mathcal{T} = 1 - \mathcal{R}_1$, transmits the useful light. The loss coefficient α_r is written as a function of $\mathcal{T}$ by substituting in (16.1-6) the loss coefficient associated with mirror 1,

$$\alpha_{m1} = \frac{1}{2d} \ln\frac{1}{\mathcal{R}_1} = -\frac{1}{2d} \ln(1 - \mathcal{T}), \tag{16.2-5}$$

which leads to

$$\alpha_r = \alpha_s + \alpha_{m2} - \frac{1}{2d} \ln(1 - \mathcal{T}), \tag{16.2-6}$$

where the loss coefficient associated with mirror 2 is

$$\alpha_{m2} = \frac{1}{2d} \ln\frac{1}{\mathcal{R}_2}. \tag{16.2-7}$$

We now use (16.2-1), (16.2-3), and (16.2-6) to obtain an equation for the transmitted photon-flux density ϕ_o as a function of the mirror transmittance,

$$\phi_o = \tfrac{1}{2}\phi_s \mathcal{T} \left[\frac{g_0}{L - \ln(1 - \mathcal{T})} - 1\right], \qquad g_0 = 2\gamma_0(\nu)d, \quad L = 2(\alpha_s + \alpha_{m2})d, \tag{16.2-8}$$

which is plotted in Fig. 16.2-3. Note that the transmitted photon-flux density is directly related to the small-signal gain coefficient. The optical transmittance $\mathcal{T}_{\text{op}}$ is determined by setting the derivative of ϕ_o with respect to $\mathcal{T}$ equal to zero. When $\mathcal{T} \ll 1$ we can make use of the approximation $\ln(1 - \mathcal{T}) \approx -\mathcal{T}$ to obtain the expression

$$\mathcal{T}_{\text{op}} \approx \sqrt{g_0 L} - L. \tag{16.2-9}$$

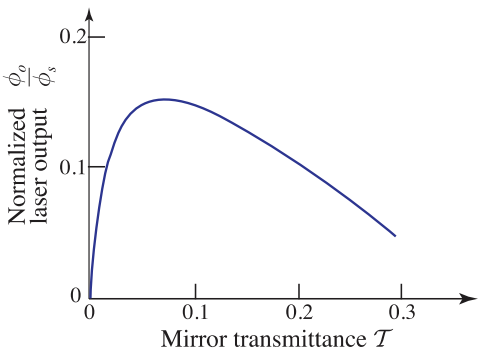


Figure 16.2-3 Dependence of the transmitted steady-state photon-flux density ϕ_o on the mirror transmittance $\mathcal{T}$. For the purposes of this illustration, the gain factor $g_0 = 2\gamma_0 d$ has been chosen to be 0.5 and the loss factor $L = 2(\alpha_s + \alpha_{m2})d$ is 0.02 (2%). The optical transmittance $\mathcal{T}_{\text{op}}$ turns out to be 0.08.

Internal Photon-Number Density

The steady-state number of photons per unit volume inside the resonator $\mathfrak{n}$ is related to the steady-state internal photon-flux density ϕ (for photons traveling in both directions) by the simple relation

$$\boxed{\mathfrak{n} = \frac{\phi}{c}\,.} \tag{16.2-10}$$

This is readily visualized by considering a cylinder of area A, length c, and volume cA (c is the velocity of light in the medium), whose axis lies parallel to the axis of the resonator. For a resonator containing $\mathfrak{n}$ photons per unit volume, the cylinder contains $cA\mathfrak{n}$ photons. These photons travel in both directions, parallel to the axis of the resonator, half of them crossing the base of the cylinder in each second. Since the base of the cylinder also receives an equal number of photons from the other side, however, the photon-flux density (photons per second per unit area in both directions) is $\phi = 2(\frac{1}{2}cA\mathfrak{n})/A = c\mathfrak{n}$, from which (16.2-10) follows.

The photon-number density corresponding to the steady-state internal photon-flux density in (16.2-2) is

$$\boxed{\mathfrak{n} = \mathfrak{n}_s\left(\frac{N_0}{N_\mathrm{t}} - 1\right), \qquad N_0 > N_\mathrm{t}\,,} \tag{16.2-11}$$

Steady-State Photon-Number Density

where $\mathfrak{n}_s = \phi_s(\nu)/c$ is the photon-number density saturation value. Using the relations $\phi_s(\nu) = [\tau_s\,\sigma(\nu)]^{-1}$, $\alpha_r = \gamma(\nu)$, $\alpha_r = 1/c\tau_p$, and $\gamma(\nu) = N\sigma(\nu) = N_\mathrm{t}\,\sigma(\nu)$, (16.2-11) may be written in the form

$$\boxed{\mathfrak{n} = (N_0 - N_\mathrm{t})\frac{\tau_p}{\tau_s}\,, \qquad N_0 > N_\mathrm{t}\,.} \tag{16.2-12}$$

Steady-State Photon-Number Density

This relation admits a simple and direction interpretation: $(N_0 - N_\mathrm{t})$ is the population difference (per unit volume) in excess of threshold, and $(N_0 - N_\mathrm{t})/\tau_s$ represents the rate at which photons are generated which, by virtue of steady-state operation, is equal to the rate at which photons are lost, $\mathfrak{n}/\tau_p$. The fraction τ_p/τ_s is the ratio of the rate at which photons are emitted to the rate at which they are lost.

Under ideal pumping conditions in a four-level laser system, (15.2-13) and (15.2-14) provide that $\tau_s \approx t_\mathrm{sp}$ and $N_0 \approx Rt_\mathrm{sp}$, where R is the rate (s^{-1}-cm^{-3}) at which atoms are pumped. Equation (16.2-12) can thus be written as

$$\frac{\mathfrak{n}}{\tau_p} = R - R_\mathrm{t}\,, \qquad R > R_\mathrm{t}\,, \tag{16.2-13}$$

where $R_\mathrm{t} = N_\mathrm{t}/t_\mathrm{sp}$ is the threshold value of the pumping rate. Under steady-state conditions, therefore, the overall photon-density loss rate $\mathfrak{n}/\tau_p$ is precisely equal to the excess pumping rate $R - R_\mathrm{t}$.

Output Photon Flux and Efficiency

If transmission through the laser output mirror is the only source of resonator loss (which is accounted for in τ_p), and V is the volume of the active medium, (16.2-13)

provides that the total output photon flux Φ_o (photons per second) is

$$\Phi_o = (R - R_t)V, \qquad R > R_t. \tag{16.2-14}$$

If there are loss mechanisms other than through the output laser mirror, the output photon flux can be written as

$$\boxed{\Phi_o = \eta_e (R - R_t)V,} \tag{16.2-15}$$
Laser Output Photon Flux

where the extraction efficiency η_e is the ratio of the loss arising from the extracted useful light to all of the total losses in the resonator α_r.

If the useful light exits only through mirror 1, (16.1-8) and (16.2-5) for α_r and α_{m1} may be used to write η_e as

$$\eta_e = \frac{\alpha_{m1}}{\alpha_r} = \frac{c}{2d}\,\tau_p \ln\frac{1}{\mathcal{R}_1}. \tag{16.2-16}$$

If, furthermore, $\mathcal{T} = 1 - \mathcal{R}_1 \ll 1$, (16.2-16) provides

$$\boxed{\eta_e \approx \frac{\tau_p}{T_F}\,\mathcal{T},} \tag{16.2-17}$$
Extraction Efficiency

where we have defined $1/T_F = c/2d$, indicating that the extraction efficiency η_e can be understood in terms of the ratio of the photon lifetime to its round-trip travel time, multiplied by the mirror transmittance. The output laser power is then

$$P_o = h\nu\,\Phi_o = \eta_e h\nu (R - R_t)V. \tag{16.2-18}$$

With the help of a few algebraic manipulations it can be confirmed that this expression accords with that obtained from (16.2-4).

Losses result from other sources as well, such as inefficiency in the pumping process. Overhead functions, such as cooling and monitoring, also consume power. The **power-conversion efficiency** η_c (also called the **overall efficiency** or the **wall-plug efficiency**) is generally defined as the ratio of the output optical power P_o to the input electrical power P_e:

$$\boxed{\eta_c = \frac{P_o}{P_e}.} \tag{16.2-19}$$
Power-Conversion Efficiency

Representative values of η_c for various types of lasers are provided in Table 16.3-1. Because the laser output power increases linearly with pump power above threshold, in accordance with (16.2-18), the **differential power-conversion efficiency** (also called the **slope efficiency**) is another oft-encountered measure of performance:

$$\boxed{\eta_s = \frac{dP_o}{dP_e}.} \tag{16.2-20}$$
Slope Efficiency

The slope efficiency η_s is typically larger than the power-conversion efficiency η_c.

Optically pumped lasers are often characterized by efficiencies analogous to those provided in (16.2-19) and (16.2-20), with the optical pump power P_p replacing the electrical power P_e. The **optical-to-optical efficiency** is thus defined as

$$\eta_\mathrm{o} = \frac{P_o}{P_\mathrm{p}}, \tag{16.2-21}$$

while the **optical-to-optical slope efficiency** is expressed as

$$\eta_\mathrm{s} = \frac{dP_o}{dP_\mathrm{p}}. \tag{16.2-22}$$

The optical-to-optical slope efficiency η_s is typically larger than the optical-to-optical efficiency η_o, which is itself bounded by $\eta_\mathrm{o} \leq 1 - q$, where q is the quantum defect set forth in (15.2-32).

B. Spectral Distribution

The spectral distribution of the generated laser light is determined both by the atomic lineshape of the active medium (including whether it is homogeneous or inhomogeneously broadened) and by the resonator modes. This is illustrated in terms of the two conditions for laser oscillations:

1. The gain condition requiring that the initial gain coefficient of the amplifier be greater than the loss coefficient $[\gamma_0(\nu) > \alpha_r]$ is satisfied for all oscillation frequencies lying within a continuous spectral band of width B centered about the atomic resonance frequency ν_0, as illustrated in Fig. 16.2-4(a). The bandwidth B increases with the atomic linewidth $\Delta\nu$ and the ratio $\gamma_0(\nu_0)/\alpha_r$; the precise relation depends on the shape of the function $\gamma_0(\nu)$.
2. The phase condition requires that the oscillation frequency be one of the resonator modal frequencies ν_q (assuming, for simplicity, that frequency pulling is negligible). The FWHM linewidth of each mode is $\delta\nu \approx \nu_F/\mathcal{F}$ [Fig. 16.2-4(b)].

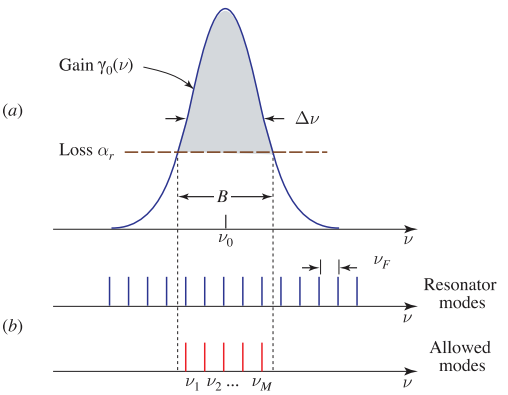


Figure 16.2-4 (a) Laser oscillation can occur only at frequencies for which the gain coefficient is greater than the loss coefficient (shaded region). (b) Oscillation can occur only within a band $\delta\nu$ of the resonator modal frequencies (which are represented as lines for simplicity of illustration).

It follows that only a finite number of oscillation frequencies $(\nu_1, \nu_2, \ldots, \nu_M)$ are

possible. The number of possible laser oscillation modes is therefore

$$M \approx \frac{B}{\nu_F}, \tag{16.2-23}$$

Number of Possible Laser Modes

where $\nu_F = c/2d$ is the approximate spacing between adjacent modes. However, of these M possible modes, the number of modes that actually carry optical power depends on the nature of the atomic line broadening mechanism. It will be shown below that for an inhomogeneously broadened medium all M modes oscillate (albeit at different powers), whereas for a homogeneously broadened medium these modes engage in some degree of competition, making it more difficult for as many modes to oscillate simultaneously. Multimode lasers give rise to intensity-noise fluctuations that arise from modal beating and modal competition.

EXERCISE 16.2-1

Number of Modes in a Gas Laser. A Doppler-broadened gas laser has a gain coefficient with a Gaussian spectral profile (see Sec. 14.3D and Exercise 14.3-2) that can be written as $\gamma_0(\nu) = \gamma_0(\nu_0)\exp[-(\nu-\nu_0)^2/2\sigma_D^2]$, where $\Delta\nu_D = \sqrt{8\ln 2}\,\sigma_D$ is the FWHM linewidth.

(a) Derive an expression for the allowed oscillation band B as a function of $\Delta\nu_D$ and the ratio $\gamma_0(\nu_0)/\alpha_r$, where α_r is the resonator loss coefficient.

(b) A He–Ne laser has a Doppler linewidth $\Delta\nu_D = 1.5$ GHz and a midband gain coefficient $\gamma_0(\nu_0) = 2 \times 10^{-3}\ \text{cm}^{-1}$. The length of the laser resonator is $d = 100$ cm, and the reflectances of the mirrors are 100% and 97% (all other resonator losses are negligible). Assuming that the refractive index $n = 1$, determine the number of laser modes M.

Homogeneously Broadened Medium

Immediately after being turned on, all laser modes for which the initial gain is greater than the loss begin to grow, as portrayed in Fig. 16.2-5(a), and photon-flux densities $\phi_1, \phi_2, \ldots, \phi_M$ are created in the M modes.

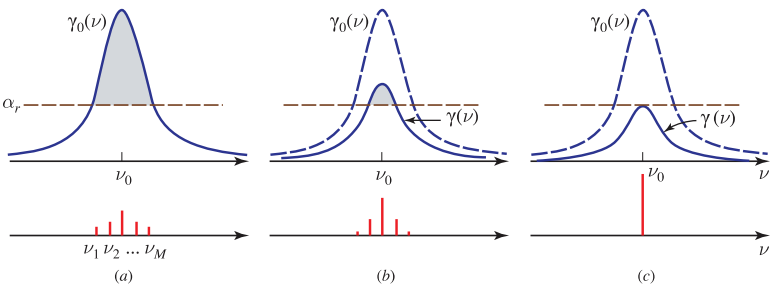


Figure 16.2-5 Growth of oscillation in an ideal homogeneously broadened medium. (a) Immediately following laser turn-on, all modal frequencies $\nu_1, \nu_2, \ldots, \nu_M$, for which the gain coefficient exceeds the loss coefficient, begin to grow, with the central modes growing at the highest rate. (b) After a short time the gain saturates so that the central modes continue to grow while the peripheral modes, for which the loss has become greater than the gain, are attenuated and eventually vanish. (c) In the absence of spatial hole burning, only a single mode survives.

Modes whose frequencies lie closest to the transition central frequency ν_0 grow most quickly and acquire the highest photon-flux densities. These photons interact with the medium and reduce the gain by depleting the population difference. The saturated gain coefficient may be written as

$$\gamma(\nu) = \frac{\gamma_0(\nu)}{1 + \sum_{j=1}^{M} \phi_j / \phi_s(\nu_j)}, \tag{16.2-24}$$

where $\phi_s(\nu_j)$ is the saturation photon-flux density associated with mode j. The validity of (16.2-24) may be verified by carrying out an analysis similar to that which led to (15.4-3). The saturated gain is displayed in Fig. 16.2-5(b).

Because the gain coefficient for a homogeneously broadened medium is reduced uniformly, for modes sufficiently distant from the line center the loss becomes greater than the gain; these modes lose power while the more central modes continue to grow, albeit at a slower rate. Ultimately, only a single surviving mode (or two modes in the symmetrical case) maintains a gain equal to the loss, with the loss exceeding the gain for all other modes. Under ideal steady-state conditions, the power in this preferred mode remains stable, while laser oscillation on all other modes vanishes [Fig. 16.2-5(c)]. The surviving mode has the frequency that lies closest to ν_0; values of the gain for its competitors lie below the loss line. Given the frequency of the surviving mode, its photon-flux density may be determined by means of (16.2-2).

In practice, however, homogeneously broadened lasers do oscillate on multiple modes because the different modes occupy different spatial portions of the active medium. When oscillation on the most central mode in Fig. 16.2-5 is established, the gain coefficient can still exceed the loss coefficient at locations where the standing-wave electric field of the most central mode vanishes. This phenomenon is called **spatial hole burning**. It allows another mode, whose peak fields are located near the energy nulls of the central mode, an opportunity to lase as well.

Inhomogeneously Broadened Medium

In an inhomogeneously broadened medium, the gain $\overline{\gamma}_0(\nu)$ represents the composite envelope of the gains of different species of atoms (Sec. 14.3D), as illustrated in Fig. 16.2-6.

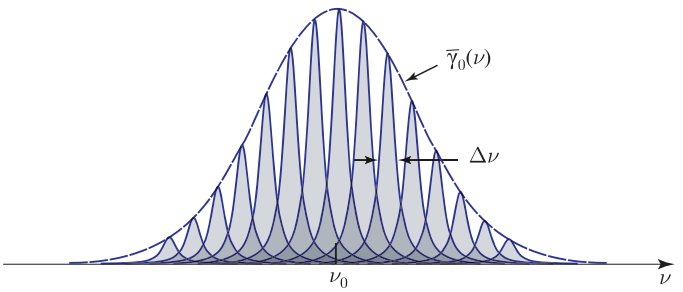


Figure 16.2-6 The lineshape function of an inhomogeneously broadened medium is a composite of numerous constituent atomic lineshape functions associated with different properties or different environments.

The situation immediately after laser turn-on is the same as in the homogeneously broadened medium. Modes for which the gain is larger than the loss begin to grow and the gain decreases. If the spacing between the modes is larger than the width $\Delta\nu$ of the constituent atomic lineshape functions, different modes interact with different

atoms. Atoms whose lineshapes fail to coincide with any of the modes are ignorant of the presence of photons in the resonator. Their population difference is therefore not affected and the gain they provide remains the small-signal (unsaturated) domain. Atomic species whose frequencies coincide with modes deplete their inverted populations and their gains saturate, creating "holes" in the gain spectral profile [Fig. 16.2-7(a)]. This process is known as **spectral hole burning**. The width of a spectral hole increases with the photon-flux density in accordance with the square-root law associated with (15.4-16): $\Delta\nu_s = \Delta\nu(1 + \phi/\phi_s)^{1/2}$.

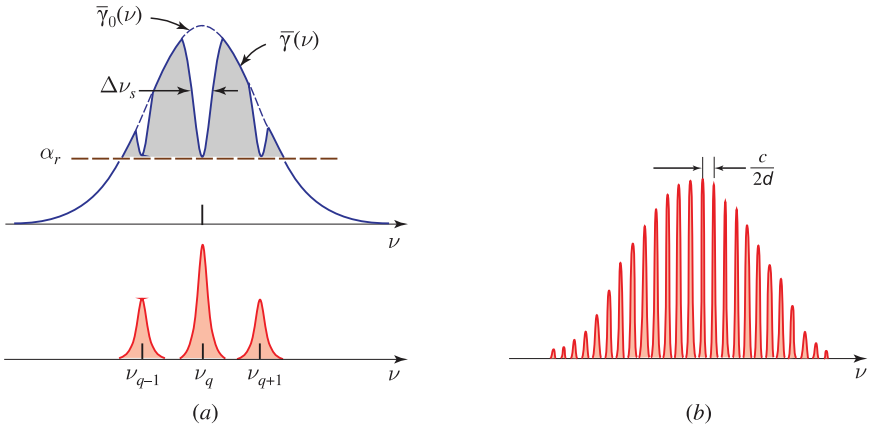


Figure 16.2-7 (a) Laser oscillation occurs in an inhomogeneously broadened medium by each mode independently burning a hole in the overall spectral gain profile. The gain provided by the medium to one mode does not influence the gain it provides to other modes. The central modes garner contributions from more atoms, and therefore carry more photons than do the peripheral modes. (b) Spectrum of a typical inhomogeneously broadened multimode gas laser.

The process of saturation by hole burning progresses independently for the different modes until the gain is equal to the loss for each mode in steady state. Modes do not compete because they draw power from different, rather than shared, atoms. Many modes oscillate independently, with the central modes burning deeper holes and growing larger, as illustrated in Fig. 16.2-7(a). The spectrum of a typical multimode inhomogeneously broadened gas laser is displayed in Fig. 16.2-7(b). The number of modes is typically larger than that in homogeneously broadened media since spatial hole burning generally sustains fewer modes than spectral hole burning.

*Spectral Hole Burning in a Doppler-Broadened Medium

The lineshape function of a gas at temperature T arises from the collection of Doppler-shifted emissions from the individual atoms, which move at different velocities (Sec. 14.3D and Exercise 14.3-2). A stationary atom interacts with radiation of frequency ν_0. An atom moving with velocity v toward the direction of propagation of the radiation interacts with radiation of frequency $\nu_0(1 + v/c)$ whereas an atom moving away from the direction of propagation of the radiation interacts with radiation of frequency $\nu_0(1 - v/c)$. Because a radiation mode of frequency ν_q travels in both directions as it bounces back and forth between the mirrors of the resonator, it interacts with atoms in two velocity classes: those traveling with velocity $+v$ and those traveling with velocity $-v$, such that $\nu_q - \nu_0 = \pm\nu_0 v/c$. It follows that the mode ν_q saturates the populations of atoms on both sides of the central frequency and burns two holes in the gain profile, as portrayed in Fig. 16.2-8. If $\nu_q = \nu_0$, only a single hole is burned in the center of the profile, of course.

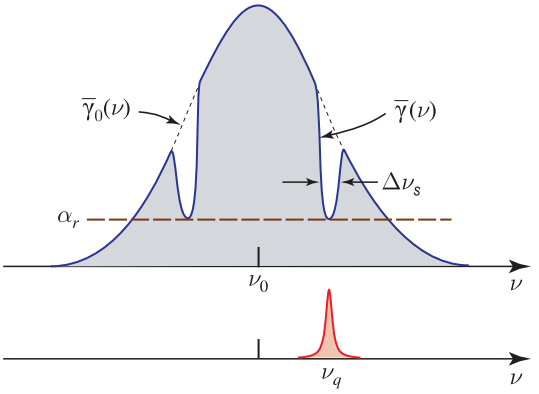


Figure 16.2-8 Hole burning in a Doppler-broadened medium. A probe wave at frequency ν_q saturates those atomic populations with velocities $v = \pm c(\nu_q/\nu_0 - 1)$ on both sides of the central frequency, burning two holes in the gain profile.

The steady-state power of a mode increases with the depth of the hole(s) in the gain profile. As the frequency ν_q moves toward ν_0 from either side, the depth of the holes increases, as does the power in the mode. As the modal frequency ν_q begins to approach ν_0, however, the mode begins to interact with only a single group of atoms instead of two, so that the two holes collapse into one. This decrease in the number of available active atoms when $\nu_q = \nu_0$ causes the power of the mode to decrease slightly. Thus, the power in a mode, plotted as a function of its frequency ν_q, takes the form of a bell-shaped curve with a central depression, known as the **Lamb dip**, at its center (Fig. 16.2-9).

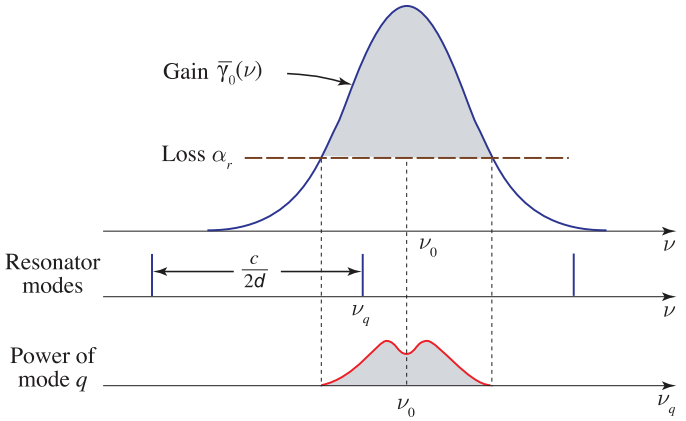


Figure 16.2-9 Power in a single laser mode of frequency ν_q in a Doppler-broadened medium whose gain profile is centered about ν_0. Rather than providing maximum power at $\nu_q = \nu_0$, it exhibits the Lamb dip.

C. Spatial Distribution and Polarization

Spatial Distribution

The spatial distribution of the emitted laser light depends on the geometry of the resonator and on the shape of the active medium. In the laser theory developed to this point we have ignored transverse spatial effects by assuming that the resonator is constructed of two parallel planar mirrors of infinite extent and that the space between them is filled with the active medium. In this idealized geometry the laser output is a plane wave propagating along the axis of the resonator. But as is evident from Chapter 11, this planar-mirror resonator is highly sensitive to misalignment and laser

resonators usually have spherical mirrors. As indicated in Sec. 11.2, the spherical-mirror resonator supports a Gaussian beam, which was studied in detail in Sec. 3.1. A laser using a spherical-mirror resonator may therefore give rise to an output that takes the form of a Gaussian beam.

It was also shown (in Sec. 11.2D) that the spherical-mirror resonator supports a hierarchy of transverse electric and magnetic modes denoted $\mathrm{TEM}_{l,m,q}$. Each pair of indices (l, m) defines a transverse mode with an associated spatial distribution. The $(0, 0)$ transverse mode is the Gaussian beam (Fig. 16.2-10). Modes of a higher l and m form Hermite–Gaussian beams (Sec. 3.3 and Fig. 3.3-2). For a given (l, m), the index q defines a number of longitudinal (axial) modes of the same spatial distribution but of different frequencies ν_q (which are always separated by the longitudinal-mode spacing $\nu_F = c/2d$, regardless of l and m). The resonance frequencies of two sets of longitudinal modes belonging to two different transverse modes are, in general, displaced with respect to each other by some fraction of the mode spacing ν_F, as indicated in (11.2-34).

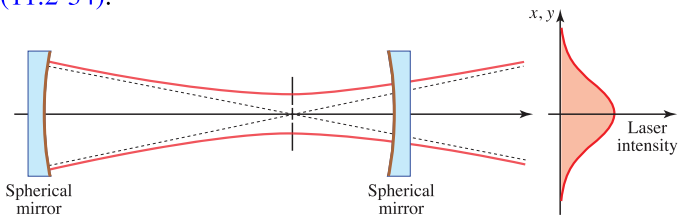


Figure 16.2-10 The laser output for the $(0, 0)$ transverse mode of a spherical-mirror resonator takes the form of a Gaussian beam.

Because of their different spatial distributions, different transverse modes experience different gains and losses. The $(0, 0)$ Gaussian mode, for example, is the most confined about the optical axis and therefore suffers the least diffraction loss at the mirror boundaries. The $(1, 1)$ mode vanishes at points on the optical axis (Fig. 3.3-2); thus if the laser mirror were blocked by a small central obstruction, the $(1, 1)$ mode would be completely unaffected, whereas the $(0, 0)$ mode would suffer significant loss. Higher-order modes occupy a larger volume and therefore can have larger gain. This disparity between the losses and/or gains of different transverse modes in different geometries determines their competitive edge in contributing to the laser oscillation, as illustrated in Fig. 16.2-11.

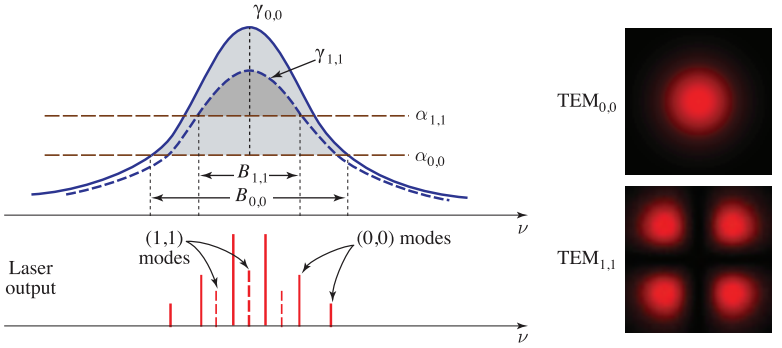


Figure 16.2-11 The gains and losses for two transverse modes, say $(0, 0)$ and $(1, 1)$, usually differ because of their different spatial distributions. A mode can contribute to the output if it lies in the spectral band of width B within which the gain coefficient exceeds the loss coefficient. Allowed longitudinal modes associated with each transverse mode are shown.

In a homogeneously broadened laser, the strongest mode tends to suppress the gain for the other modes, but spatial hole burning can permit a few longitudinal modes to oscillate. Transverse modes can have substantially different spatial distributions so that they can readily oscillate simultaneously. A mode whose energy is concentrated in a given transverse spatial region saturates the atomic gain in that region, thereby burning a spatial hole there. Two transverse modes that do not spatially overlap can coexist without competition because they draw their energy from different atoms. Partial spatial overlap between different transverse modes and atomic migrations (as occur in gases) allow for mode competition.

Lasers are often designed to operate on a single transverse mode; this is usually the $(0, 0)$ Gaussian mode because it has the smallest beam diameter and can be focused to the smallest spot size (Sec. 3.2). Oscillation on higher-order modes can be desirable, on the other hand, for purposes such as generating large optical power.

Polarization

Each (l, m, q) mode has two degrees of freedom, corresponding to two independent orthogonal polarizations. These two polarizations are regarded as two independent modes. Because of the circular symmetry of the spherical-mirror resonator, the two polarization modes of the same l and m have the same spatial distributions. If the resonator and the active medium provide equal gains and losses for both polarizations, the laser will oscillate on the two modes simultaneously, independently, and with the same intensity. The laser output is then unpolarized (Sec. 12.4).

Unstable Resonators

Though our discussion has focused on laser configurations that make use of stable resonators (Fig. 11.2-3), the use of **unstable resonators** offers a number of advantages in the operation of high-power lasers, including the following: (1) a greater portion of the gain medium contributes to the laser output power as a result of the availability of a larger modal volume; (2) higher output powers may be obtained from operation on the lowest-order transverse mode, rather than on higher-order transverse modes as in the case of stable resonators; and (3) high output power can be attained by making use of purely reflective water-cooled optics that permits the laser light to spill out around the mirror edges.

D. Mode Selection

A multimode laser may be operated on a single mode by making use of an element inside the resonator to provide loss sufficient to prevent oscillation on undesired modes.

Selection of a Laser Line

An active medium with multiple transitions (atomic lines) whose populations are inverted by the pumping mechanism will produce a multiline laser output. A particular line may be selected for oscillation by placing a prism inside the resonator, as shown schematically in Fig. 16.2-12.

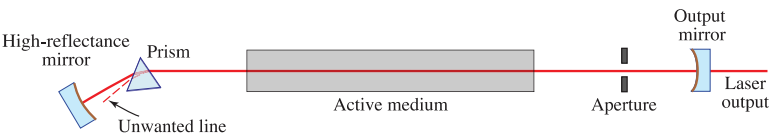


Figure 16.2-12 A particular atomic line may be selected by the use of a prism placed inside the resonator. A transverse mode may be selected by means of a spatial aperture of carefully chosen shape and size.

The prism is adjusted such that only light of the desired wavelength strikes the highly reflecting mirror at normal incidence and can therefore be reflected back to complete the feedback process. By rotating the prism, one wavelength at a time may be selected. Argon-ion lasers, as an example, often contain a rotatable prism in the resonator to allow the choice of one of six common laser lines, stretching from 488 nm in the blue to 514.5 nm in the blue–green. A prism can only be used to select a line if the other lines are well separated from it. It cannot be used, for example, to select one longitudinal mode from another; adjacent modes are too closely spaced for the dispersive refraction provided by the prism to distinguish them.

Selection of a Transverse Mode

Different transverse modes have different spatial distributions, so that an aperture of controllable shape placed inside the resonator may be used to selectively attenuate undesired modes (Fig. 16.2-12). The laser mirrors may also be designed to favor a particular transverse mode.

Selection of a Polarization

A polarizer may be used to convert unpolarized light into polarized light. It is advantageous, however, to place the polarizer inside the resonator rather than outside it since an external polarizer wastes half the output power generated by the laser. The light transmitted by an external polarizer can also suffer from noise arising from the fluctuation of power between the two polarization modes (mode hopping). An internal polarizer creates high losses for one polarization so that oscillation on its corresponding mode never even begins and the atomic gain is devoted solely to the surviving polarization. An internal polarizer is usually implemented by means of Brewster windows (Sec. 6.2 and Exercise 6.2-1), as illustrated in Fig. 16.2-13.

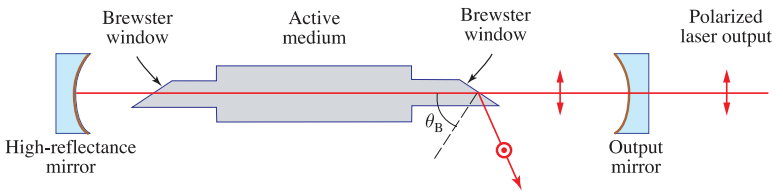


Figure 16.2-13 The use of Brewster windows in a gas laser provides a linearly polarized output laser beam. Light polarized in the plane of incidence (the TM wave) is transmitted without reflection loss through a window placed at the Brewster angle. The orthogonally polarized (TE) mode suffers reflection loss and therefore does not oscillate.

Selection of a Longitudinal Mode

A single longitudinal mode can be selected by appropriately modifying the resonator. The number of longitudinal modes in an inhomogeneously broadened laser (e.g., a Doppler-broadened gas laser) is the number of resonator modes contained in a frequency band B within which the atomic gain is greater than the loss (Fig. 16.2-4). At first blush, there are two obvious alternatives for operating a laser on a single longitudinal mode, but neither turns out to be satisfactory for the reasons indicated: 1) the loss can be increased sufficiently so that only the mode with the largest gain oscillates, but the surviving mode is then itself weak; 2) the longitudinal-mode spacing, $\nu_F = c/2d$, can be increased by reducing the resonator length, thereby leaving only a single mode within the band B, but this reduces the volume of the active medium and hence also results in a weak surviving mode.

Rather, intracavity frequency-selective elements can be conveniently used to alter the frequency spacings of the allowed resonator modes and thereby permit single longitudinal-mode operation. Two commonly used configurations are detailed below: the use of an intracavity tilted etalon and the use of multiple-mirror resonators.

Intracavity tilted etalon. A Fabry–Perot resonator whose mirror separation d_1 is much shorter (thinner) than the laser resonator, i.e., an *intracavity tilted etalon*, may be used for longitudinal-mode selection, as illustrated in Fig. 16.2-14. Modes of the thin etalon have large spacing $c/2d_1 > B$, so that only one etalon mode can fit within the laser amplifier bandwidth. The etalon is designed so that one of its modes coincides with the resonator longitudinal mode exhibiting the highest gain (or any other desired mode). The etalon may be fine-tuned by means of a slight rotation, by changing its temperature, or by slightly changing its width d_1 with the help of a piezoelectric (or other) transducer. The etalon is slightly tilted with respect to the resonator axis to prevent reflections from its surfaces from reaching the resonator mirrors and creating undesired additional resonances. The etalon is usually temperature stabilized to assure frequency stability.

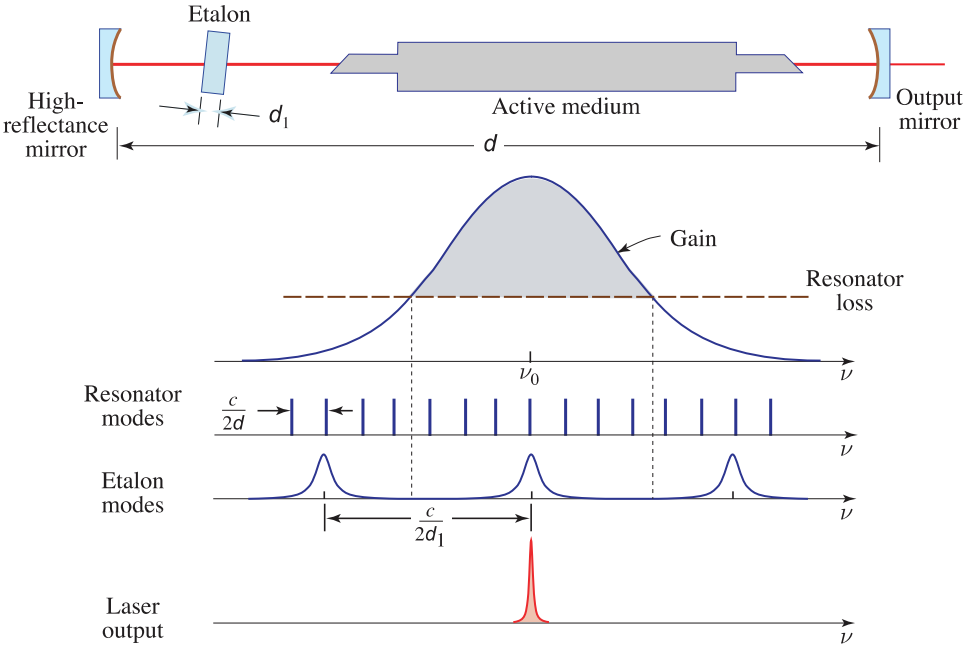


Figure 16.2-14 Longitudinal mode selection by use of a thin intracavity etalon. Oscillation occurs at frequencies where a mode of the resonator coincides with an etalon mode; both must, of course, lie within the spectral window where the gain of the medium exceeds the loss.

Multiple-mirror resonators. Mode selection can also be achieved by making use of *multiple-mirror resonators*. Several configurations are illustrated in Fig. 16.2-15. Mode selection may be achieved by means of two coupled resonators of different lengths [Fig. 16.2-15(a)]. The resonator in Fig. 16.2-15(b) consists of two coupled cavities, each with its own gain — in essence, two coupled lasers. Yet another configuration makes use of a resonator coupled with an interferometer [Fig. 16.2-15(c)].

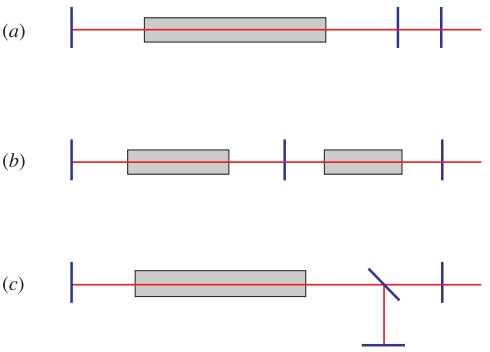


Figure 16.2-15 Longitudinal mode selection achieved with the help of: (*a*) two coupled resonators (one passive and one active); (*b*) two coupled active resonators; (*c*) a coupled resonator–interferometer.

Laser Linewidth

The FWHM spectral width $\Delta\nu_{\mathrm{L}}$ of the output of a single-mode laser can in principle be far smaller than the spectral width of a resonator mode, $\delta\nu \approx \nu_F/\mathcal{F}$. The width $\Delta\nu_{\mathrm{L}}$ is limited by the so-called Schawlow–Townes linewidth, which accommodates random-phase spontaneous-emission contributions that combine with the laser-oscillation mode. The FWHM value of the Schawlow–Townes linewidth for a four-level system, which is given by $\Delta\nu_{\mathrm{ST}} = \pi h\nu\,(\delta\nu)^2/P_o$, can be minimized by: 1) increasing the resonator length d, which decreases the free spectral range $\nu_F = c_o/2nd$ and thus $\delta\nu$; 2) reducing the resonator loss, which increases $\mathcal{F}$ and hence reduces $\delta\nu$; and 3) increasing the laser output power P_o. In carefully controlled experiments, it is possible to approach the Schawlow–Townes laser-linewidth limit. However, most lasers have linewidths substantially greater than this limit as a result of external effects such as resonator mechanical vibrations, active-medium temperature fluctuations, and pump-power fluctuations. Free-running lasers usually exhibit linewidths that lie in the kHz to GHz range. Single-mode laser linewidths can be reduced to the mHz domain by various stabilization techniques, although attaining a linewidth below 1 Hz is challenging.

16.3 TYPES OF LASERS

Laser amplification and oscillation is ubiquitous; it occurs in an enormous variety of media, including solids (crystals, glasses, fibers, powders), gases (atomic, ionic, molecular, excimeric), and liquids (organic-dye solutions). A single biological cell can serve as a laser when genetically programmed to produce green fluorescent protein. Plasmas support laser action in the extreme-ultraviolet and X-ray regions, and relativistic electrons wiggling in a magnetic field serve as an active medium for the free-electron laser. We discuss a number of examples of lasers in these various categories.

Parameters associated with lasers assume a wide range of values and extend over many orders of magnitude:

- ***Physical sizes*** range from nanometers to kilometers; interstellar molecular clouds exhibiting maser action extend over terameters.
- ***Emission frequencies*** span nearly 10 orders of magnitude, from GHz in the microwave to EHz in the hard-X-ray.
- ***Spectral linewidths*** extend over more than 15 orders of magnitude, from mHz to THz.
- ***Peak powers*** stretch 24 orders of magnitude, from nanowatts to petawatts.
- ***Pulse durations*** of lasers and laser-based systems reach from tens of attoseconds to CW.

A. Solid-State Lasers

The energy-level diagrams of several solid-state laser materials (ruby, alexandrite, Nd^{3+}:YAG, and Nd^{3+}:glass) were displayed in Figs. 14.1-4 and 14.1-5, and the operation of several solid-state laser amplifiers (ruby, Nd^{3+}:glass, and Er^{3+}:silica fiber) were examined in Figs. 15.3-1, 15.3-3, and 15.3-6, respectively. A number of characteristics of the principal laser transitions in these, and other, doped dielectric media are summarized in Table 15.3-1. When suitably pumped and placed in an optical resonator providing feedback, all of these solid-state materials behave as laser oscillators. When ground into powders, some solid-state laser materials function as random lasers.

Crystalline, ceramic, and glass hosts. There are numerous varieties of solid-state lasers since dozens of transparent dielectric media are commonly used as host materials to accommodate many kinds of active dopant ions. Crystalline hosts include oxides, garnets, fluorides, vanadates, and double-tungstates. The most common host materials are Al_2O_3 (sapphire), $Y_3Al_5O_{12}$ (yttrium aluminum garnet or YAG), $Lu_3Al_5O_{12}$ (lutetium aluminum garnet or LuAG), $YLiF_4$ (yttrium lithium fluoride or YLF), YVO_4 (yttrium vanadate, also known as yttrium orthovanadate), $KY(WO_4)_2$ (potassium yttrium tungstate or KYW), and $KGd(WO_4)_2$ (potassium gadolinium tungstate or KGW). Semiconductor crystals such as ZnS and ZnSe also serve as suitable hosts for solid-state lasers.

High optical-quality transparent polycrystalline **ceramic hosts**, with the same compositions as their single-crystal counterparts, are increasingly being used because of the many salutary features they offer, including increased power and efficiency, reduced cost, and options for fabricating flexible composite structures.

Glass hosts also enjoy wide use; these include silicate-based compositions (such as noncrystalline SiO_2, which is fused silica) and phosphate-based compositions, which have long been favored for high-power and pulsed-laser applications (see, e.g., Sec. 15.3B). Because they are poor conductors of heat, however, glass lasers are principally used in systems that operate at very high powers with low duty cycles. Notable exceptions are the glass hosts used in fiber lasers, which have large area-to-volume ratios that facilitate cooling.

Comparing the characteristics of lasers that make use of crystalline or ceramic hosts with those that use glass hosts reveals that the former category typically offers homogeneous broadening (see Sec. 14.3D) with narrower linewidths and correspondingly lower laser thresholds, higher thermal conductivities, and increased resistance to solarization (darkening caused by the ultraviolet component of flashlamp light). In contrast, glass hosts, which exhibit inhomogeneous broadening, have a number of distinct merits: they are isotropic, easily fabricated with high optical quality and homogeneous doping, they retain their optical finishes, and they are readily grown in large sizes (Sec. 15.3B).

Dopant ions. The lion's share of dopant ions used as active laser media in host crystals are transition-metal or lanthanide-metal (rare-earth) ions, but actinide-metal ions are also occasionally employed (Table 14.1-1). The dopant ions are generally dispersed throughout the host and act as independent radiators, much as organic-dye ions behave in a solvent. The dopant concentration (molar percentage) typically lies in the vicinity of 1%; however, it can be as small as 0.01% or as large as 50%, depending on the dopant, host material, and application. To minimize strain, the host material is generally chosen so that the active dopant ion is comparable in atomic size to the substituted atom.

Trivalent chromium ions doped into sapphire (Cr^{3+}:Al_2O_3), i.e., ruby, was the first material to be crafted into a laser (page 657). Ruby suffers from low efficiency,

however, because it is a three-level system. Alexandrite (Cr^{3+}:$BeAl_2O_4$), another early Cr^{3+}-doped solid-state laser, finds occasional use in dermatology. When used as an active laser ion, Cr^{3+} is nowadays often doped into colquiriite materials such as $LiCaAlF_6$ (LiCAF), $LiSrAlF_6$ (LiSAF), or $LiSrGaF_6$ (LiSGaF). Alexandrite and the Cr^{3+}-doped colquiriites may be efficiently pumped by red AlInGaP laser diodes.

Of the vast array of host and dopant-ion combinations for solid-state lasers, among the most commonly encountered are Nd^{3+}:YVO_4, Nd^{3+}:YAG, Yb^{3+}:YAG, Ti^{3+}:sapphire, and Cr^{2+}:ZnS, and we consider these in turn. Many other solid-state lasers also belong to the family of rare-earth-doped dielectrics, including Er^{3+}:YAG, Ho^{3+}:YAG, and Tm^{3+}:YAG. As discussed in Sec. 14.1B, the energy levels of the rare-earth ions (but not the sublevels thereof) are essentially independent of the host material because their $4f$ electrons are well shielded from the lattice by the filled $5s$ and $5p$ subshells (Table 14.1-1).

Diode-pumped solid-state lasers. Solid-state lasers that are optically pumped by laser diodes (or bars or stacks of laser diodes) are known as **diode-pumped solid-state** (DPSS) lasers. These devices convert the relatively broadband, multimode output of laser diodes into the narrowband, single-mode output of solid-state lasers. DSPPs are compact and highly efficient devices that have excellent beam quality. They offer a variety of wavelengths, dictated by their attendant electronic and vibronic transitions, which can be augmented by harmonic generation and other forms of optical frequency conversion (Chapter 22). Diode-pumped solid-state lasers find wide application in industry, medicine, and research.

Neodymium-Doped Yttrium Vanadate

Nd^{3+}:YVO_4 is a dielectric medium with refractive index $n \approx 2.0$. The host material is transparent over a broad range of wavelengths, from 0.3 to 2.5 μm. The energy levels associated with lasing as a *four-level system* are illustrated in Fig. 16.3-1; the laser threshold is substantially lower than that of three-level ruby.

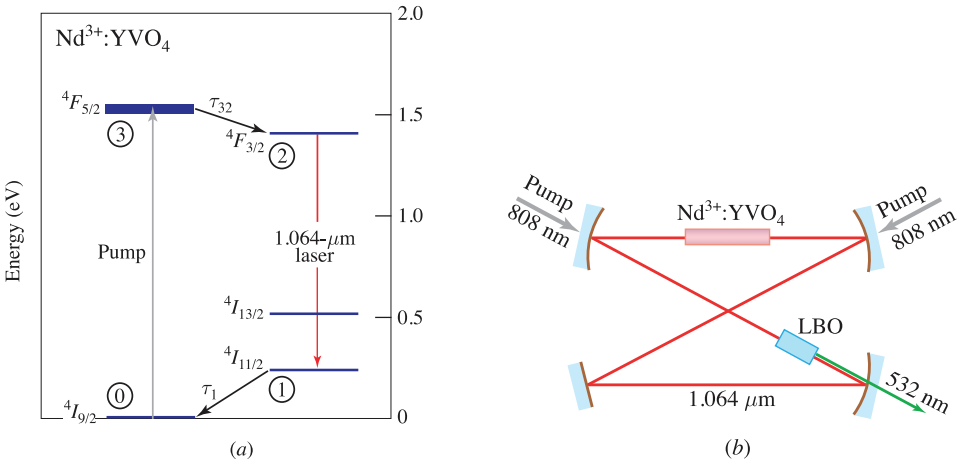


Figure 16.3-1 (a) Selected energy levels of Nd^{3+}:YVO_4. The red arrow indicates the principal laser transition at a wavelength of 1.064 μm in the near infrared. The four interacting energy levels are indicated by encircled numbers. (b) Configuration of a Nd^{3+}:YVO_4 laser with an intracavity frequency-doubling lithium-triborate (LBO) crystal that generates light at $\lambda_o/2 = 532$ nm via second-harmonic generation (Sec. 22.2A).

Optical pumping is readily implemented by making use of an 808-nm AlGaAs laser-diode array to raise the Nd^{3+} ions to level ③ ($^4F_{5/2}$) at 1.53 eV. The ②→① ($^4F_{3/2}$→

$^4I_{11/2}$) transition is responsible for laser action at the iconic wavelength of 1.064 μm. Alternatively, laser action on this transition can be attained by direct *in-band pumping* (Fig. 15.2-8) from level ⓪ to level ② at 880 nm (using AlGaAs laser diodes), or at wavelengths as long as 914 nm (using InGaAs laser diodes). In the latter case, the small quantum defect ($q = 14\%$) results in an output power > 10 W at 1.064 μm, with an optical-to-optical efficiency $\eta_o \approx 80\%$. Light from an intracavity frequency-doubled Nd^{3+}:YVO_4 laser often serves as a 532-nm pump for a Ti:sapphire laser (Fig. 16.3-3).

The $^4F_{3/2} \rightarrow {}^4I_{13/2}$ and $^4F_{3/2} \rightarrow {}^4I_{9/2}$ transitions also support laser action at wavelengths of 1.34 μm and at 914 nm, respectively, in the latter case as a *quasi-three-level system*. Frequency doubling of the stimulated emission at 914 nm generates blue light at 457 nm. Neodymium-doped yttrium vanadate is distinguished from neodymium-doped glass (Fig. 14.1-5) by its higher refractive index, homogeneous broadening, and smaller transition linewidth (Table 15.3-1).

Neodymium-Doped Yttrium Aluminum Garnet

Developed in 1964, Nd^{3+}:YAG, whose energy levels are displayed in Fig. 14.1-5, is one of the oldest and most widely used of all diode-pumped solid-state lasers. It is a compact system that provides high output power. Because the optically active $4f$ electrons in Nd^{3+} are shielded from the crystalline host, the energy levels of this material are similar to those of neodymium-doped yttrium vanadate (Fig. 16.3-1). Nd^{3+}:YAG lasers often incorporate intracavity doubling crystals, as illustrated in Fig. 16.3-1 for Nd^{3+}:YVO_4. Though it can be pumped by flashlamp, Nd^{3+}:YAG is most conveniently pumped as a *four-level system* using an AlGaAs laser-diode array at 808 nm, much as with Nd^{3+}:YVO_4. Crystals with lengths as short as a few hundred μm can serve as efficient, single-frequency, thin-disk lasers.

The most common laser line offered by Nd^{3+}:YAG is at a wavelength of $\lambda_o = 1.06415$ μm. The sublevels of the three manifolds associated with this laser transition are displayed in Fig. 14.1-6; the numbers of distinct sublevels are $(2\mathcal{J} + 1)/2 = 5$, 6, and 2, respectively. This particular laser line arises from a transition between the upper sublevel of the $^4F_{3/2}$ manifold at 1.4269 eV and the third-from-bottom sublevel of the $^4I_{11/2}$ manifold at 0.2616 eV. When frequency doubled, this transition provides green light at 532 nm. Transitions among the different sublevels within the upper and lower laser manifolds offer a multitude of possible laser wavelengths that span the wavelength range between 1.052 and 1.122 μm. In particular, lasing can be achieved at $\lambda_o = 1.12238$ μm via a transition between the lower of the two levels in the $^4F_{3/2}$ manifold at 1.4165 eV and the highest of the levels in the $^4I_{11/2}$ manifold at 0.3117 eV. This represents the longest wavelength that can be attained using a transition between these manifolds; when frequency doubled, this transition yields yellow-green light at $\lambda_o = 561$ nm.

Nd^{3+}:YAG can also be operated as a *quasi-three-level system* on the $^4F_{3/2} \rightarrow {}^4I_{9/2}$ transition, generating light at 946 nm; intracavity frequency doubling then provides blue light at 473 nm. Many other possibilities abound since oscillation on the dominant transition can be suppressed by making use of photonic-crystal filters, for example. As with Nd^{3+}:YVO_4, *in-band pumping* (Fig. 15.2-8) at 880 nm can be used to raise the Nd^{3+} ions from the ground state directly to the upper manifold, a scheme that has the merit of a small quantum defect.

The principal disadvantages of Nd^{3+}:YAG relative to Nd^{3+}:YVO_4 are its narrower $^4F_{5/2}$ absorption band (rendering it more sensitive to wavelength variations in the pump laser diodes), higher threshold, lower slope efficiency, and unpolarized output. A distinct advantage, however, is that it can tolerate approximately three times the thermal fracture stress. Indeed, when configured appropriately, Nd^{3+}:YAG can produce multi-kilowatt output powers. It continues to be the workhorse of diode-pumped solid-state lasers.

EXAMPLE 16.3-1. ***Power-Conversion Efficiency of a Diode-Pumped Nd^{3+}:YAG Laser.*** A 1-cm-long water-cooled laser-diode bar comprising 25 AlGaAs broad-area laser diodes emitting at a wavelength of 808 nm is used to pump a Nd^{3+}:YAG laser rod operating at 1064 nm on the basis of a four-level pumping scheme. Each laser diode emits 4 W and is spaced 0.4 mm from its neighbor in the form of a 1D array. The bar consumes 200 W of electrical power and delivers 100 W of optical power so its power-conversion efficiency is 50%. In accordance with the expression for the quantum defect q provided in (15.2-32), the pumping efficiency is $1 - q = 808/1064 \approx 76\%$; furthermore, the transfer of pump radiation to, and useful absorption by, the Nd^{3+}:YAG laser rod has an efficiency of about 65%. Mode matching and optical energy retention in the resonator is 60% efficient. The overall power-conversion efficiency η_e of the electrically driven Nd^{3+}:YAG laser is therefore 15%. Pumping this laser with a single laser-diode bar that consumes 200 W of electrical power results in an output of 30 W at 1064 nm. The output power can be increased by pumping with stacks of laser-diode bars. CW and pulsed Nd^{3+}:YAG lasers are widely used in manufacturing, materials processing, and medicine, and have myriad other applications, such as rangefinding.

Ytterbium-Doped Yttrium Aluminum Garnet

Yb^{3+}:YAG thin-disk lasers, which operate on the basis of *quasi-three-level pumping*, make use of a 940-nm laser-diode pump (Fig. 16.3-2).

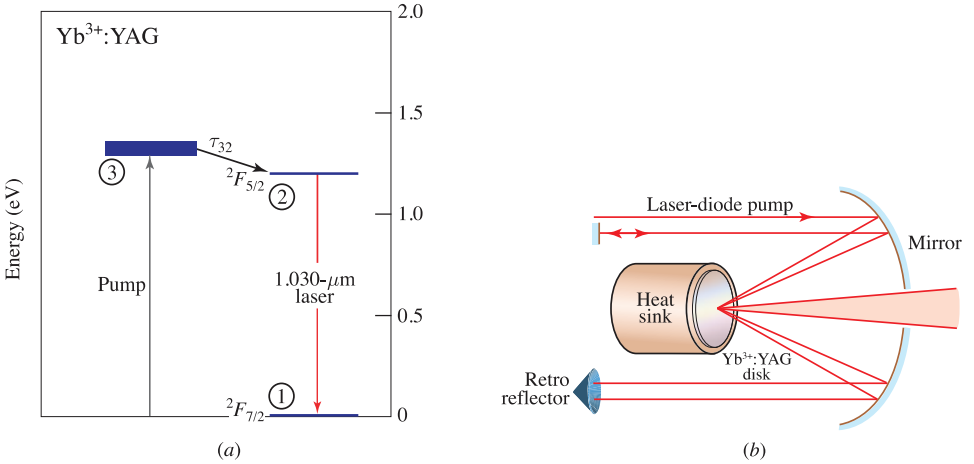


Figure 16.3-2 (*a*) Energy levels pertinent to the ytterbium-doped YAG laser transition at $\lambda_o = 1.030\ \mu$m. Yb^{3+}:YAG behaves as a quasi-three-level system at $T = 300°$ K and as a four-level system when cooled to $T = 77°$ K, which serves to reduce the thermal population in the lower laser level. (*b*) Schematic of a single-frequency, single-mode Yb^{3+}:YAG thin-disk laser. The active medium typically has a thickness of several hundred μm. The pump light is passed through the active medium some 20 times by an optical system that includes a parabolic mirror and a retroreflector. High gain is achieved by using Yb^{3+} doping levels $\approx 25\%$.

High-efficiency absorption of the pump light is achieved by passing it through the active medium multiple times with the help of suitably designed optics. High gain is attained by using high Yb^{3+} doping levels. The pump wavelength $\lambda_o = 940$ nm, usually provided by an InGaAs laser-diode array, is quite close to the laser wavelength $\lambda_o = 1030$ nm so that the quantum defect q is small. Hence, little heat is generated in the crystal.

Moreover, the thin-disk configuration allows the residual heat to be effectively removed by heat-sink mounting or fluid immersion, thereby permitting the TEM_{00} spatial mode to be maintained. Thin-disk lasers can generate hundreds of watts of single-mode

CW optical power and many kilowatts of multimode CW power at 1.030 μm. When frequency-doubled, the ytterbium-doped YAG laser provides a strong source of green light at 515 nm that has replaced the Ar^+ laser in many applications.

Yb^{3+} ions are also welcome dopants in hosts such as yttrium vanadate, as well as in various double-tungstates, borates, sesquioxides, and glasses. In particular, ytterbium-doped double-tungstates such as Yb^{3+}:KYW and Yb^{3+}:KGW make efficient use of *in-band pumping* at 981 nm (rather than at 940 nm), further reducing the quantum defect q and therefore the heat dissipated. Moreover, the large transition linewidths and good thermal properties associated with these hosts render them ideal for use as thin-disk, mode-locked lasers that can generate average powers in excess of 100 W and pulses with durations below 100 fs. Thin-disk lasers may also be fabricated using Nd^{3+}:YVO_4 and Nd^{3+}:YAG.

Titanium-Doped Sapphire

The Ti^{3+}:sapphire laser is widely used because it is tunable over a substantial range of wavelengths. Another of its merits is that it can be mode-locked to provide ultrashort pulses (Sec. 16.4D). The energy levels relevant to lasing in this *four-level system* are displayed in Fig. 16.3-3.

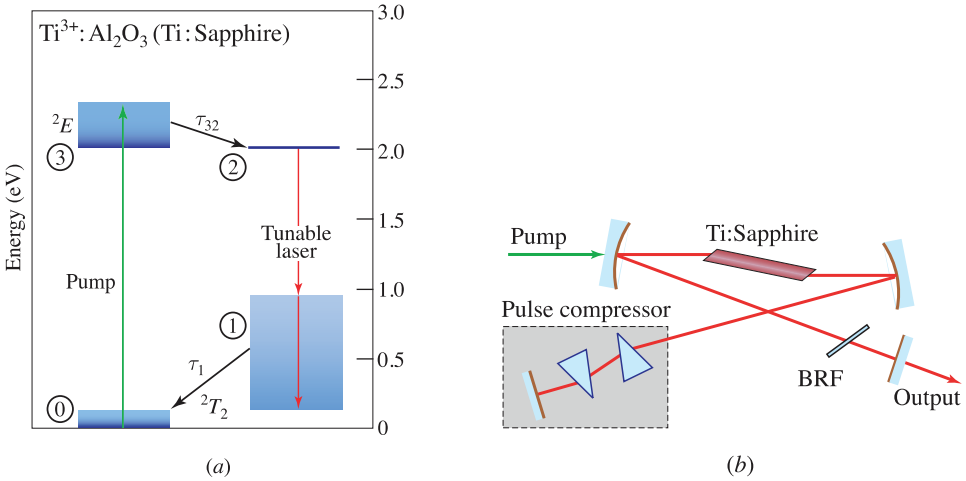


Figure 16.3-3 (*a*) Selected energy bands of Ti^{3+}:Al_2O_3. The red arrow indicates the principal laser transition of this vibronic system, which is tunable between 700 and 1050 nm. Dark-to-light shading in the bands indicates a decrease in relative occupancy. (*b*) Schematic diagram of a Ti^{3+}:Al_2O_3 mode-locked laser. The two prisms within the shaded box provide intracavity dispersion compensation (Sec. 23.2). Wavelength tuning over tens of nm is achieved by means of a rotatable birefringent filter (BRF) that acts as a bandpass filter for the polarized intracavity beam; tuning over a larger range is effected by adjusting one of the prisms. The green pump light is often provided by a frequency-doubled Nd^{3+}:YVO_4 laser, such as that illustrated in Fig. 16.3-1.

During the course of crystal growth, a small fraction ($\approx$ 1%) of the Al^{3+} ions in sapphire are replaced by Ti^{3+} ions. Like ruby, the material is principally sapphire and therefore has a refractive index $n \approx 1.76$. Optical pumping can be provided by a frequency-doubled Nd^{3+}:YVO_4 or Nd^{3+}:YAG DPSS laser operating at 532 nm (Fig. 16.3-1), or by a frequency-doubled Yb^{3+}:YAG DPSS laser operating at 515 nm (Fig. 16.3-2). Alternatively, direct pumping with a green laser-diode array can be used.

Each titanium ion, which has a single $3d^1$ active electron (Table 14.1-1), is surrounded by six oxygen atoms at an octahedral site. This ion is therefore subjected to

significant crystal-field and orbital interactions. As with other transition-metal ions in dielectric hosts, the titanium-doped sapphire energy levels displayed in Fig. 16.3-3 are designated by group-theoretical, rather than by term symbols (Sec. 14.1B). Moreover, the electronic energy levels are strongly coupled to the lattice vibrations, resulting in broad bands of vibronic states. Stimulated emission is thus accompanied by the simultaneous emission of one or more phonons. The occupancy of the ${}^2T_2$ band follows a Boltzmann distribution so that its upper reaches are essentially unoccupied and the system behaves as a four-level laser, as shown in Fig. 16.3-3(a).

The laser transition indicated by a red arrow in Fig. 16.3-3(a) can be tuned over a few tens of nm by making use of a rotatable birefringent filter installed at the Brewster angle within the cavity [Fig. 16.3-3(b)], which acts as a bandpass filter for the polarized intracavity beam. Greater changes in wavelength are effected by adjusting the internal optics since the cavity group velocity dispersion changes with wavelength. The net result is that a broad range of wavelengths, from 700 nm in the red to 1050 nm in the near infrared, can be accessed. A typical Ti^{3+}:Al_2O_3 laser provides ≈ 5 W of optical power when operated CW. When mode-locked, it can generate a sequence of 10-fs, 50-nJ pulses with peak power $P_p \approx 5$ MW, at a repetition rate of ≈ 80 MHz; repetition rates exceeding 10 GHz can also be attained. Applications range from multiphoton imaging to nonlinear optics to petawatt physics.

Because of the importance of lattice vibrations in the tunability of this laser, it is described as a **phonon-terminated** or **vibronic laser**. In general, a vibronic transition comprises a simultaneous change in the electronic and vibrational states of a system. The chromium-doped forsterite laser and the alexandrite laser (Fig. 14.1-4) also fall in this class, as does the dye laser (Sec. 16.3E) since molecular vibrations play the same role as lattice vibrations. It is often convenient to use *in-band pumping* with vibronic lasers.

Transition-Ion-Doped Zinc Chalcogenides

A family of vibronic lasers with continuous tunability, analogous to the Ti:sapphire laser but in the mid infrared, makes use of zinc-chalcogenide ceramic hosts doped with transition-metal ions. Among the most widely used members of this family are Cr^{2+}:ZnS and Cr^{2+}:ZnSe. Both ZnS and ZnSe are wide-bandgap II–VI semiconductor materials (Fig. 17.1-8) that are readily doped with Cr^{2+} ions that substitute for a fraction of the Zn^{2+} ions comprising the lattice. (This substitutional doping is to be distinguished from doping with ions of valences other than two for purposes of creating n- and p-type semiconductors.) These *four-level solid-state lasers* offer single-mode, linearly polarized, CW operation on the ${}^5E \rightarrow {}^5T_2$ transition with $P_o \approx 100$ W, $\mathbb{M}^2 \leq 1.1$, linewidth $< 1/2$ nm, and $\eta_c \approx 25\%$. Using an external cavity, they can be tuned over a wavelength range stretching from 1.9 to 3.0 μm. Though pumping can be effected by means of a laser-diode array, as with traditional DPSS lasers, Er^{3+}:silica-fiber pumps offer greater optical power in this wavelength range. Much as with the Ti^{3+}:sapphire laser, a mode-locked Cr^{2+}:ZnS laser can provide a sequence of 30-fs, 50-nJ pulses at a repetition rate between 80 MHz and 1 GHz, with $P_p \approx 1.7$ MW and $P_o \approx 5$ W.

Other members of this family of materials include Fe^{2+}:ZnS and Fe^{2+}:ZnSe, which can be tuned over a wavelength range stretching from 3.8 to 4.8 μm. These longer-wavelength lasers offer single-mode, linearly polarized CW operation with $P_o \approx 100$ mW, $\mathbb{M}^2 \leq 1.2$, and linewidths < 1 nm. In the current state of their development, however, the iron-doped devices require cooling to achieve CW operation. Efficient pumping is provided by a Tm^{3+}:silica-fiber laser. In spite of the fact that transition-ion-doped zinc-chalcogenide lasers operate only over a limited range in the mid IR, they have the merit of being continuously tunable. Other commonly encountered sources of radiation in the mid infrared are considered in Sec. 18.4D.

B. Fiber Lasers

With suitable feedback, rare-earth-doped fiber amplifiers can be made to operate as highly efficient **fiber lasers**. The most commonly used dopant ions for fiber lasers are neodymium (Nd^{3+}), ytterbium (Yb^{3+}), erbium (Er^{3+}), and thulium (Tm^{3+}). These ions offer useful laser transitions at near-infrared wavelengths, in the vicinity of 1.06, 1.07, 1.55, and 2.00 μm, respectively (among many other wavelengths). Silica-glass fibers are generally preferred as hosts because they are endowed with optical and mechanical properties superior to those of most other glass fibers. However, the transparency of silica glass diminishes appreciably for wavelengths longer than about 2.2 μm, so that fibers fabricated from fluorides and other glasses are also used (Sec. 10.5). To first order, the energy levels of rare-earth-ion laser transitions are little affected by the host material, as explained in Sec. 14.1B. Consequently, the energy-level diagrams for Nd^{3+}, Yb^{3+}, and Er^{3+} ions embedded in silica-fiber hosts are nearly indistinguishable from those presented in Figs. 14.1-5, 16.3-2, and 15.3-6, respectively, at the illustrated resolution. Fiber lasers, like diode-pumped solid-state (DPSS) lasers, are usually pumped by laser-diode arrays, although they are sometimes pumped by other fiber lasers. The first fiber laser, developed by Elias Snitzer in 1961 (one year after Maiman demonstrated the ruby laser), made use of a core of neodymium-doped glass.

Fiber lasers offer many salutary features, including:

- High optical power (the large area-to-volume ratio facilitates cooling).
- High power-conversion efficiency.
- Diffraction-limited beam quality.
- Stability against temperature variations and vibrations.
- Compact, robust, and maintenance-free construction.
- Ability to operate on low-gain transitions (the gain region can be arbitrarily long).

A simplified schematic illustrating a typical fiber laser that uses diode-laser pumping and **fiber Bragg-grating (FBG) reflectors** is displayed in Fig. 16.3-4. FBGs have the merit that they are robust and do not require realignment.

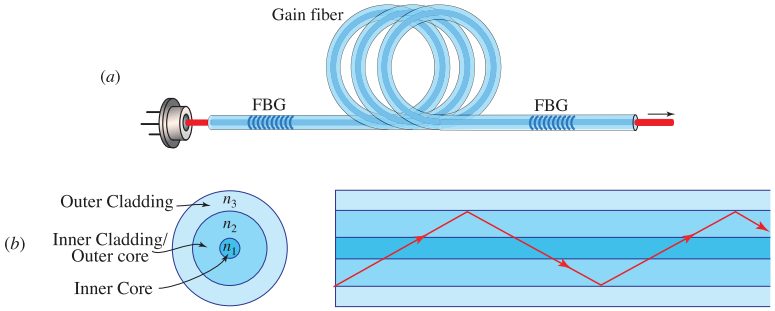


Figure 16.3-4 (*a*) Simplified schematic of a laser-diode-pumped fiber laser that makes use of fiber Bragg gratings (FBGs) as reflectors. Pumping often involves arrays of broad-area multimode laser diodes whose light is coupled into the outer core of the fiber via multimode couplers, sometimes in both the forward and backward directions. (*b*) Concentric double-clad fiber configuration. A single-mode inner core fosters single-transverse-mode oscillation. Many double-clad configurations are designed to increase the overlap between the inner core and the skew rays of the outer core (Fig. 10.1-2). For example, the inner core may be shifted off-center (toward the edge of the outer core), or the outer core may be rectangular, hexagonal, octagonal, or D-shaped.

A double-clad fiber configuration enables the laser mode to propagate in the inner core while the multimode pump light circulates in the inner cladding/outer core. This configuration is widely used to avoid the deleterious nonlinear effects that would arise

from a high-power pump concentrated in the small inner core.Fiber laser operation has been achieved in many configurations, and makes use of various forms of end pumping and side pumping. **Slab-waveguide fiber lasers**, also known as **ribbon fiber lasers**, have a rectangular cross section and can provide output powers greater than those offered by traditional fibers. **Photonic-bandgap fiber lasers** can be configured so that the light emerges radially from the full circumferential surface of the fiber, rather than axially.

Master-Oscillator Power-Amplifier (MOPA)

A **master-oscillator power-amplifier (MOPA)** is a configuration consisting of a master-oscillator (seed laser) followed by an optical amplifier to boost its output power. MOPAs may contain various combinations of fiber, diode-pumped solid-state, and semiconductor lasers and amplifiers. In the special case when the power-amplifier is a fiber-amplifier, a MOPA is also referred to as a **master-oscillator fiber-amplifier (MOFA)**. Though it is admittedly more complex to implement a MOPA than to use a single high-power laser to directly produce a given output power, the MOPA approach enjoys a number of distinct advantages. These stem from the relative ease of controlling a low-power seed laser, in which devices such as filters and modulators can be readily inserted because of the low-power intracavity environment; such devices can alternatively be placed between the oscillator and amplifier. This in turn facilitates the ability to optimize the MOPA output parameters, such as wavelength tuning range, laser linewidth, beam quality, and pulse duration. Moreover, the power-amplifier segment of the MOPA is subject only to the limited output power of the seed laser and not to the large intracavity optical power associated with full-fledged laser oscillation.

Another advantage of the MOPA configuration is its modularity. This feature offers a template for using concatenated amplifier stages with successively greater capacities, such as modal area and optical power. An example is provided by the MOPA at the National Ignition Facility (NIF), as described in Sec. 15.3B. A diode-pumped Yb^{3+}-doped fiber laser, by virtue of its stability and the high quality of its beam, serves as the master-oscillator. The 1-nJ, 5-ns pulse that it generates is split and sent to 192 main beamlines, where cascades of Nd^{3+}-doped glass power amplifiers increase the pulse energy to 20 kJ in each beamline.

MOPAs do have some disadvantages, aside from increased complexity. These include the possibility of amplifier-to-oscillator feedback and increased noise arising from amplified spontaneous emission (ASE), but these effects can often be mitigated by making use of optical isolators and by operating at saturation, respectively.

Ytterbium-Doped Silica-Fiber Lasers

Ytterbium-doped silica-fiber lasers offer excellent performance. Operating on the ${}^2F_{5/2} \rightarrow {}^2F_{7/2}$ transition (Fig. 16.3-2) as *quasi-three-level systems*, double-clad Yb^{3+}:silica-fiber lasers pumped by InGaAs laser-diode arrays at $\lambda_o = 940$ nm generate light in the wavelength range 1020–1200 nm. At very low powers ($P_o < 1$ W), the use of short (few-cm-long) fibers provides linearly polarized single longitudinal- and transverse-mode operation with laser linewidths in the kHz range. The laser linewidth increases with increasing output power, reaching ≈ 100 kHz at 100 W. Single fibers of modest lengths can provide multi-kW output powers in the form of single-mode, near-diffraction-limited beams ($\mathbb{M}^2 \approx 1.1$). The 9% quantum defect leads to large power-conversion efficiencies, $\eta_c \approx 40\%$. These lasers can also be operated at high average powers in Q-switched and mode-locked configurations (Sec. 16.4D; Example 16.4-3).

Tandem pumping. An effective way of increasing the power of a fiber laser beyond the multi-kW level is to make use of **tandem pumping**. In this technique, the output of one or more double-clad fiber lasers pumps another fiber laser or a fiber amplifier.

Using *in-band pumping*, the wavelength of the pump laser can be very close to that of the pumped fiber, resulting in a small quantum defect. This, along with the low divergence (high brightness) of the fiber laser pump, reduces heating and increases the system wall-plug efficiency. The net result is the availability of high-power operation using a reduced length of pumped fiber that offers a concomitant reduction in fiber nonlinearity.

EXAMPLE 16.3-2. ***High-Power, Tandem-Pumped, CW Yb^{3+}:Silica-Fiber MOFA.*** An example of tandem pumping in the form of a single-stage MOFA is provided by a 15-m-long Yb^{3+}:silica-fiber power-amplifier seeded by a 1-kW Yb^{3+}:silica-fiber master-oscillator. Using *in-band pumping* in the forward and backward directions, with 13 kW from Yb^{3+}:silica-fiber lasers operating at 1018 nm, this system produces 10 kW of optical power at 1070 nm in the form of a single-mode, near-diffraction-limited optical beam. The power can readily be scaled up to 20 kW CW and **coherent beam combination** can further boost the power to more than 30 kW CW in a single mode. **Incoherent beam combination** yields in excess of 100 kW CW in multimode operation, with $\mathbb{M}^2 \approx 50$.

Applications. Because of the many salutary features of ytterbium-doped fiber lasers and MOFAs, they are commercially available with a range of output powers and they find application in a broad variety of scientific and industrial venues. At low powers ($P_o \approx$ 1–100 W) these lasers are used for holography, interferometry, metrology, spectroscopy, sensing, 3D lidar, and optical trapping. At medium powers ($P_o \approx$ 0.1–1 kW) they find use for precision cutting, micro-drilling, micro/nanoscale machining, and additive manufacturing. At high powers ($P_o \approx$ 1–20 kW), ytterbium-doped fiber devices excel for materials processing and for cutting high-reflectance metals, while at ultra-high power levels ($P_o >$ 20 kW) they are invaluable for welding, drilling, precision cutting, annealing, and brazing.

High-power fiber lasers and MOFAs also find use as directed-energy weapons (DEWs). They are generally superior to other types of lasers (e.g., chemical, solid-state, and free-electron lasers) in this connection because of their high power, high efficiency, good beam quality, compactness, and immunity to vibrations and temperature fluctuations. The efficacy of Yb^{3+}:silica-fiber lasers in this capacity is demonstrated by the U.S. Navy Laser Weapon System (LaWS). This system comprises six 5.5-kW fiber lasers whose outputs are incoherently combined into a single 33-kW beam and transmitted through a beam director. An advanced version delivers a 150-kW beam.

Erbium-Doped Silica-Fiber Lasers

Erbium-doped silica-fiber lasers owe their prominence to their progenitors, the erbium-doped silica-fiber amplifiers (EDFAs) widely used in optical fiber communication systems (Secs. 15.3C and 25.1C). Operating on the ${}^4I_{13/2} \rightarrow {}^4I_{15/2}$ transition as *quasi-three-level systems* (Fig. 15.3-6), Er^{3+}:silica-fiber lasers pumped by strained-layer InGaAs laser-diode arrays at $\lambda_o = 980$ nm lase in the vicinity of the conventional (C) and long (L) telecommunications bands. The fibers are often co-doped with ytterbium to increase efficiency and reduce fiber length. By virtue of its large cross section and high doping density, Yb^{3+} is highly effective for absorbing the laser-diode pump radiation and transferring the excitation energy to the Er^{3+} ions. Co-doping is also useful for extending the wavelength range and for preventing erbium-ion clustering.

Narrow laser linewidths ($\approx$ 1 kHz) are available when CW erbium-doped silica-fiber lasers are operated at low power levels ($P_o \approx$ 10–1000 mW). At modest power levels ($P_o <$ 10 W), single-frequency (FWHM laser linewidth $\approx$ 50 kHz), linearly polarized, diffraction-limited ($\mathbb{M}^2 \approx 1.1$) operation is available in the wavelength range 1530–1625 nm, with power-conversion efficiencies $\eta_c \approx 10\%$. Multimode operation ($\mathbb{M}^2 \approx 10$) of an erbium-doped MOFA in the vicinity of 1567 nm (FWHM

linewidth $\approx$ 400 GHz) offers substantially higher output powers ($P_o \approx$ 2–5 kW) along with higher efficiencies ($\eta_c \approx 20\%$). Mode-locked operation at 1550 nm provides 5-μJ pulses of 500-fs duration at repetition rates up to 2 MHz, with peak powers $\approx$ 10 MW, average powers $\approx$ 10 W, and $\mathbb{M}^2 \approx 1.4$. These devices have applications in telecommunications, metrology, sensing, polymer welding, non-metal cutting, and low-loss power transmission.

The output power of the Er^{3+}:silica-fiber laser is limited in comparison with that available from the Yb^{3+}:silica-fiber laser, in part because of the difference in the quantum defect q [see (15.2-32)], which is far larger for erbium (37%) than for ytterbium (9%). Though the quantum defect for erbium can be reduced to 5% by using *in-band pumping* at 1480 nm, rather than quasi-three-level pumping at 980 nm, there are concomitant disadvantages to doing so.

Thulium-Doped Silica-Fiber Lasers

Thulium-doped silica-fiber lasers operate at yet longer wavelengths, namely in the 1.8–2.1-μm range, as *quasi-three-level systems* on the ${}^3F_4 \rightarrow {}^3H_6$ transition. Double-clad Tm^{3+}:silica-fiber lasers and MOFAs pumped by AlGaAs laser-diode arrays at $\lambda_o =$ 793 nm can generate output powers in excess of 500 W CW in the form of single-mode (FWHM laser linewidth $\approx$ 75 GHz), near-diffraction-limited ($\mathbb{M}^2 \approx 1.1$) beams; linear polarization can be obtained by operating at powers $<$ 200 W. The output power can be scaled up by making use of coherent beam combination. Multimode operation leads to optical powers $>$ 1 kW CW. Tunable pulsed operation in the wavelength range 1900–2050 nm can deliver 1-mJ pulses of 1-ns duration at repetition rates up to 50 kHz, with peak powers $\approx$ 1 MW, average powers $\approx$ 20 W, and $\mathbb{M}^2 \approx 1.1$.

Cross-relaxation, in this case a salutary effect, results in the creation of two Tm^{3+} ions in the upper laser level for each absorbed pump photon. This in turn leads to a high power-conversion efficiency, $\eta_c \approx 35\%$, notwithstanding the large (96%) quantum defect. Since light at a wavelength of ≈ 2 μm is strongly absorbed by water and biological soft tissue, the thulium-doped fiber laser is useful for surgery and lithotripsy. It is also suitable for remote sensing and for the processing of plastics that are transparent in the visible region.

Comparison of Fiber and DPSS Lasers

Fiber lasers are most notably distinguished from diode-pumped solid-state (DPSS) lasers (Sec. 16.3A) in that fiber resonators impose narrow lateral confinement and have long lengths (Chapter 10). This latter feature can yield high optical gain even for transitions with small gain coefficients. Another distinction is that the inhomogeneous broadening associated with glass-fiber hosts generally leads to broader transition linewidths than does the homogeneous broadening associated with crystalline or ceramic hosts.

Both classes of lasers provide high performance and both are expected to enjoy substantial further advances. It is useful, however, to highlight a few specific distinctions between the two classes of lasers in the current state of their development:

In general, fiber lasers are superior to DSPP lasers by virtue of their:

- Higher output power (CW and pulsed).
- Higher power-conversion efficiency.
- Superior beam quality that persists to high-power operation.
- Superior immunity to thermal and vibrational effects.
- Superior performance on low-gain transitions.

On the other hand, DPSS lasers have the edge over fiber lasers because they offer:

- Reduced nonlinearities (smaller length and larger spatial extent of active region).
- Reduced stimulated Raman and stimulated Brillouin scattering.
- Broad wavelength tunability associated with vibronic lasers (e.g., Ti^{3+}:sapphire).
- Possibility of using pump sources with poor beam quality.

C. Raman Lasers

Raman fiber lasers (RFLs). These devices operate on the basis of stimulated Raman scattering (SRS), a process considered in Sec. 14.5C and revisited in Sec. 15.3D in connection with Raman fiber amplifiers (RFAs). The process of stimulated Raman scattering is illustrated in Fig. 15.3-7: A signal photon of energy $h\nu_s$ stimulates the emission of a clone signal photon that is obtained by Stokes-shifting the pump photon by the Raman vibrational energy $h\nu_R$ so that the energy of the clone photon precisely matches that of the initial signal photon. The optical gain of a RFA is governed by the Raman gain coefficient γ_R set forth in (22.3-15); its bandwidth is determined by the vibrational spectrum of the glass host, as described in Sec. 15.3D and illustrated in Fig. 15.3-7.

Just as a rare-earth-doped fiber amplifier can be converted into a fiber laser by introducing optical feedback as shown in Fig. 16.3-4, so too can a Raman fiber amplifier be converted into a Raman fiber laser (RFL). Fiber Bragg gratings serve as reflectors that define the resonator, fostering oscillation at those frequencies where their reflectance is large (Sec. 7.1C). Distributed-feedback (DFB) resonator configurations can also be used (Sec. 18.3C). The oscillation frequency ν_s is reduced below the pump frequency ν_p by the Stokes frequency ν_R, which can take on any value within the vibrational spectrum of the glass host, as shown in Fig. 15.3-7. Though bulk Raman lasers were demonstrated early on, fiber implementation brought Raman technology to the fore for several reasons: 1) fibers offer long lengths and therefore large gains; 2) fibers can support large intensities in a single-mode core; 3) fibers are efficiently pumped by diode-pumped solid-state lasers; and 4) fibers readily accommodate multiple fiber Bragg gratings. We emphasize that the Raman laser makes use of Raman gain, rather than a population inversion and stimulated emission, so it differs from the usual laser in an essential way.

Cascaded Raman fiber lasers. A unique feature of the Raman interaction is that the Stokes shift is independent of the pump frequency. Indeed, the Stokes-shifted RFL oscillation frequency generated using a resonator comprising a particular pair of fiber Bragg gratings, say FBG1, can itself serve as a pump within the same fiber. As shown in Fig. 16.3-5(a), this second pump, which has reduced frequency $\nu_{p1} = \nu_p - \nu_R$, can then create a second-order Stokes-shifted oscillation, established at the frequency of maximum reflectance ν_{p2} of a second pair of fiber Bragg gratings, FBG2.

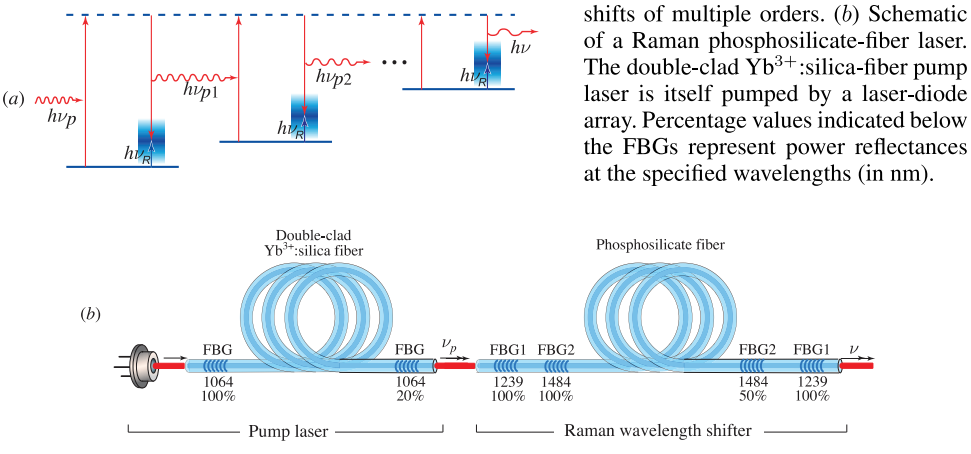


Figure 16.3-5 (a) Cascaded Stokes shifts of multiple orders. (b) Schematic of a Raman phosphosilicate-fiber laser. The double-clad Yb^{3+}:silica-fiber pump laser is itself pumped by a laser-diode array. Percentage values indicated below the FBGs represent power reflectances at the specified wavelengths (in nm).

The cascade can continue, using nested pairs of FBGs, until terminated by the use of an output coupler that directs light out of the fiber at the desired frequency. Raman fiber

lasers comprising multiple orders of Stokes shifts thus offer a greatly expanded range of possible wavelengths. They are known as **cascaded Raman fiber lasers**.

EXAMPLE 16.3-3. ***Cascaded Raman Phosphosilicate-Fiber Laser.*** A cascaded Raman fiber laser can be constructed by using a double-clad ytterbium-doped fiber laser that emits in the 1050–1120-nm region (Sec. 16.3B) as a pump for a 1-km length of phosphosilicate-fiber Raman wavelength shifter, as depicted in Fig. 16.3-5(b). The Yb^{3+}:silica-fiber pump laser, which is itself pumped by a laser-diode array operating in the vicinity of 960 nm, emits single-mode light that can be coupled into the single-mode Raman wavelength shifter more efficiently than can multimode light from a laser-diode array. As a specific example, assume we wish to convert ytterbium-laser pump light at 1064 nm (282 THz) to a longer wavelength, say 1484 nm (202 THz), so that the Raman fiber laser can be used to pump a Raman fiber amplifier or an erbium-doped fiber amplifier (Sec. 15.3D). Since we seek a significant frequency shift (80 THz), we make use of phosphosilicate fiber, which has a large Stokes shift, $\nu_R \approx 40$ THz, thereby enabling the desired frequency conversion using only two Stokes orders. We could alternatively use germanium-doped silica fiber, but this would require six Stokes orders since the Stokes shift for silica fiber is far smaller, $\nu_R \approx 13$ THz, as is clear from Fig. 15.3-7. As portrayed in Fig. 16.3-5(b), a first pair of fiber Bragg gratings, FBG1, is used to shift the 1064-nm pump light down by 40 THz in frequency to 1239 nm, while a second pair, FBG2, shifts the light down another 40 THz in frequency to the desired wavelength of 1484 nm. The FBG1 pair have reflectances of 100%, while the reflectance of one member of the FGB2 pair is reduced to 50% so that light can be coupled out of the RFL at 1484 nm. A compact, commercially available RFL, such as that shown in Fig. 16.3-5(b), can deliver tens of watts of single-mode CW optical power with a bandwidth of $\approx$ 2 nm at any desired wavelength within the range of 1100–1700 nm. The power-conversion efficiencies of the ytterbium-doped fiber pump laser and the pump/Raman-shifter combination are approximately 40% and 10%, respectively.

The levels of optical power obtainable from Raman lasers can be extended to hundreds of watts by making use of novel configurations, such as seed lasers in combination with Raman fiber amplifiers. Optical powers can be further increased to the multi-kW regime by employing mixed rare-earth/Raman amplification in conjunction with Yb^{3+}-doped fibers. Raman fiber lasers are useful for many applications, including laser pumping, optical fiber communications, supercontinuum generation, materials processing, and clinical medicine. They are particularly suitable for pumping Raman fiber amplifiers in dense wavelength-division-multiplexed (DWDM) systems and for the remote pumping of erbium-doped fiber amplifiers. Ultralow-threshold, high-efficiency, single-mode CW Raman lasers have also been fabricated using ultrahigh-Q toroidal silica microcavities. The most attractive feature of a RFL is that oscillation over a broad range of wavelengths is supported by suitable choice of the pump wavelength, fiber material, and fiber Bragg gratings.

Stimulated Brillouin scattering [Fig. 14.5-5(d)] can be used in an analogous way to construct Brillouin fiber lasers. The Brillouin frequency shift and bandwidth for silica fibers are in the vicinity of 10 GHz and 100 MHz, respectively — these values are orders of magnitude lower than those for the Raman frequency shift and bandwidth in the same material.

EXAMPLE 16.3-4. ***Silicon Raman Laser.*** Low-threshold, CW silicon Raman lasers that operate at room temperature have been fabricated using integrated-optic-ring resonators on silicon chips. Though Si has a high Raman gain coefficient, the substantial losses associated with free-carrier absorption induced by two-photon absorption must be limited to achieve high gain. A Si Raman laser of this kind has been implemented by using a 3-cm-long racetrack-shaped ring resonator consisting of a silicon-on-insulator (SOI) rib waveguide [Fig. 9.3-5(d)] with a rib width of 1.5 μm.[†]

[†] See H. Rong, S. Xu, Y.-H. Kuo, V. Sih, O. Cohen, O. Raday, and M. Paniccia, Low-Threshold Continuous-Wave Raman Silicon Laser, *Nature Photonics*, vol. 1, pp. 232–237, 2007.

Racetrack-shaped p-type and n-type regions, separated from each other by about 6 μm, hug the outside and inside contours of the i-type rib-waveguide resonator, respectively. A reverse-bias field applied between the doped regions serves to sweep out the electron–hole pairs generated by two-photon absorption that are responsible for loss. A bus waveguide, connected to the resonator via an integrated-photonic directional coupler, couples pump light at 1550 nm into the resonator and Raman laser light at 1686 nm out of the resonator. An output power $P_o = 50$ mW is obtained at a reverse bias of 25 V, with an optical-to-optical slope efficiency $\eta_s = 28\%$ and a threshold optical pump power $P_t = 20$ mW. Even under zero-bias conditions, the output power is greater than 10 mW and the threshold pump power is 26 mW. Silicon Raman lasers such as these can access wavelengths substantially longer than the bandgap wavelength of Si ($\lambda_g = 1.11$ μm), but they require auxiliary pump lasers, which is challenging for silicon (see Sec. 17.1B).

Cascaded silicon Raman lasers have also been fabricated. In one example, a pump at 1550 nm produces a Raman output at 1686 nm by virtue of a Stokes shift at $\nu_R = 15.6$ THz, along with an output at 1848 nm resulting from a second-order Stokes shift.[†] The single-mode, CW output at 1848 nm has an optical power $P_o = 5$ mW, a spectral width $\Delta\nu < 2.5$ MHz, an optical-to-optical slope efficiency $\eta_s \approx 3\%$, and an optical-to-optical efficiency $\eta_o \approx 1\%$. The threshold optical pump power is $P_t = 120$ mW.

D. Random Lasers

As discussed in Secs. 16.1 and 16.2, the oscillation frequencies of conventional lasers are determined by the Fabry–Perot resonator modes together with the gain profile of the active-medium resonant transition. The output light, transmitted through a partially reflecting exit mirror, typically has a narrow spectrum, strong directionality, and a high degree of temporal and spatial coherence. Scattering from the laser medium introduces loss and is assiduously avoided.

When scattering in the active medium is very strong, however, it itself can provide feedback. Random lasers operate on the basis of feedback provided by multiple scattering within a disordered gain medium, which serves as a closed 3D cavity. Photons traveling within the medium can be viewed as executing a random walk in 3D or, alternatively, the medium may be considered as a collection of low-Q cavities coupled by the randomly scattered photons [Fig. 16.3-6(a)]. Because strong scattering is associated with disordered media, lasers that operate on this principle are known as **random lasers**. In distinction to conventional lasers, the radiation scattered back to any location in the active medium has a random phase. The feedback is thus incoherent and intensity-based, rather than coherent and field-based. Inasmuch as resonant feedback is absent in such random lasers, the central oscillation frequency is governed by the active-medium gain profile. Random lasers can be implemented in 1D and 2D geometries as well, corresponding to scattering media in the form of fibers and plates, respectively.

When ground into powders, conventional solid-state laser materials such as ruby, Nd^{3+}:YAG, Nd^{3+}:glass, Ti^{3+}:sapphire, Cr^{2+}:ZnSe, and GaAs can function as random lasers known as **powder lasers** or **plasers**. In many such powders, the gain and scattering media are one-and-the-same, but this is not the only possible configuration. For example, rhodamine-6G dye molecules serve well as an active medium, while Al_2O_3 microparticles function as a scattering medium, when both are placed in a solution of methanol. Random-laser active media encompass inorganic dielectrics, polymers, liquids, dye solutions, dye-doped liquid crystals, disordered semiconductor nanostructures, and even biological tissues.

Substantial gain can be attained in random lasers because of the large overall path-length in the active medium that is engendered by the multiple scattering. The onset of lasing is characterized by two lengths: d_{st}, which represents the mean distance

[†] See H. Rong, S. Xu, O. Cohen, O. Raday, M. Lee, V. Sih, and M. Paniccia, A Cascaded Silicon Raman Laser, *Nature Photonics*, vol. 2, pp. 170–174, 2008.

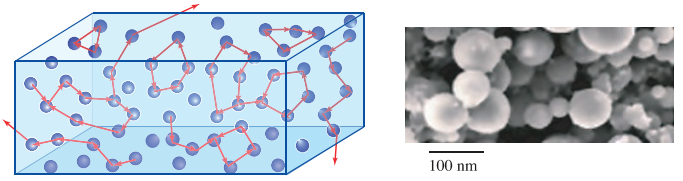


Figure 16.3-6 (*a*) A random laser relies on incoherent and nonresonant feedback provided by multiple scattering, as well as a long pathlength within the gain medium. The modes can be spatially localized or extended, and both versions can contemporaneously coexist. Recurrent scattering (illustrated schematically as loops) represents an alternative scenario in which the field repeatedly retraces one or more local paths that collectively serve as a local cavity, thereby providing coherent and resonant feedback. (*b*) Close-packed ZnO nanocrystallites serve as both the active medium and the scattering feedback elements in a microrandom laser.

traveled by a photon before stimulating the net emission of a clone photon; and d_{ex}, which is the mean distance traveled by a photon before it exits the medium. As the scattering strength increases, so too does d_{ex}. When the scattering increases to a level such that d_{ex} just equals d_{st}, each photon generates a clone photon, on average, before escaping from the medium and a chain reaction of induced photon generation becomes sustainable. Thus, the stronger the scattering, the lower the threshold pump power.

The hallmark of random lasers is the absence of directionality and spatial coherence of the emitted light. Indeed, the spatial lasing pattern from the face of a cuvette containing a powdered active medium can resemble that of a surface-emitting LED [Fig. 18.1-11(*a*)]. However, random lasers share many properties in common with conventional lasers, including their diversity. They can be pumped optically, electrically, or by electron beam. Lasing can take place over a broad range of wavelengths. The sizes of active regions can stretch from the microscopic to the macroscopic. Indeed, feedback via scattering appears to play an important role in astrophysical masers, such as those observed from molecular clouds of H_2O, OH, and SiO.

If the constituent particles of the active medium are sufficiently large and regularly shaped so that they support resonator modes, they can behave instead as random collections of individual microlasers, each with its own emission direction. Alternatively, a local configuration of scatterers can support resonances. If the scattering is recurrent [Fig. 16.3-6(*a*)], those return paths for which the optical gain exceeds the loss can serve as cavities. The ensuing laser emission is then sharply peaked at these fortuitous cavity-mode frequencies. Since different regions of the random medium support different collections of such return paths, the oscillation frequencies depend on the particular regions of the material being pumped and observed. Such lasers are sometimes called **coherent random lasers** because of their coherent feedback and spatially random cavity configurations; they are similar to collections of microlasers with random directions of emission (conventional microcavity lasers are considered in Sec. 18.5).

Clusters of scatterers can be used to fabricate individual **microrandom lasers**, in which light is confined to a volume of the order of a cubic wavelength by strong scattering rather than by reflection, as illustrated in Example 16.3-5.

EXAMPLE 16.3-5. ***ZnO Microrandom Laser.*** A closely packed collection of several thousand ZnO nanocrystallites [Fig. 16.3-6(*b*)], each of the order of tens of nm in diameter, can coalesce into a microcluster ≈ 1 μm in diameter. The nanocrystallites serve as both the gain medium and the scattering feedback elements of a microrandom laser. The emission wavelength $\lambda_o \approx 380$ nm lies near the bandgap wavelength of ZnO. Because the optical confinement arises from scattering, rather than from reflection at the surface of the microcluster, such microlasers need not have regular shapes and smooth surfaces.

E. Gas and Dye Lasers

Atomic and Ionic Gas Lasers

Atomic and ionic **gas lasers** such as He–Ne, Ar^+, and Kr^+, produce the beautiful multicolored beams that were long a staple of optics laboratories (Table 16.3-1). The Kr^+-ion laser, in particular, generates hundreds of milliwatts of optical power at wavelengths ranging from $\lambda_o = 350$ nm in the near-ultraviolet to 676 nm in the red. It can be operated simultaneously on a number of lines to produce "white laser light." Many other monoatomic species, and their ions, also serve as active laser media and operate at innumerable wavelengths in the near-infrared and visible regions. Nevertheless, atomic and ionic gas lasers are now used principally for specialized applications since diode-pumped solid-state lasers and laser diodes have far superior performance, aside from being tunable and physically more robust.

Molecular Gas Lasers

Molecular gas lasers such as the CO_2 laser (Table 16.3-1 and Fig. 14.1-7), which lases in the vicinities of $\lambda_o = 9.6$ and 10.6 μm in the mid infrared, can produce kilowatts of CW power with high efficiency, and has applications such as cutting, welding, and engraving. In years past, a favorite in the far-infrared was the methanol (CH_3OH) laser, which oscillates at $\lambda_o = 119$ and 124 μm as well as at myriad other wavelengths (Table 16.3-1). Indeed, most molecular transitions in the infrared region can be made to lase; even simple water vapor (H_2O) generates laser radiation at many wavelengths in the far infrared (Table 16.3-1). In recent years, however, quantum cascade lasers (QCLs) have come to the fore (Sec. 18.4D). Operating at room temperature and with high power-conversion efficiencies, these devices produce watts of CW power in the mid-infrared. QCLs have supplanted molecular lasers in essentially all applications, save those requiring optical powers in excess of 10 W.

Excimer and Exciplex Lasers

Excimer and exciplex lasers are important in the ultraviolet region of the spectrum. The term **excimer**, which is a contraction of the phrase "excited dimer," refers to a short-lived molecule that contains two atoms in an excited electronic state; the term **exciplex** is preferred when the two atoms are not identical. Noble-gas halides form exciplexes because the chemical behavior of an excited noble-gas atom is similar to that of an alkali atom, which readily reacts with a halogen. An example of an excimer laser is F_2, which lases at 157 nm in the far ultraviolet. Examples of exciplex lasers, along with their principal wavelengths of operation, are: ArF (193 nm), KrF (248 nm), XeCl (308 nm), and XeF (351 nm) (Table 16.3-1). ArF exciplexes can be formed, for example, by passing a 20-ns-duration current pulse through a gas mixture of Ar and F_2 to create a gas discharge. As the exciplexes return to the ground state, they emit a 150-mJ pulse of stimulated emission as the exciplex components dissociate (the individual atoms often repel each other). Since a lower laser level does not exist, exciplex lasers enjoy a built-in population inversion.

Since ultraviolet light does not penetrate deeply into most materials, exciplex lasers find use in applications involving the processing of delicate materials. The UV light from these lasers disrupts the molecular bonds at the surface of the material, vaporizing it via *ablation* rather than by heating, burning, or cutting. This, along with the substantial energy per pulse that can be generated, makes exciplex lasers suitable for the precision micromachining of polymers and for carrying out sensitive dermatological and refractive surgeries. The most important application of ArF lasers is the fabrication of microelectronic and photonic integrated circuits with feature sizes smaller than 10 nm. This is achieved by making use of optical lithography in conjunction with multiple patterning and liquid-immersion techniques (to increase the numerical aperture). Fea-

ture sizes continue to shrink as optical lithography migrates toward extreme-ultraviolet and X-ray wavelengths. Other nanofabrication techniques include electron-beam and focused ion-beam lithography.

Chemical Lasers

Chemiluminescence, the emission of light via a chemical reaction, is observed when the reaction between two or more substances releases sufficient energy to populate the excited state of a reaction product (Sec. 14.5A). **Chemical lasers**, which comprise mixtures of gases, are self-pumped in the sense that the pump energy derives from a chemical reaction in the active medium itself. The HF laser, which delivers megawatts of optical power in the 2.7–2.9-μm wavelength range, is perhaps the best known among this class of lasers. Its construction resembles that of a rocket engine: it contains a combustion chamber, nozzles, gas-injection mechanisms, and a resonator. As a simplified explanation of its operation, a mixture of H_2 and F_2 gases is subjected to an electric discharge, which results in the production of an HF molecule in an excited vibrational state, denoted HF^*. This molecule emits an infrared photon and dissociates. Its components in turn react with the H_2 and F_2 gases to create other vibrationally excited molecules, creating a chain reaction of sorts. Other chemical lasers include the "chemical oxygen–iodine laser" (COIL) and the "all gas-phase iodine laser" (AGIL).

In earlier decades, the high power and good beam quality of chemical lasers made them prime candidates for use as directed energy weapons (DEWs). In the end, however, they were found to be too bulky, too heavy, too inefficient, and too hazardous for shipboard use. In recent years, attention has shifted toward solid-state, diode-pumped, free-electron, and fiber lasers. Fiber lasers, in particular, are currently of substantial interest for the development of DEWs because they offer high power, high efficiency, good beam quality, compactness, and immunity to vibrations (Sec. 16.3B).

EXAMPLE 16.3-6. ***Deuterium Fluoride Chemical Laser.*** The most notorious chemical laser, perhaps, is the U.S. Navy's Mid-Infrared Advanced Chemical Laser (MIRACL), located at the White Sands Missile Range in New Mexico. This formidable DEW burns ethylene (C_2H_4) with nitrogen trifluoride (NF_3). The resulting free fluorine atoms combine with injected deuterium gas to form vibrationally excited deuterium fluoride molecules, DF^*. The photon emission, molecular dissociation, and creation of new vibrationally excited molecules is similar to that in the HF laser. However, the DF device lases on multiple lines in the 3.5–4.0-μm wavelength range, which falls within the MWIR atmospheric window so that its radiation is absorbed far less than that emitted by the HF laser. The DF laser produces multi-megawatt levels of CW radiation over durations ≈ 1 minute, in the form of a 14 cm $\times$ 14 cm square beam. It is the highest power CW chemical laser in the Western Hemisphere.

Dye Lasers

We include a brief description of **organic dye lasers** for completeness. Dye lasers played a central role in optics and photonics in decades past because of their ability to be tuned over a substantial range of wavelengths. The active medium of a dye laser is generally a solution of an organic dye compound (a carbon-based fluorescent soluble stain) in alcohol or water, with a concentration $\approx 10^{-4}$ M. The dye solution is usually rapidly circulated since organic dyes tend to decompose in the presence of light.

Dyes typically exhibit large transition linewidths, accommodating broad wavelength tunability as well as ultrashort pulse generation with passive mode locking. Polymethine dyes provide oscillation in the red and near infrared (0.7–$1.5\ \mu\mathrm{m}$), xanthene dyes lase in the visible (500–700 nm), coumarin dyes lase in the violet, blue, and green (390–500 nm), and scintillator dyes lase in the ultraviolet ($< 390\ \mathrm{nm}$). Rhodamine-6G,

the quintessential example of an active dye molecule, can be tuned over the wavelength range 560–640 nm. Dye molecules can also be imbedded in a polymer, glass, or crystalline host to form a **solid-state dye laser**. Dye lasers are typically *in-band pumped* (Fig. 14.1-8) with the help of an external laser such as a doubled Nd^{3+}:YAG laser or a laser diode. Fabry–Perot and ring-laser resonators are the most prevalent designs, with a prism or diffraction grating inserted in the beam path to provide wavelength tuning.

Unfortunately, dye lasers require high maintenance, in no small part because of the limited chemical life of the dye in the solvent. As a result, diode-pumped solid-state lasers (Sec. 16.3A) and laser diodes (Sec. 18.3) have by-and-large replaced the dye laser, except in the most specialized of applications. The diode-pumped Ti^{3+}:sapphire laser, for example, offers broader tunability than a typical dye laser in the vicinity of $\lambda_o = 800$ nm, and requires little maintenance. A frequency-doubled Ti:sapphire laser offers a band of tunable radiation in the vicinity of 400 nm. Tunability near 600 nm, in the gap between 400 and 800 nm, can be achieved by frequency-doubling the output of an optical parametric oscillator operating in the 1–2-μm wavelength region (Sec. 22.2C).

F. X-Ray and Free-Electron Lasers

A number of approaches are commonly used to generate coherent light in the **extreme-ultraviolet** (**EUV**) spectral band, as well as in the **soft-X-ray** (**SXR**) and **hard-X-ray** (**HXR**) bands, which lie on the short-wavelength side of the EUV band. The **γ-ray band** comprises wavelengths yet shorter than those in the HXR band. The wavelengths and photon energies of these bands are presented in Fig. 16.3-7.

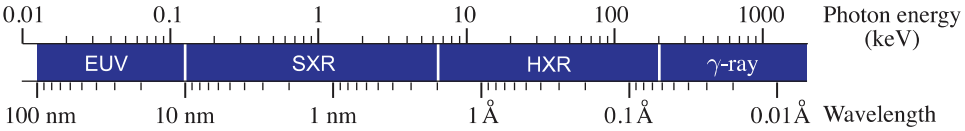


Figure 16.3-7 Wavelengths and photon energies for the extreme ultraviolet (EUV), soft-X-ray (SXR), hard-X-ray (HXR), and γ-ray spectral bands. The boundaries dividing the various bands are somewhat arbitrary. However, for concreteness we set the boundary between the EUV and SXR bands at $\lambda_o = 10$ nm and the boundary between the SXR and HXR bands at $\lambda_o = 0.2$ nm. The Ångström length scale is widely used in the X-ray literature (1 Å $\equiv 10^{-10}$ m $= 0.1$ nm). The photon energy is related to the wavelength via $E(\text{keV}) \approx 1.24/\lambda_o(\text{nm})$, in accordance with (13.1-6).

We begin with a discussion of coherent X-ray generation using **atomic EUV and X-ray lasers**. Two other methods are also widely used for generating coherent X-rays: **X-ray free-electron lasers (XFELs)**, which use particle accelerators to provide pulse energies in the mJ range, as discussed subsequently; and **high-harmonic generation (HHG)**, which provides pulse energies in the μJ range and is examined in Sec. 23.5D. HHG systems are compact and versatile; individual spectral and/or temporal components can be extracted by the use of gating techniques.

Atomic EUV and X-Ray Lasers

Using atomic transitions to achieve laser action in the X-ray region of the electromagnetic spectrum is a challenging enterprise because of the difficulty of achieving a population inversion at short wavelengths. According to (16.1-16), for a fixed value of t_{sp}, the threshold population difference $N_t \propto 1/\tau_p \lambda^2 g(\nu)$. Since a photon of energy $E = hc/\lambda$ is required to pump each atom, the threshold pump power density is proportional to $1/\tau_p \lambda^3 g(\nu)$. When Doppler broadening prevails, as is often the case

in these media, (14.3-42) and (14.3-43) provide that $g(\nu) \propto \lambda$, so that the threshold pump power density is proportional to $1/\tau_p \lambda^4$. Since $1/\lambda^4$ increases rapidly as the wavelength moves toward the X-ray region, it becomes increasingly difficult to supply sufficient pump power to attain laser threshold.

Another challenging aspect of achieving laser action at X-ray wavelengths is centered on the difficulty of constructing suitable optical components. This is the case both because the absorption coefficient is large (which decreases the photon lifetime τ_p and thus further increases N_t) and because the refractive index is close to unity in most materials in the X-ray region. Nevertheless, a number of X-ray optical components have been successfully implemented and are available:

- *Grazing-Incidence Optics.* Since X-ray frequencies lie well above the plasma frequency, metals exhibit a refractive index that lies just below unity (Sec. 8.2A and Fig. 8.2-3). Total internal reflection can therefore be achieved and metals can serve as mirrors. This is possible only near grazing incidence, however, because the small refractive-index contrast requires a large angle of incidence (the situation is analogous to that at the boundary between the core and cladding of a planar dielectric waveguide). Nevertheless, X-rays can be focused to nanometer-scale spots using grazing-incidence deformable mirrors that allow for wavefront correction, as well as by making use of other techniques.
- *Multilayer Optics.* The construction of multilayer-optical devices is more complex than in the visible region because the refractive index is close to unity and does not vary appreciably from one material to another. In accordance with (6.2-15), the normal-incidence Fresnel reflectance at a vacuum/material interface, $\mathcal{R} = [(1-n)/(1+n)]^2$, is typically $< 10^{-4}$. Nevertheless, high-reflectance multilayer mirrors can be fabricated by making use of stacks that contain a large numbers of layers (Sec. 7.1), although this number is ultimately restricted by the limited penetration depth into the stack. As an example, a multilayer mirror comprising tens of alternating layers of Si and Mo can provide normal-incidence reflectance $\approx 70\%$ at $\lambda_o = 13.5$ nm. The optimization of reflectance at certain wavelengths can be fostered by constructing stacks of three materials rather than two, and by relying on biperiodic and aperiodic structures. Other special design criteria are necessary for the design of components that accommodate femtosecond and attosecond X-ray pulses. Multilayer optics can also be incorporated in grazing-incidence configurations and in zone-plate structures.
- *Gratings.* Gratings are useful for selecting narrowband wavelength slices from continuous-wavelength X-ray sources (e.g., synchrotron radiation) as well as for selecting individual harmonic components from line-spectrum X-ray sources such as those generated by high-harmonic generation. They are also useful for pulse shaping and compression (Sec. 23.2). Plane and concave gratings are often ruled with triangular groove profiles.

In spite of the severity of these limitations, they were overcome in 1980 when X-ray laser action was first achieved in a dramatic experiment carried out by researchers at the Lawrence Livermore National Laboratory (LLNL). An underground nuclear detonation, dubbed *Dauphin*, created a flux of incoherent X-rays sufficiently strong that it was able to achieve a population inversion by pumping the atoms in an assembly of metal rods. A coherent X-ray laser pulse was generated and detected — just before the apparatus was vaporized by the detonation. It is worth mentioning that since HHG sources are not lasers, the generation of coherent X-rays via HHG is not subject to the limitations discussed above.

We now examine two practical approaches for implementing atomic EUV and X-ray lasers in the laboratory: 1) atomic ionization using focused laser beams; and 2) atomic inner-shell photopumping. This is followed by a brief discussion of a number of applications that make use of atomic X-ray lasers.

Focused-laser-beam atomic ionization. Highly ionized atoms can be formed by focusing short-duration, high-intensity laser pulses in the IR, visible, or UV onto a solid target to directly generate a hot ionized-atom plasma. Coherent X-ray radiation can then be generated via **recombination** or **collisional excitation**. A downward electron transition in a highly ionized atom produces a high-energy photon that in turn can induce the emission of a clone photon from a nearby ion via stimulated emission. Because it is difficult to construct resonators with X-ray feedback at these short wavelengths, many X-ray lasers operate on the basis of amplified stimulated emission (ASE, Sec. 15.5). An EUV laser of this form that operates at a wavelength of 18.2 nm via plasma recombination was first reported in hydrogen-like carbon (C^{5+}) in the mid-1980s, as detailed in Example 16.3-7.

EXAMPLE 16.3-7. ***Hydrogen-Like Carbon EUV Laser.*** The carbon-plasma laser provides a didactic example of an EUV laser that operates on the basis of focused-laser-beam atomic ionization and recombination. In an experiment carried out at Princeton University in 1985, a 10.6-μm-wavelength CO_2 laser pulse, of 300-J energy and 50-ns duration, with a peak power of 6 GW, was focused onto a solid carbon disk.[†] The infrared-laser pump pulse generated sufficient heat to strip all of the electrons from some of the carbon atoms, thereby creating a plasma of ionized carbon, which was radially confined by the use of a magnetic field. The cooling of the plasma at the termination of the pump pulse led to the capture of electrons in the $n = 3$ shells, and simultaneously to a dearth of electrons in the $n = 2$ shells because of fast radiative decay to the ground state. The net result was a collection of hydrogen-like C^{5+} ions with a population inversion (Fig. 14.1-1).

As expected from (14.1-4), the decay of an electron from the $n = 3$ to the $n = 2$ shell (the $3d \rightarrow 2p$ transition has the largest cross section) is accompanied by the emission of a photon of energy

$$E = \frac{M_r Z^2 e^4}{(4\pi\epsilon_o)^2\, 2\hbar^2}\left(\frac{1}{2^2} - \frac{1}{3^2}\right). \tag{16.3-1}$$

With $Z = 6$ this corresponds to an EUV photon of energy 68 eV at a wavelength $\lambda_o = 18.2$ nm. In the ionized-carbon experiment, a spontaneously emitted photon ($t_{sp} \approx 12$ ps) initiated the stimulated emission of EUV photons from other ions, resulting in ASE (other transition parameters are provided in Table 15.3-1). The single-pass gain-coefficient–length product γd was found to be ≈ 6 so that, in accordance with (15.1-7), the gain was $G \approx e^6$. The output was a 20-ns pulse of EUV ASE with a power of 100 kW, an energy of 2 mJ, and a divergence of ≈ 5 mrad. Similar results were obtained by using Nd^{3+}:glass laser excitation at 1.06 μm.

Typical active media that are used for hot-plasma atomic EUV and X-ray lasers are Ne-like or Ni-like highly ionized atoms, which offer the most stable electron configurations. In 1985, EUV lasers that made use of collisional excitation were reported in Ne-like selenium (Se^{24+}) and in Ne-like yttrium (Y^{29+}), at wavelengths near 21 nm and 16 nm, respectively. Not too long thereafter, in the late 1980s, the 1.06-μm NOVA Nd^{3+}:glass laser system at Lawrence Livermore National Laboratory was used to vaporize thin foils of tantalum and tungsten metal, creating Ni-like Ta^{45+} and Ni-like W^{46+} ions respectively, and generating 250-ps SXR laser pulses at wavelengths as short as $\lambda_o = 4.3$ nm.

A common pumping configuration employs a cylindrical lens that focuses the pump light onto the target in the form of a thin line, generating a column of plasma that serves as a length of active region. Pumping is often provided by a Ti^{3+}:sapphire laser operating near 800 nm, or by the fundamental, second, or third harmonic of a Nd^{3+}:glass laser, at 1053, 526, or 351 nm, respectively. The use of sequential pump pulses can enhance the population inversion, thereby improving efficiency and permitting laser operation in the saturated-gain regime. Delivering the main pump pulse at

[†] See S. Suckewer, C. H. Skinner, H. Milchberg, C. Keane, and D. Voorhees, Amplification of Stimulated Soft X-Ray Emission in a Confined Plasma Column, *Physical Review Letters*, vol. 55, pp. 1753–1756, 1985.

grazing incidence increases the absorption and reduces the required pump energy since it effectively provides traveling-wave pumping, thereby increasing the path length of the light in the plasma gain medium. Furthermore, injecting light from a low-noise, coherent seed laser into the device causes it to behave as an amplifier and to produce *amplified coherent* emission rather than *amplified spontaneous* emission. Desirable characteristics of the seed, such as high coherence, can then be transferred to the X-ray laser, resulting in output pulses with high spatial and temporal coherence, low divergence, and a defined polarization. The salutary features discussed here are all exhibited by the Ni-like Ag^{19+} EUV laser considered in Example 16.3-8.

EXAMPLE 16.3-8. ***Nickel-Like Silver EUV Laser.*** Plasma-based, collisionally excited, atomic EUV lasers have been operated at 13.9 nm on the $4d\,^1S_0 \rightarrow 4p\,^1P_1$ transition in Ni-like Ag^{19+}.

In the *unseeded* configuration, pumping was provided by a mode-locked, amplified, 800-nm Ti^{3+}:sapphire laser that generated ps-duration main heating pulses of ≈ 1 J energy. These pulses were directed to a silver target at grazing incidence to form a 4 mm $\times$ 30 μm line of Ag^{19+} plasma.[†] The small-signal gain coefficient was 67.5 cm^{-1} and the gain-coefficient–length product was $\gamma d \approx 16.8$. The resulting 13.9-nm gain-saturated ASE pulses had duration $\approx$ 5 ps, energy $\approx$ 85 μJ, average power ≈ 2 μW, divergence ≈ 8 mrad, and a repetition rate of 5 Hz.

When *injection-seeded* with pulses from a source of high-harmonic generation (Sec. 23.5D), a similar device generated amplified coherent emission with properties superior to those obtained in the unseeded configuration. The dense Ag^{19+} plasma was created by irradiating a polished silver slab with a main heating pulse of 6.7-ps duration and $\approx$ 1 J energy.[‡] Produced by an amplified Ti^{3+}:sapphire laser operating at 815 nm, the heating pulse impinged on the sample at grazing incidence to form a 4.1 mm $\times$ 30 μm line of plasma. The HHG seed was generated by focusing 20-mJ laser pulses compressed to 50-fs duration, obtained from the Ti^{3+}:sapphire pump laser, into Ne gas at 20 torr (Example 23.5-4). The seed pulse at 13.9 nm, consisting of the 59th harmonic of the HHG, was injected into the amplifier 1 ps after the heating pulse; gain saturation was reached after 3 mm of propagation. The EUV laser amplified the seed pulse by a factor of 200 and generated ps-duration output pulses with an energy of $\approx$ 50 nJ and a divergence $\approx$ 1 mrad. The repetition rate was 5 Hz. The pulse duration can be reduced below 1 ps by making use of collisional-ionization gating.

Other approaches for achieving ionized-atom-plasma lasing include current-pulse and field-induced atomic ionization. The medium can be excited by a strong electrical pulse to create a hot plasma; capillary-confined plasmas can be used to produce saturated lasing in a compact configuration. The active medium can alternatively be directly ionized by laser-driven optical field effects and multiphoton processes, which give rise to a cold, dense collection of ionized atoms surrounded by a hot electron distribution, in which case collisional excitation is initiated by the emitted electrons rather than by the thermalized particles in a heated plasma.

Inner-shell photopumping. The techniques for achieving lasing in the EUV and SXR regions considered above are based on recombination or collisional excitation into excited states of highly ionized atoms in a plasma. An alternative approach, made possible by the advent of the X-ray free-electron laser (XFEL, Fig. 16.3-8), relies on the photopumping of neutral atoms that results in the photoionization of atomic inner-shell electrons. Photopumping leaves the lower laser level unpopulated, and as such results in the direct production of a population inversion. However, since the population

[†] See Y. Wang, M. A. Larotonda, B. M. Luther, D. Alessi, M. Berrill, V. N. Shlyaptsev, and J. J. Rocca, Demonstration of High-Repetition-Rate Tabletop Soft-X-Ray Lasers with Saturated Output at Wavelengths Down to 13.9 nm and Gain Down to 10.9 nm, *Physical Review A*, vol. 72, 053807, 2005.

[‡] See Y. Wang, E. Granados, F. Pedaci, D. Alessi, B. Luther, M. Berrill, and J. J. Rocca, Phase-Coherent, Injection-Seeded, Table-Top Soft-X-Ray Lasers at 18.9 nm and 13.9 nm, *Nature Photonics*, vol. 2, pp. 94–98, 2008.

inversion disappears at the expiration of the lifetime of the core-excited state, the laser action is *self-terminating*. Though this lifetime is short as a result of fast radiative decay from higher energy levels and nonradiative Auger recombination (Fig. 17.1-18), a population inversion can nevertheless be attained by making use of an XFEL, which is an ultrafast, coherent X-ray-laser pump of sufficient power to allow inner-shell photoionization to take place on a timescale comparable with the lifetime of the core-excited state. This pumping scheme bears some similarity to in-band pumping.

Though it requires an XFEL pump, which is hardly found in every laboratory, the inner-shell photopumping approach has a number of merits: 1) the pump energy is directed exclusively to the transition of interest, thereby avoiding inefficiencies associated with the excitation of extraneous states; 2) the laser radiation is spatially well-matched to the pump; and 3) the pump XFEL, when configured for dual-frequency operation, can provide an auxiliary seed.

EXAMPLE 16.3-9. ***XFEL-Pumped Neon Inner-Shell SXR Laser.*** Laser emission at 1.46 nm ($E_{\text{photon}} = 849$ eV) is generated by an inner-shell-photopumped atomic Ne X-ray laser pumped by an XFEL (Example 16.3-11) that operates at a wavelength of 1.29 nm ($E_{\text{photon}} = 960$ eV).[†] Neon is a closed-shell noble gas, with ten electrons ($Z = 10$) and the ground-state electron configuration Ne:$1s^2 2s^2 2p^6$. An XFEL pump photon photoionizes an electron from the inner ($\text{n} = 1$) shell, also known as the K shell. This results in a singly ionized neon ion (Ne^+:$1s^1 2s^2 2p^6$) that serves as the upper laser level. The laser transition Ne^+:$1s^1 2s^2 2p^6 \rightarrow \text{Ne}^+$:$1s^2 2s^2 2p^5$ is associated with the radiative recombination of an electron from the outer ($\text{n} = 2$) shell, also called the L shell, with the *core hole* in the K shell. X-rays resulting from $\text{n} = 2 \rightarrow \text{n} = 1$ transitions are, by convention, called Kα X-rays. Though a competing Auger transition (which has a short 2.4-fs lifetime and thus is exceptionally strong) is available from the upper laser level, Ne^+:$1s^1 2s^2 2p^6 \rightarrow \text{Ne}^{2+}$:$1s^2 2s^2 2p^4$, a residual 1.8% probability of spontaneous radiative decay at 1.46 nm on the laser transition is nevertheless sufficient to achieve lasing. The traditional atomic Ne laser lines in the visible and infrared regions of the spectrum (Fig. 14.1-2) arise from valence-electron transitions lying exclusively within the $\text{n} = 2$ shell.

In an experiment carried out in 2012 at the SLAC National Accelerator Laboratory, XFEL pump pulses (energy ≈ 0.02 to 0.27 mJ, duration ≈ 50 fs, and maximum intensity $\approx 2 \times 10^{17}$ W/cm^2) were focused (spot size ≈ 1 μm) into a gas cell containing a volume of neon atoms (pressure ≈ 500 torr) to create a long, narrow column of transiently core-excited ions. Photons that were spontaneously emitted ($t_{\text{sp}} \approx 130$ fs) near the front end of the column initiated ASE that increased exponentially along the path of the XFEL pulse, which served to prepare atoms in the excited state just as the Ne Kα ASE from previously excited atoms arrived. Some characteristic parameters associated with this transition are provided in Table 15.3-1.

The emitted SXR ASE pulses had energy ≈ 1 μJ, duration ≈ 5 fs, divergence ≈ 1 mrad, and an energy conversion efficiency $\approx 4 \times 10^{-3}$. The single-pass gain coefficient was ≈ 65 cm^{-1} and the gain-coefficient–length product was $\gamma d \approx 18$. Doubling the incident XFEL pulse energy from 0.12 to 0.24 mJ increased the output pulse energy by a factor of 10^4. The laser line exhibited a Lorentzian lineshape with a width of 0.27 eV that resulted principally from lifetime broadening associated with the Auger transitions. The low operating temperature and low Ne gas density rendered the effects of Doppler and collisional broadening negligible; indeed, the observed linewidth was substantially narrower than that attainable with plasma-based X-ray lasers. Globally speaking, the neon inner-shell-photopumped X-ray laser served to convert the fluctuating self-amplified spontaneous emission (SASE) radiation from the pump XFEL into a highly stable, narrowband, coherent X-ray source, a transformation reminiscent of the operation of diode-pumped solid-state (DPSS) lasers (Sec. 16.3A).

EXAMPLE 16.3-10. ***XFEL-Pumped and Seeded Copper Inner-Shell HXR Laser.*** The operation of the XFEL-pumped copper inner-shell laser is similar to that of the XFEL-pumped neon inner-shell laser (Example 16.3-9), but it offers two additional important features: 1) it operates in the HXR rather than in the SXR band; and 2) it can be operated in a seeded configuration by making

---

[†] See N. Rohringer, D. Ryan, R. A. London, M. Purvis, F. Albert, J. Dunn, J. D. Bozek, C. Bostedt, A. Graf, R. Hill, S. P. Hau-Riege, and J. J. Rocca, Atomic Inner-Shell X-Ray Laser at 1.46 Nanometres Pumped by an X-Ray Free-Electron Laser, *Nature*, vol. 481, pp. 488–491, 2012.

use of a dual-frequency XFEL pump. Laser emission on the Cu Kα line at 1.54 Å ($E_{\text{photon}} = 8.0$ keV) from a 20-μm-thick copper foil is obtained by inner-shell photopumping with a HXR XFEL operating at 1.4 Å ($E_{\text{photon}} = 8.9$ keV).[‡] Various parameters associated with this transition are set forth in Table 15.3-1. When an auxiliary spectral component provided by the XFEL was used to seed the laser, the amplification was augmented by a factor of about 100 and the temporal coherence was improved considerably.

Applications of EUV and X-ray lasers. EUV and X-ray laser applications include nanolithography for semiconductor manufacturing, nanopatterning, nanoimaging, plasma diagnostics, medical imaging, high-resolution spectroscopy, nonlinear X-ray optics, and the dynamic imaging and holography of biological samples.

Free-Electron Lasers

The **free-electron laser** (FEL) makes use of an accelerator-generated relativistic electron beam that is passed through a channel between two opposing rows of stationary magnets of alternating polarity known as an **undulator** or **wiggler** (Fig. 16.3-8). The electron beam serves as the pump for the FEL and its interaction with the generated electromagnetic field serves as the active medium. The appellation "free-electron laser" signifies that, unlike the situation in most other lasers, the electrons are not bound in atoms or molecules. Since the motion of the electrons is affected by both the wiggler and the generated field, however, the description "free" is not totally apt.

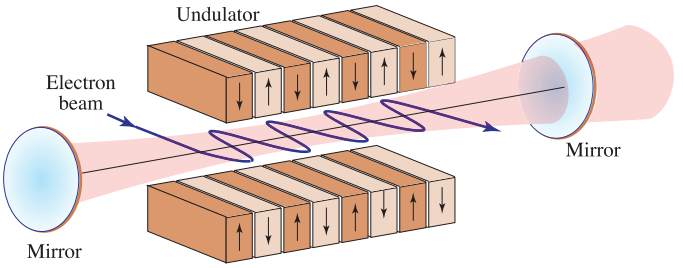


Figure 16.3-8 Schematic of a free-electron laser (FEL) oscillator. The undulator creates a periodic magnetic flux density, typically with a magnitude in the vicinity of 1 T. It usually has a centimeter-scale spatial period and contains anywhere from tens to thousands of periods, so that its total longitudinal length is somewhere between meters and tens of meters. The electron beam, with a radius of millimeters, is guided into the undulator by bending magnets. The electron-beam current typically ranges from amperes to kiloamperes while the electron energy varies from MeV to GeV. The pulse duration of the emitted radiation, which follows that of the electron pulse, varies from femtoseconds to microseconds. Both spontaneous and stimulated emission occur; in wavelength regions where resonators can be fabricated, mirrors can be used to foster oscillation. Such an FEL oscillator might use a mirror separation d about twice the length of the undulator and might support an optical beam with a waist of millimeters. However, FELs can also be operated on the basis of a single pass through an undulator, an approach that is particularly useful in the X-ray region, where it is difficult to fabricate mirrors.

The undulator produces a periodic transverse static magnetic **wiggler field**, which, with proper design of the pole pieces, can be made nearly sinusoidal in space near the axis of the undulator. The wiggler field in turn causes the electrons to undergo nearly sinusoidal transverse oscillations. Relativistic electrons forced to accelerate or decelerate under the action of a magnetic field emit **synchrotron radiation** in a narrow

[‡] See H. Yoneda, Y. Inubushi, K. Nagamine, Y. Michine, H. Ohashi, H. Yumoto, K. Yamauchi, H. Mimura, H. Kitamura, T. Katayama, T. Ishikawa, and M. Yabashi, Atomic Inner-Shell Laser at 1.5-Ångström Wavelength Pumped by an X-Ray Free-Electron Laser, *Nature*, vol. 524, pp. 446–449, 2015.

cone tangential to the electron trajectory in the forward direction. This radiation, while rendered narrowband by the undulator, is incoherent because of constructive and destructive interference among the waves emitted by the randomly distributed electrons.

The lasing process can be initiated by the spontaneous emission at the entrance to the undulator associated with the inherent electron-position fluctuations (electron shot noise). Because the electrons are highly relativistic and move essentially at the speed of light, they copropagate with the radiation they generate and remain coupled to it over long distances, thereby allowing the spontaneous synchrotron radiation to grow. When this radiation becomes sufficiently strong, its transverse field interacts with the transverse component of the electron current. This causes some electrons to lose energy to the field, while others gain energy, via the ponderomotive force associated with the electric field, $\mathcal{F}_\mathrm{p} = -(e^2/4m_0\omega^2)\nabla(\mathcal{E}^2)$, where ω and $\mathcal{E}$ are the angular frequency and amplitude of the oscillating field, respectively. This in turn gives rise to successive microbunches of electrons separated by an optical wavelength along the axis of the undulator, and hence to emissions that are coherent with the phase of the field. The net result is a saturated power many orders of magnitude greater than that of the synchrotron radiation. This route to lasing is known as a **self-amplified spontaneous emission** (**SASE**). Alternatively, the lasing process can be initiated by seeding the FEL with a radiation field in resonance with it, whereupon the FEL acts as an amplifier. The seed may be generated by the FEL itself (self-seeding) or by an external source, such as high-harmonic generation.

The wavelength of the on-axis light emitted by a free-electron laser is expressible as

$$\lambda_\mathrm{FEL} \approx \frac{\Lambda_\mathrm{u}}{2\gamma^2}\left(1+\frac{K^2}{2}\right), \tag{16.3-2}$$

Free-Electron Laser Emission Wavelength

where Λ_u is the spatial period of the undulator and $\gamma = 1/\sqrt{1-v^2/c_o^2} \approx E_\mathrm{beam}/m_0c_o^2$ is the relativistic *Lorentz factor*; the electron rest energy $m_0c_o^2 \approx 0.511$ MeV. The undulator (magnetic-deflection) parameter is expressible as $K = e\mathcal{B}\Lambda_\mathrm{u}/2\pi m_0 c_o$, where $\mathcal{B}$ is the magnetic flux density. The FEL emission wavelength expressed in (16.3-2) is of the form $\lambda_\mathrm{FEL} \propto \Lambda_\mathrm{u}/\gamma^2$. The physical underpinnings of this relation can be understood by recognizing that λ_FEL is matched to a scaled version of the undulator period Λ_u, where the scaling factor $1/\gamma^2$ is imposed by the requirements of relativity. One factor of $1/\gamma$ arises from the length contraction of the undulator spatial period in the rest frame of the relativistic electron, while the other factor of $1/\gamma$ arises from the time dilation of the emitted radiation in the laboratory frame, corresponding to the relativistic Doppler shift.

In accordance with (16.3-2), decreasing the spatial period of the undulator Λ_u results in a decrease of the wavelength of the emitted light λ_FEL, as does decreasing the strength of the magnetic flux density $\mathcal{B}$ embedded in the undulator parameter K. As the energy of the electron beam E_beam increases, so too does the electron velocity v and the Lorentz factor γ, which also lead to a decrease in the wavelength of the emitted radiation. Individual FELs can thus provide a broad range of operating wavelengths by tuning the electron-beam energy E_beam, undulator period Λ_u, and undulator parameter K. Because they operate in a vacuum, high peak powers can be attained without incurring material damage and encountering thermal lensing effects. Moreover, the gain medium is transparent at all wavelengths. Though they are highly complex and require large and expensive facilities, FELs can offer unrivaled power and performance, particularly in wavelength regions that are difficult to reach using other lasers.

Free-electron lasers operate across much of the electromagnetic spectrum, including the millimeter-wave, infrared, visible, ultraviolet, and X-ray bands. Many facilities

have multiple FELs and beamlines. Representative examples of FEL facilities in various spectral regions include:

- *mm-wave/FIR*: The electrostatic FEL at the University of California at Santa Barbara operates at wavelengths that stretch from 2.5 mm to 63 μm.
- *IR/Visible/UV*: The iFEL 1–5 LINAC FEL at Osaka University operates at wavelengths extending from 100 μm to 230 nm.
- *NUV/MUV*: The OK4 storage-ring FEL at Duke University operates at wavelengths that extend from 400 nm to 193 nm.
- *EUV*: The FERMI LINAC FEL at Elettra Sincrotrone Trieste operates at wavelengths between 100 nm and 4 nm.
- *SXR/HXR*: The Linac Coherent Light Source (LCLS) LINAC FEL at the SLAC National Accelerator Laboratory at Stanford University operates at wavelengths that stretch from 44 Å to 1.1 Å.

X-ray free-electron lasers (XFELs). As discussed above, free-electron lasers can operate in the X-ray region by making use of high electron-beam energies and small undulator periods. Because of the difficulty of constructing laser resonators in the X-ray band, XFELs often operate using a single pass through a long undulator and rely on self-amplified spontaneous emission (SASE) or injection seeding; the latter approach typically provides superior performance in terms of output power and temporal coherence. Today's XFELs are typically driven by kilometer-scale linear accelerators (LINACs), which offer more energetic and brighter X-ray beams than those available from devices in which the electrons circulate, such as synchrotrons and electron storage rings. Efforts are underway to reduce the lengths of XFELs from kilometers to meters by coupling tabletop accelerators that make use of **plasma wakefield acceleration (PWFA)** or **laser wakefield acceleration (LWFA)**[†] with undulators and compact electron guns. Though ultrafast SXR pulses are available from high-harmonic generation, as mentioned earlier, XFELs can provide harder, and far more energetic, X-ray pulses.

EXAMPLE 16.3-11. ***Linac Coherent Light Source (LCLS) XFEL.*** The Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory (SLAC) is a 1-km-long SASE free-electron laser that generates ultrafast hard-X-ray pulses at a repetition rate of 120 Hz.[‡] Its (Nd-Fe-B) permanent-magnet undulator, comprising 33 3.4-m-long sections, has an overall active length of 112 m. The undulator period is $\Lambda_u = 3$ cm and the magnetic-deflection parameter is $K = 3.5$, corresponding to a peak magnetic wiggler field of 1.25 T. The LCLS makes use of an electron beam whose energy is adjustable over the range $2.5 \leq E_{\text{beam}} \leq 15.9$ GeV, corresponding to a Lorentz factor $\gamma \approx E_{\text{beam}}/(0.511\ \text{MeV})$ that can be adjusted over the range $4\,892 \leq \gamma \leq 31\,115$. Using (16.3-2), this corresponds to laser operation over the wavelength range $4.4 \geq \lambda_{\text{FEL}} \geq 0.11$ nm ($44 \geq \lambda_{\text{FEL}} \geq 1.1$ Å). The LCLS produces coherent X-ray pulses of energy ≈ 2 mJ, duration between 2 and 500 fs (FWHM), peak power ≈ 20 GW, intensity $\approx 10^{21}$ W/cm^2, and average power ≈ 200 mW. The peak brightness is some 10 orders of magnitude greater than that available using a conventional, incoherent synchrotron source. Multipulse and multicolor lasing, with fixed or variable time and/or wavelength separation, is incorporated in the system, as is polarization control. Self-seeding and external-seeding capabilities provide increased power and improved temporal coherence.

The pulse repetition rate of the LCLS is limited to 120 Hz to avoid damage to the LINAC's copper cavities. High pulse repetition rate greatly facilitates imaging by increasing the number of image

[†] See, e.g., F. Albert, Laser Wakefield Accelerators: Next Generation Light Sources, *Optics & Photonics News*, vol. 29, no. 1, pp. 42–49, 2018.

[‡] See P. Emma, R. Akre, J. Arthur, R. Bionta, C. Bostedt, J. Bozek, A. Brachmann, P. Bucksbaum, R. Coffee, F.-J. Decker, Y. Ding, D. Dowell, S. Edstrom, A. Fisher, J. Frisch, S. Gilevich, J. Hastings, G. Hays, Ph. Hering, Z. Huang, R. Iverson, H. Loos, M. Messerschmidt, A. Miahnahri, S. Moeller, H.-D. Nuhn, G. Pile, D. Ratner, J. Rzepiela, D. Schultz, T. Smith, P. Stefan, H. Tompkins, J. Turner, J. Welch, W. White, J. Wu, G. Yocky, and J. Galayda, First Lasing and Operation of an Ångstrom-Wavelength Free-Electron Laser, *Nature Photonics*, vol. 4, pp. 641–647, 2010.

samples that can be collected in a fixed period of time. The European XFEL at Deutsches Elektronen-Synchrotron (DESY) Hamburg, commissioned in 2017, operates at an average repetition rate of 27 kHz by virtue of its superconducting linear accelerator. This FEL generates HXR radiation in the form of clusters of 2 700 pulses separated by 220 ns, periodically repeated at a rate of 10 Hz. Because of its high pulse repetition rate, this 3.4-km-long, 17.5-GeV, SASE FEL has an average power and brightness nearly 1000 times greater than that of the LCLS. In 2021, a new superconducting-LINAC FEL, dubbed LCLS-II, will come online at the SLAC National Accelerator Laboratory. Among other advances, it will support a pulse repetition rate of 1 MHz.

XFEL applications. The XFEL serves as a source of energetic, ultrafast, and coherent X-rays. Because the short-wavelength pulses emitted are brief and bright, XFELs have facilitated the imaging and filming of physical, chemical, and biological structures and processes at unprecedented spatial and temporal scales. The XFEL has also expedited the study of non-periodic systems, non-crystalline states, non-equilibrium dynamical processes, and nonlinear X-ray phenomena. With spatial resolution at the Å-scale and temporal resolution at the fs-scale, XFEL imaging has elucidated the inner workings of processes in photonics, materials science, and medicine.

A pulsed radiation source suitable for high-resolution, spatiotemporal imaging has a number of requirements that are accommodated by XFELs:

- *Temporal and spatial coherence for coherent imaging.* Spatial coherence is accommodated for SASE operation and temporal coherence is accommodated for seeded operation.
- *Temporal resolution at the atomic scale (Bohr period $T_0 \approx 150$ as).* Pulse durations ≈ 1 fs are currently accommodated; the generation of attosecond pulses is under development.
- *High energy and high peak power.* Pulses with mJ energies and peak powers of tens of GW are currently accommodated; beams at photon energies > 50 keV are being planned.
- *High average power for rapid image accumulation.* Pulse repetition rates of tens of kHz are currently accommodated; rates up to 1 MHz are under development.
- *Spatial resolution at the atomic scale (Bohr radius $a_0 \approx 0.53$ Å).* Focusing to nm-scale spot sizes is currently accommodated.
- *High intensity to facilitate extreme nonlinear X-ray interactions.* Pulse intensities as high as $\approx 10^{21}$ W/cm^2 are currently accommodated.
- *Multicolor and multipulse operation.* Pulses with fixed or variable wavelength and/or time separation are accommodated.

XFELs offer unrivaled spatiotemporal resolution in a number of imaging configurations and have ushered in new imaging methodologies such as **serial femtosecond X-ray coherent diffractive imaging**. A single, ultrafast, high-energy, XFEL pulse forms a diffractive image of a target particle before photoionizing it, whereupon the particle vaporizes via a "Coulomb explosion." This approach is referred to as *diffraction-before-destruction*. Many such individual diffractive patterns are recorded from individual particles that are serially injected into the XFEL beam via a jet. Patterns with the same orientation are combined to form a full set of 3D diffraction data, from which the image is extracted by use of phase retrieval. This technique is suitable for a broad variety of targets, including nanocrystals, proteins, and viruses.

Valuable information about the internal workings of molecules can also be obtained by **optical/X-ray pump–probe experiments**. Ultrafast optical-laser pulses are used to pump a target to manipulate its internal electronic state and XFEL pulses are used to probe it after an adjustable delay time. The collected data is fashioned into a movie that tracks femtosecond-scale changes to the electronic states and molecular structure. The reorganization of electron clouds in the course of making and breaking molecular bonds has been visualized.

G. Tabulation of Selected Laser Characteristics

Table 16.3-1 lists, in order of increasing wavelength, several characteristics for various lasers. The broad range of transition wavelengths, wall-plug efficiencies, and output powers is noteworthy. The transition cross section, spontaneous lifetime, and transition linewidth for a number of these lasers are provided in Table 15.3-1. The laser linewidth $\Delta\nu_{\mathrm{L}}$ is generally orders of magnitude smaller than the transition linewidth $\Delta\nu$ because of the additional frequency selectivity imposed by the optical resonator.

Table 16.3-1 Parameters for some well-known laser media, in order of increasing wavelength.

Laser Medium[a]	Transition Wavelength λ_o	Single Mode (S) or Multimode (M)	CW or Pulsed[b]	Approx. Wall-Plug Efficiency $\eta_c(\%)$[c]	Max. Output Power or Energy[d]	Energy-Level Diagram (Fig.)
X-ray free-electron laser (LCLS)	1.1–44 Å	M	Pulsed	10^{-6}	2 mJ	
Ne^{+} Kα (g)	14.6 Å	S	Pulsed	10^{-9}	1 μJ	
Ag^{19+} (p)	13.9 nm	M	Pulsed	10^{-4}	85 μJ	
C^{5+} (p)	18.2 nm	M	Pulsed	10^{-3}	2 mJ	14.1-1
ArF Exciplex (g)	193 nm	M	Pulsed	0.2	1 J	
KrF Exciplex (g)	248 nm	M	Pulsed	0.4	1.5 J	
Ar^{+} (g)	515 nm	S/M	CW	0.05	15 W	
Rhodamine-6G (l)	560–640 nm	S/M	CW	10	100 mW	14.1-8
Ne (He–Ne) (g)	633 nm	S/M	CW	0.1	50 mW	14.1-2
Kr^{+} (g)	647 nm	S/M	CW	0.01	1 W	
Cr^{3+}:Al_2O_3 (ruby) (s)	694 nm	M	CW	4	1 W	15.3-1
Cr^{3+}:$BeAl_2O_4$ (alexandrite) (s)	700–820 nm	M	CW	40	25 W	14.1-4
Ti^{3+}:Al_2O_3 (s)	700–1050 nm	S/M	CW	0.1	5 W	16.3-3
Yb^{3+}:YAG (thin-disk) (s)	1030 nm	S/M	CW	30.	1 kW	16.3-2
Nd^{3+}:Glass (phosphate) (s)	1053 nm	S/M	Pulsed	1.	50 J	15.3-3
Nd^{3+}:YAG (s)	1064 nm	S/M	CW	15.	200 W	14.1-5
Nd^{3+}:YVO_4 (s)	1064 nm	S/M	CW	30.	30 W	16.3-1
Yb^{3+}:Silica fiber (s)	1070 nm	S/M	CW	40.	10 kW	
Cr^{4+}:Mg_2SiO_4 (forsterite) (s)	1100–1400 nm	M	CW	0.5	1 W	
Er^{3+}:Silica fiber (s)	1550 nm	S/M	CW	20.	2 kW	15.3-6
Tm^{3+}:Silica fiber (s)	1800–2100 nm	S/M	CW	35.	500 W	
Cr^{2+}:ZnS (s)	1900–3000 nm	S/M	CW	25.	100 W	
Ne (He–Ne) (g)	3.39 μm	S/M	CW	0.1	20 mW	14.1-2
CO_2 (g)	10.6 μm	S/M	CW	15.	500 W	14.1-7
H_2O (g)	28 μm	S/M	CW	0.05	250 mW	
CH_3OH (methanol) (g)	118.8 μm	S/M	CW	0.03	150 mW	
HCN (hydrogen cyanide) (g)	336.8 μm	S/M	CW	0.05	120 mW	

[a] Gas (g), solid (s), liquid (l), plasma (p).

[b] Lasers designated "CW" can be operated in pulsed mode as well. Lasers that cannot sustain a continuous population inversion can operate only in pulsed mode and are denoted "Pulsed."

[c] The wall-plug efficiency η_c (also known as the power-conversion efficiency or overall efficiency) is the ratio of the output optical power P_o to the input electrical power P_e (for pulsed lasers, it is the ratio of the energies per pulse). Values reported have substantial uncertainty since they sometimes include the electrical power consumed for overhead functions such as cooling and monitoring. Laser diodes exhibit the highest efficiencies, which can exceed 70%, as discussed in Sec. 18.4.

[d] The maximum output power for CW systems varies over a substantial range, as does the maximum output energy per pulse for pulsed systems (in part because of the wide range of pulse durations). Representative values are listed. Achievable values for single-mode operation with no amplification are provided where available; multimode output powers are generally significantly higher. Output powers delivered by laser systems used for industrial applications can be orders of magnitude higher.

16.4 PULSED LASERS

It is sometimes desirable to operate lasers in a pulsed mode since the optical power can be greatly increased when the output pulse has limited duration. Lasers can be made to emit optical pulses with durations as short as femtoseconds; the durations can often be further compressed by making use of nonlinear optical techniques (Sec. 23.2). Pulse repetition rates (PRRs) can extend to the THz range and individual laser pulses can carry enormous peak powers and intensities. A 10-fs-duration pulse of 10-mJ energy, for example, exhibits a peak power of 1 TW. Focusing such a pulse to a 3-μm-radius spot provides a peak intensity of 7 EW/cm^2. Some lasers must be operated in a pulsed mode since CW operation cannot be sustained, as is evident in Table 16.3-1.

A. Methods of Pulsing Lasers

The most direct method of obtaining pulsed light from a laser is to use a continuous-wave (CW) laser in conjunction with an external modulator (switch) that transmits the light only during selected short time intervals. This simple method has two distinct disadvantages, however. First, the scheme is inefficient since it blocks (and therefore wastes) the light energy during the off-time of the pulse train. Second, the peak power of the pulses cannot exceed the steady power of the CW source, as illustrated in Fig. 16.4-1(a).

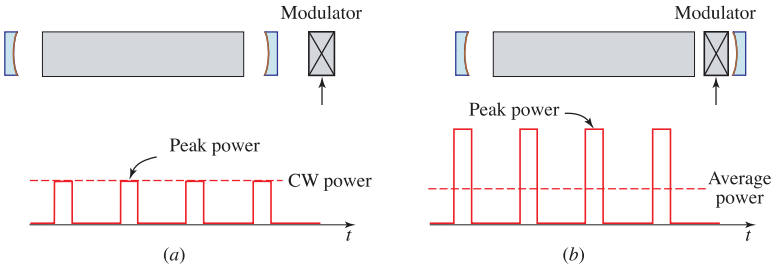


Figure 16.4-1 Comparison of attainable peak laser output powers using (a) an external modulator and (b) an internal modulator.

More efficient pulsing schemes are based on turning the laser itself on and off by means of an internal modulator, designed so that energy is stored during the off-time and released during the on-time. Energy may be stored in the resonator in the form of light that is periodically permitted to escape, or in the atomic system in the form of a population inversion that is periodically released by allowing the system to oscillate. These schemes permit short laser pulses to be generated with peak powers far in excess of the constant power deliverable by a CW laser, as illustrated in Fig. 16.4-1(b).

Four common methods are used for the internal modulation of laser light: 1) gain switching, 2) Q-switching, 3) cavity dumping, and 4) mode locking. These are considered in turn.

Gain Switching

Gain switching is a rather direct approach in which the gain is controlled by turning the laser pump on and off (Fig. 16.4-2). In a flashlamp-pumped pulsed ruby laser, for example, the pump (flashlamp) is switched on periodically for brief periods of time by a sequence of electrical pulses. During the on-times, the gain coefficient exceeds the loss coefficient and laser light is produced. Pulsed laser diodes are generally gain

switched because it is easy to modulate the electric current that provides the pumping (Sec. 18.3). The laser-pulse rise and fall times achievable with gain switching are derived in Sec. 16.4B.

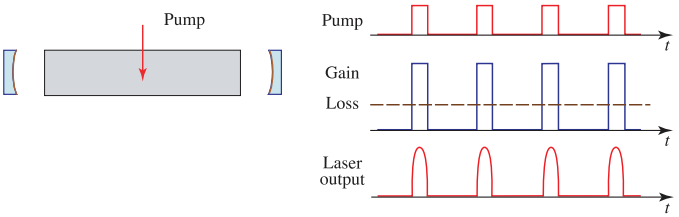


Figure 16.4-2 Gain switching. The laser pump is switched on and off periodically.

Q-Switching

In Q**-switching**, the laser output is turned off by periodically increasing the resonator loss (spoiling the resonator quality factor Q) with the help of a modulated absorber inside the resonator (Fig. 16.4-3). Thus, Q-switching is loss switching. Because the pump continues to deliver constant power at all times, energy is stored in the atoms in the form of an accumulated population difference during the high-loss off-times. When the losses are reduced during the on-times, the large accumulated population difference is released, generating intense (usually short) pulses of light. An analysis of this method is provided in Sec. 16.4C.

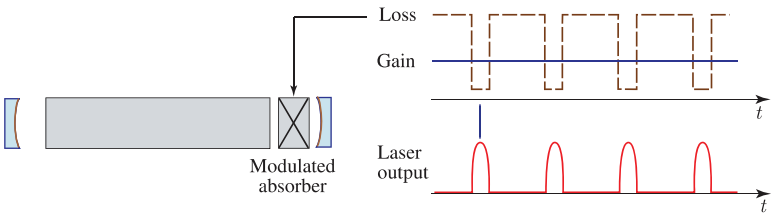


Figure 16.4-3 Q-switching. During the off-times, energy builds up as an accumulated population inversion in the active medium. The resonator loss is modulated by use of an absorber.

Cavity Dumping

Cavity dumping is a technique based on storing photons (rather than a population difference) in the resonator during the off-times, and releasing them during the on-times. It differs from Q-switching in that the resonator loss is modulated by altering the mirror transmittance (Fig. 16.4-4). The system operates like a bucket into which water is poured from a hose at a constant rate. The bucket represents the resonator, the water hose represents the constant pump, and the bucket bottom represents the laser output mirror. After a period of time of accumulating water, the bottom of the bucket is suddenly removed so that the water is "dumped." The bucket bottom is subsequently returned and the process is repeated. A constant flow of water is therefore converted into a pulsed flow. The leakage of light from the resonator, including useful light, is not permitted during the off-times; photons are stored in the resonator and cannot escape. This results in negligible resonator loss, thereby increasing the optical power inside the laser resonator. The mirror is suddenly removed altogether (e.g., by rotating it out of alignment), increasing its transmittance to 100% during the on-times. The result is a strong pulse of laser light. As the accumulated photons leave the resonator, the sudden

increase in the loss arrests the oscillation. A detailed analysis for cavity dumping is not provided in the sequel inasmuch as it is closely related to that of Q-switching. This is because the variation of the gain and loss with time are similar, as may be seen by comparing Fig. 16.4-4 with Fig. 16.4-3.

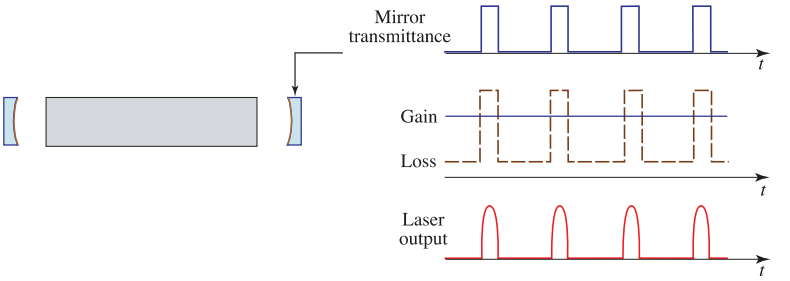


Figure 16.4-4 Cavity dumping. During the off-times, energy builds up as an increase of the photon-number density in the resonator. One of the mirrors is periodically removed (e.g., by rotating it out of alignment) to dump the stored photons as useful light.

Mode Locking

The pulse-generation approaches discussed above are based on the transient dynamics of a laser medium. **Mode locking** differs from these approaches in that it is a dynamic steady-state process. It is the most important of all the techniques for generating trains of ultrashort laser pulses. Pulsed laser action is attained by locking the phases of the modes of a laser to each other. An example is provided by the longitudinal modes of a multimode laser, which oscillate at frequencies that are equally spaced by the intermodal frequency $\nu_F = c/2d$. When the phases of these components are locked together, they behave like the Fourier components of a periodic function of time, and therefore form a periodic pulse train. The mode coupling is achieved by periodically modulating the losses inside the resonator. Mode locking is examined in Sec. 16.4D.

*B. Analysis of Transient Effects

An analytical description of the operation of pulsed lasers requires an understanding of the dynamics of the laser oscillation process, i.e., the time course of laser oscillation onset and termination. The steady-state solutions presented earlier in this chapter are inadequate for this purpose. The lasing process is governed by two variables: the number of photons per unit volume in the resonator, $\mathfrak{n}(t)$, and the atomic population difference per unit volume, $N(t) = N_2(t) - N_1(t)$; both are functions of the time t.

Rate Equation for the Photon-Number Density

The photon-number density $\mathfrak{n}$ is governed by the rate equation

$$\frac{d\mathfrak{n}}{dt} = -\frac{\mathfrak{n}}{\tau_p} + NW_i\,. \tag{16.4-1}$$

The first term on the right-hand side represents photon loss caused by leakage from the resonator, at a rate given by the inverse photon lifetime $1/\tau_p$. The second term represents net photon gain, at a rate NW_i, arising from stimulated emission and absorption. Here, the quantity $W_i = \phi\,\sigma(\nu) = c\,\mathfrak{n}\,\sigma(\nu)$ is the probability density for induced absorption/emission. Spontaneous emission is assumed to be negligible. With the help

of the relation $N_t = \alpha_r/\sigma(\nu) = 1/c\tau_p\,\sigma(\nu)$, where N_t is the threshold population difference [see (16.1-15)], we write $\sigma(\nu) = 1/c\tau_p\,N_t$, from which

$$W_i = \frac{\mathfrak{n}}{N_t\,\tau_p}. \tag{16.4-2}$$

Substituting this into (16.4-1) provides a simple differential equation for the photon number density $\mathfrak{n}$,

$$\boxed{\frac{d\mathfrak{n}}{dt} = -\frac{\mathfrak{n}}{\tau_p} + \frac{N}{N_t}\frac{\mathfrak{n}}{\tau_p}.} \tag{16.4-3}$$

(16.4-3) Photon-Number Rate Equation

As long as $N > N_t$, $d\mathfrak{n}/dt$ will be positive and $\mathfrak{n}$ will increase. When steady state ($d\mathfrak{n}/dt = 0$) is reached, $N = N_t$.

Rate Equation for the Population Difference

The dynamics of the population difference $N(t)$ depends on the pumping configuration. We proceed to analyze a three-level pumping scheme (Sec. 15.2B). The rate equation for the population of the upper energy level of the transition is, according to (15.2-8),

$$\frac{dN_2}{dt} = R - \frac{N_2}{t_{sp}} - W_i(N_2 - N_1), \tag{16.4-4}$$

where it is assumed that $\tau_2 = t_{sp}$. The pumping rate R is taken to be independent of the population difference N. Denoting the total atomic number density $N_2 + N_1$ by N_a, so that $N_1 = \frac{1}{2}(N_a - N)$ and $N_2 = \frac{1}{2}(N_a + N)$, we obtain a differential equation for the population difference $N = N_2 - N_1$:

$$\frac{dN}{dt} = \frac{N_0}{t_{sp}} - \frac{N}{t_{sp}} - 2W_i N, \tag{16.4-5}$$

where the small-signal population difference $N_0 = 2Rt_{sp} - N_a$ [see (15.2-27)]. Substituting the relation $W_i = \mathfrak{n}/N_t\tau_p$ from (16.4-2) into (16.4-5) then yields

$$\boxed{\frac{dN}{dt} = \frac{N_0}{t_{sp}} - \frac{N}{t_{sp}} - 2\frac{N}{N_t}\frac{\mathfrak{n}}{\tau_p}.} \tag{16.4-6}$$

(16.4-6) Population-Difference Rate Equation (Three-Level System)

The third term on the right-hand side of (16.4-6) is twice the second term on the right-hand side of (16.4-3), and of opposite sign. This reflects the fact that the generation of one photon by an induced transition reduces the population of level ② by one atom while increasing the population of level ① by one atom, thereby decreasing the population difference by two atoms.

Equations (16.4-3) and (16.4-6) are a pair of coupled nonlinear differential equations whose solution determines the transient behaviors of the photon number density $\mathfrak{n}(t)$ and the population difference $N(t)$. Setting $d\mathfrak{n}/dt = 0$ and $dN/dt = 0$ leads to $N = N_t$ and $\mathfrak{n} = (N_0 - N_t)(\tau_p/2t_{sp})$, respectively. These are indeed the steady-state values of N and $\mathfrak{n}$ obtained previously, as is evident from (16.2-12) with $\tau_s = 2t_{sp}$, as provided by (15.2-28) for a three-level pumping scheme.

EXERCISE 16.4-1

Population-Difference Rate Equation for a Four-Level System. Obtain the population-difference rate equation for a four-level system for which $\tau_1 \ll t_{sp}$. Provide a rationale for the absence of the factor of 2 that appears on the right-hand side of (16.4-6).

Gain Switching

Gain switching is implemented by turning the pumping rate R on and off, which is equivalent to modulating the small-signal population difference $N_0 = 2Rt_{sp} - N_a$, as provided in (15.2-27). A schematic illustration of the typical time evolution of the population difference $N(t)$ and the photon-number density $n(t)$, as the laser is pulsed by varying N_0, is provided in Fig. 16.4-5.

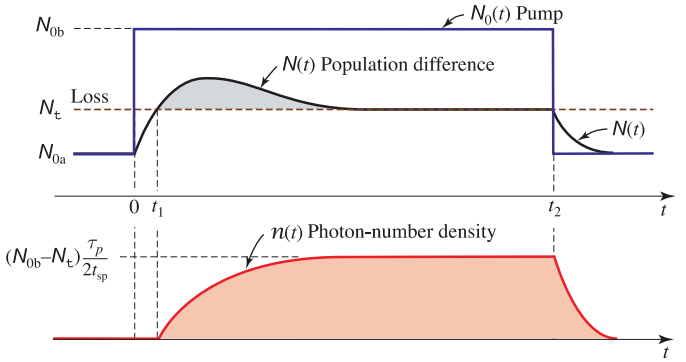


Figure 16.4-5 Variation of the population difference $N(t)$ and the photon-number density $n(t)$ with time as a square pump pulse causes N_0 to suddenly increase from a low value N_{0a} to a high value N_{0b} at $t = 0$, and then to return back to the low value N_{0a} at $t = t_2$.

The following regimes are evident in the process:

- For $t < 0$, the population difference $N(t) = N_{0a}$ lies below the threshold N_t and oscillation does not occur.
- The pump is turned on at $t = 0$, which increases N_0 from a value N_{0a} below threshold to a value N_{0b} above threshold in step-function fashion. The population difference $N(t)$ begins to increase as a result. As long as $N(t) < N_t$, however, the photon-number density $n = 0$. In this region (16.4-6) therefore becomes $dN/dt = (N_0 - N)/t_{sp}$, indicating that $N(t)$ grows exponentially toward its equilibrium value N_{0b} with time constant t_{sp}.
- Once $N(t)$ crosses the threshold N_t at $t = t_1$, laser oscillation begins and $n(t)$ increases. The population inversion then begins to deplete so that the rate of increase of $N(t)$ slows. As $n(t)$ becomes larger, the depletion becomes more effective so that $N(t)$ begins to decay toward N_t. $N(t)$ finally reaches N_t, at which time $n(t)$ reaches its steady-state value.
- The pump is turned off at time $t = t_2$, so that $N_0(t)$ returns to its initial value N_{0a}. $N(t)$ and $n(t)$ proceed to decay to the values N_{0a} and 0, respectively.

The detailed profile of the buildup and decay of $n(t)$ is obtained by numerically solving (16.4-3) and (16.4-6). The precise shape of the solution depends on the values of t_{sp}, τ_p, and N_t, as well as on N_{0a} and N_{0b} (see Prob. 16.4-4).

*C. *Q*-Switching

Q-switched laser pulsing is achieved by switching the resonator loss coefficient α_r from a large value during the off-time to a small value during the on-time. This may be accomplished in any number of ways, such as by placing a modulator into the resonator that periodically introduces large losses. Since the lasing threshold population difference N_t is proportional to the resonator loss coefficient α_r [see (16.1-14) and (16.1-6)], the result of switching α_r is to decrease N_t from a high value N_{ta} to a low value N_{tb}, as illustrated in Fig. 16.4-6. Therefore, in Q-switching N_t is modulated while N_0 remains fixed, whereas in gain switching N_0 is modulated while N_t remains fixed (see Fig. 16.4-5).

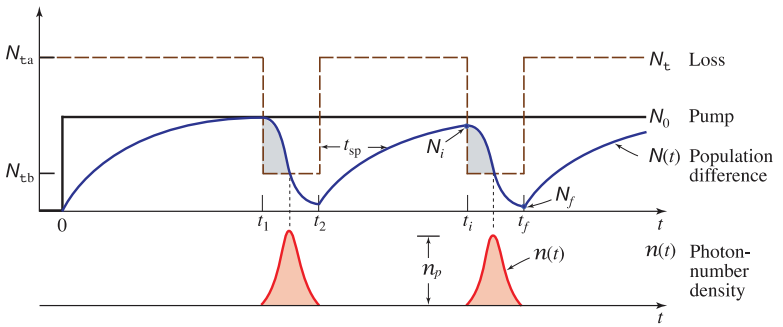


Figure 16.4-6 Operation of a Q-switched laser. Behavior of the threshold population difference N_t (which is proportional to the resonator loss coefficient α_r), the pump parameter N_0, the population difference $N(t)$, and the photon number density $n(t)$.

The population and photon-number densities behave as follows:

- At $t = 0$, the pump is turned on so that N_0 follows a step function. The loss is maintained at a level that is sufficiently high ($N_t = N_{ta} > N_0$) so that laser oscillation cannot begin. The population difference $N(t)$ therefore builds up (with time constant t_{sp}). The medium behaves as a high-gain amplifier in this region but the loss is sufficiently large that oscillation is prevented.
- At $t = t_1$, the loss is suddenly decreased so that N_t diminishes to a value $N_{tb} < N_0$. In accordance with (16.1-13), oscillation therefore begins and the photon-number density rises sharply. The presence of the radiation causes a depletion of the population inversion (gain saturation) whereupon $N(t)$ begins to decrease. When $N(t)$ falls below N_{tb}, the loss again exceeds the gain, resulting in a rapid decrease of the photon-number density (with a time constant of the order of the photon lifetime τ_p).
- At $t = t_2$, the loss is reinstated, ensuring the availability of a long period of population-inversion buildup to prepare for the next pulse. The process is repeated periodically so that a periodic optical pulse train is generated.

We now undertake an analysis to determine the peak power, energy, duration, and shape of the optical pulse generated by a Q-switched laser in the steady pulsed state. We rely on the basic three-level-system rate equations (16.4-3) and (16.4-6) for $n(t)$ and $N(t)$, respectively, which we solve during the on-time t_i to t_f indicated in Fig. 16.4-6. The problem can, of course, be solved numerically. However, the equations simplify sufficiently to permit an analytical solution if we assume that the first two terms on the right-hand side of (16.4-6) are negligible. This assumption is suitable if both the pumping and the spontaneous emission are negligible in comparison with the effects of induced transitions during the short time interval from t_i to t_f. This

approximation turns out to be reasonable if the duration of the generated optical pulse is much shorter than $t_{\rm sp}$. When this is the case, (16.4-3) and (16.4-6) become, respectively,

$$\boxed{\frac{d\mathfrak{n}}{dt} = \left(\frac{N}{N_{\rm t}} - 1\right)\frac{\mathfrak{n}}{\tau_p}} \tag{16.4-7}$$

$$\boxed{\frac{dN}{dt} = -2\,\frac{N}{N_{\rm t}}\,\frac{\mathfrak{n}}{\tau_p}\,.} \tag{16.4-8}$$

These are two coupled differential equations in $\mathfrak{n}(t)$ and $N(t)$, with initial conditions $\mathfrak{n} = 0$ and $N = N_i$ at $t = t_i$. During the time interval from t_i to t_f, $N_{\rm t}$ is fixed at its low value $N_{\rm tb}$.

Dividing (16.4-7) by (16.4-8), we obtain a single differential equation relating $\mathfrak{n}$ and N,

$$\frac{d\mathfrak{n}}{dN} \approx \frac{1}{2}\left(\frac{N_{\rm t}}{N} - 1\right), \tag{16.4-9}$$

which we integrate to obtain

$$\mathfrak{n} \approx \tfrac{1}{2}N_{\rm t}\ln(N) - \tfrac{1}{2}N + \text{constant}. \tag{16.4-10}$$

Using the initial condition $\mathfrak{n} = 0$ when $N = N_i$ finally leads to

$$\boxed{\mathfrak{n} \approx \tfrac{1}{2}N_{\rm t}\ln\frac{N}{N_i} - \tfrac{1}{2}(N - N_i)\,.} \tag{16.4-11}$$

Pulse Power

According to (16.2-10) and (16.2-3), the internal photon-flux density (comprising both directions) is given by $\phi = \mathfrak{n}c$, while the external photon-flux density emerging from the output mirror of transmittance $\mathcal{T}$ is $\phi_o = \frac{1}{2}\mathcal{T}\mathfrak{n}c$. Assuming that the photon-flux density is uniform over the cross-sectional area A of the emerging beam, the corresponding optical output power is

$$P_o = h\nu\,A\,\phi_o = \tfrac{1}{2}h\nu\,c\,\mathcal{T}A\mathfrak{n} = h\nu\,\mathcal{T}\frac{c}{2d}\,V\mathfrak{n}\,, \tag{16.4-12}$$

where $V = Ad$ is the volume of the resonator. According to (16.2-17), if $\mathcal{T} \ll 1$ the fraction of the resonator loss that contributes to useful light at the output is $\eta_e \approx \mathcal{T}(c/2d)\tau_p$, where η_e is the extraction efficiency, which leads to

$$P_o = \eta_e\,h\nu\,\frac{\mathfrak{n}V}{\tau_p}\,. \tag{16.4-13}$$

Equation (16.4-13) is easily interpreted since the factor $\mathfrak{n}V/\tau_p$ is the number of photons lost from the resonator per unit time.

Peak Pulse Power

As discussed earlier, and illustrated in Fig. 16.4-6, $\mathfrak{n}$ reaches its peak value $\mathfrak{n}_p$ when $N = N_\mathrm{t} = N_\mathrm{tb}$. This is corroborated by setting $d\mathfrak{n}/dt = 0$ in (16.4-7), which leads immediately to $N = N_\mathrm{t}$. Substituting this into (16.4-11) therefore provides

$$\mathfrak{n}_p = \tfrac{1}{2} N_i \left(1 + \frac{N_\mathrm{t}}{N_i} \ln \frac{N_\mathrm{t}}{N_i} - \frac{N_\mathrm{t}}{N_i} \right). \tag{16.4-14}$$

Using this result in conjunction with (16.4-12) provides the peak power

$$P_p = h\nu \, \mathcal{T} \frac{c}{2d} \, V \mathfrak{n}_p \, . \tag{16.4-15}$$

When $N_i \gg N_\mathrm{t}$, as must be the case for pulses with large peak powers, we have $N_\mathrm{t}/N_i \ll 1$ so that (16.4-14) provides

$$\mathfrak{n}_p \approx \tfrac{1}{2} N_i \, . \tag{16.4-16}$$

In this limit, the peak photon-number density is equal to one-half of the initial population density difference whereupon the peak power assumes a particularly simple form:

$$\boxed{P_p \approx \tfrac{1}{2} h\nu \, \mathcal{T} \frac{c}{2d} \, V N_i \, .} \tag{16.4-17}$$

Peak Pulse Power

Pulse Energy

The pulse energy E is given by

$$E = \int_{t_i}^{t_f} P_o \, dt, \tag{16.4-18}$$

which, in accordance with (16.4-12), can be written as

$$E = h\nu \, \mathcal{T} \frac{c}{2d} \, V \int_{t_i}^{t_f} \mathfrak{n}(t) \, dt = h\nu \, \mathcal{T} \frac{c}{2d} \, V \int_{N_i}^{N_f} \mathfrak{n}(t) \frac{dt}{dN} \, dN. \tag{16.4-19}$$

Inserting (16.4-8) in (16.4-19), we obtain

$$E = \tfrac{1}{2} h\nu \, \mathcal{T} \frac{c}{2d} \, V N_\mathrm{t} \, \tau_p \int_{N_f}^{N_i} \frac{dN}{N} \, , \tag{16.4-20}$$

which integrates to

$$E = \tfrac{1}{2} h\nu \, \mathcal{T} \frac{c}{2d} \, V N_\mathrm{t} \, \tau_p \ln \frac{N_i}{N_f} \, . \tag{16.4-21}$$

The final population difference N_f is determined by setting $\mathfrak{n} = 0$ and $N = N_f$ in (16.4-11), which provides

$$\ln \frac{N_i}{N_f} = \frac{N_i - N_f}{N_\mathrm{t}} \, . \tag{16.4-22}$$

Finally, substitution of (16.4-22) into (16.4-21) yields

$$E = \tfrac{1}{2} h\nu\, \mathcal{T} \frac{c}{2d}\, V \tau_p \,(N_i - N_f)\,. \tag{16.4-23}$$

Q-Switched Pulse Energy

When $N_i \gg N_f$, we obtain $E \approx \frac{1}{2} h\nu\, \mathcal{T}(c/2d) V \tau_p\, N_i$, as expected. It remains to solve (16.4-22) for N_f. One approach is to let $X = N_i/N_\mathrm{t}$ and $Y = N_f/N_\mathrm{t}$, whereupon (16.4-22) becomes $\ln(X/Y) = X - Y$ or, equivalently, $\ln Y - Y = \ln X - X$. Exponentiating both sides of this equation gives $\exp(\ln Y - Y) = \exp(\ln X - X)$, which in turn provides $Y \exp(-Y) = X \exp(-X)$. Thus, given $X = N_i/N_\mathrm{t}$, we can solve for $Y = N_f/N_\mathrm{t}$ numerically or by using the graph provided in Fig. 16.4-7.

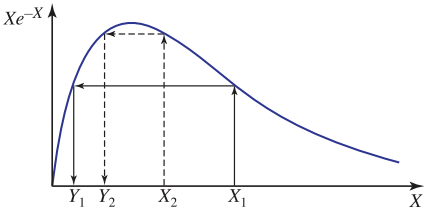


Figure 16.4-7 Graphical construction for determining N_f from N_i, where $X = N_i/N_\mathrm{t}$ and $Y = N_f/N_\mathrm{t}$. For $X = X_1$ the ordinate represents the value $X_1 \exp(-X_1)$. Since the corresponding solution Y_1 obeys $Y_1 \exp(-Y_1) = X_1 \exp(-X_1)$, it must have the same value of the ordinate.

Pulse Duration

A rough estimate of the pulse duration is provided by the ratio of the pulse energy to the peak pulse power. Using (16.4-14), (16.4-15), and (16.4-23), we obtain

$$\tau_\mathrm{pulse} = \tau_p \frac{N_i/N_\mathrm{t} - N_f/N_\mathrm{t}}{N_i/N_\mathrm{t} - \ln(N_i/N_\mathrm{t}) - 1}\,. \tag{16.4-24}$$

Pulse Duration

When $N_i \gg N_\mathrm{t}$ and $N_i \gg N_f$, the pulse duration reduces to $\tau_\mathrm{pulse} \approx \tau_p$.

Pulse Shape

The optical pulse shape, along with the various pulse characteristics described above, can be determined by numerically integrating (16.4-7) and (16.4-8). Examples of the resulting pulse shapes are displayed in Fig. 16.4-8.

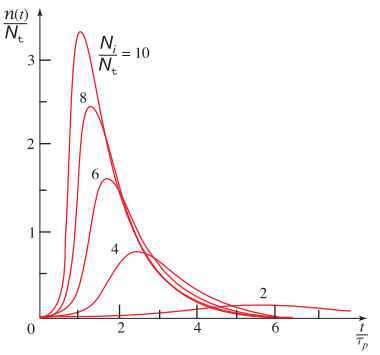


Figure 16.4-8 Q-switched pulse shapes obtained by numerically integrating the approximate rate equations provided in (16.4-7) and (16.4-8). The photon-number density $\mathfrak{n}(t)$ is normalized to the threshold population difference $N_\mathrm{t} = N_\mathrm{tb}$ and the time t is normalized to the photon lifetime τ_p. As the ratio N_i/N_t increases, the pulse narrows and attains a higher peak value. In the limit $N_i/N_\mathrm{t} \gg 1$, the peak value of $\mathfrak{n}(t)$ approaches $\frac{1}{2} N_i$.

EXAMPLE 16.4-1. ***Q-Switched Frequency-Doubled Nd^{3+}:YAG Microchip Laser.*** A slice of Nd^{3+}:YAG is brought together with a saturable absorber and an intracavity frequency-doubling crystal (see Sec. 22.2A) to form a 1-mm-long cavity. When pumped with 1 W of light from a fiber-coupled 808-nm laser diode, this microchip laser generates Q-switched optical pulses at 532 nm. Each pulse has an energy of 30 μJ and a duration of 250 ps. The repetition rate is $\approx$ 10 kHz and the average power is 300 mW.

EXERCISE 16.4-2

Pulsed Ruby Laser. Consider the ruby-laser example discussed in Exercise 16.1-1. Assume that the laser is now Q-switched and that at $t = t_i$ [see Fig. 16.4-6] the population difference $N_i = 6N_t$. Use Fig. 16.4-8 to estimate the shape and duration of the laser pulse. Calculate the approximate peak power, energy, and duration of the laser pulse.

D. Mode Locking

A laser can oscillate on many longitudinal modes, with frequencies that are equally separated by the Fabry–Perot intermodal spacing $\nu_F = c/2d$. Though these modes normally oscillate independently (they are then called free-running modes), techniques are available for coupling them and locking their phases together. The modes can then be regarded as the components of a Fourier-series expansion of a periodic function of time with period $T_F = 1/\nu_F = 2d/c$, which constitute a periodic pulse train. This was the approach taken in Sec. 2.6B, where we considered the interference of M monochromatic waves with equal intensities and equally spaced frequencies. We discuss in turn the properties of a mode-locked pulse train and methods of achieving mode locking, and then provide several examples of mode-locked lasers.

Properties of a Mode-Locked Pulse Train

If each of the laser modes is approximated by a uniform plane wave propagating in the z direction with velocity $c = c_o/n$, where n is the refractive index of the laser medium, the total complex wavefunction of the field may be written in the form of a sum:

$$U(z,t) = \sum_q A_q \exp\left[j2\pi\nu_q\left(t - \frac{z}{c}\right)\right], \tag{16.4-25}$$

where

$$\nu_q = \nu_0 + q\nu_F, \qquad q = 0, \pm1, \pm2, \ldots \tag{16.4-26}$$

is the frequency of mode q and A_q is its complex envelope. For convenience we assume that the $q = 0$ mode coincides with the central frequency ν_0 of the atomic lineshape. The magnitudes $|A_q|$ may be determined from knowledge of the spectral profile of the gain and the resonator loss (see Sec. 16.2B). In an inhomogeneously broadened medium, the modes interact with different groups of atoms so that their phases $\arg\{A_q\}$ are random and statistically independent.

Substituting (16.4-26) into (16.4-25) provides

$$U(z,t) = \mathcal{A}\left(t - \frac{z}{c}\right)\exp\left[j2\pi\nu_0\left(t - \frac{z}{c}\right)\right], \tag{16.4-27}$$

where the complex envelope $\mathcal{A}(t)$ is the function

$$\mathcal{A}(t) = \sum_q A_q \exp\left(\frac{jq2\pi t}{T_F}\right) \tag{16.4-28}$$

and

$$T_F = \frac{1}{\nu_F} = \frac{2d}{c}\,. \tag{16.4-29}$$

The complex envelope $\mathcal{A}(t)$ in (16.4-28) is a periodic function of t of period T_F while $\mathcal{A}(t - z/c)$ is a periodic function of z of period $cT_F = 2d$. If the magnitudes and phases of the complex coefficients A_q are properly chosen, $\mathcal{A}(t)$ may be made to take the form of a sequence of periodic narrow pulses.

Consider, for example, M modes ($q = 0, \pm1, \ldots, \pm S$, so that $M = 2S + 1$) whose complex coefficients are all equal: $A_q = A$ for $q = 0, \pm1, \ldots, \pm S$. Equation (16.4-28) then becomes

$$\mathcal{A}(t) = A\sum_{q=-S}^{S} \exp\left(\frac{jq2\pi t}{T_F}\right) = A\sum_{q=-S}^{S} x^q = A\,\frac{x^{S+1} - x^{-S}}{x-1} = A\,\frac{x^{S+1/2} - x^{-S-1/2}}{x^{1/2} - x^{-1/2}}\,, \tag{16.4-30}$$

where $x = \exp(j2\pi t/T_F)$ (see Sec. 2.6B for more details). After a few algebraic manipulations, $\mathcal{A}(t)$ can be cast in the form

$$\mathcal{A}(t) = A\,\frac{\sin(M\pi t/T_F)}{\sin(\pi t/T_F)}\,. \tag{16.4-31}$$

The optical intensity $I = |U|^2$ is then given by $I(t, z) = |\mathcal{A}(t - z/c)|^2$ so that

$$I(t, z) = |A|^2\,\frac{\sin^2\left[M\pi(t - z/c)/T_F\right]}{\sin^2\left[\pi(t - z/c)/T_F\right]}\,, \tag{16.4-32}$$

which is a periodic function of time, as illustrated in Fig. 16.4-9.

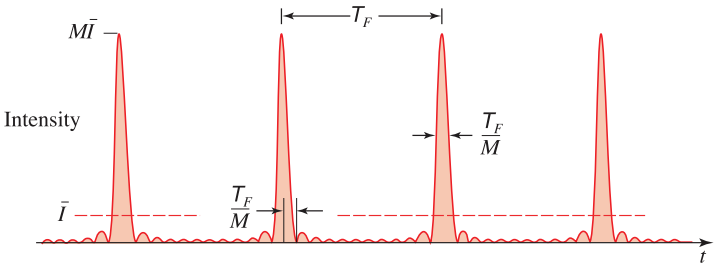


Figure 16.4-9 Intensity of the periodic pulse train resulting from the sum of M laser modes of equal magnitudes and phases. Each pulse has a duration τ_{pulse} that is M times smaller than the period T_F and a peak intensity that is M times greater than the mean intensity.

The behavior of the mode-locked laser pulse train is therefore dependent on the number of modes M, which is proportional to the transition linewidth $\Delta\nu$. If $M \approx \Delta\nu/\nu_F$, then the pulse duration $\tau_{\text{pulse}} = T_F/M \approx 1/\Delta\nu$. Since τ_{pulse} is inversely

proportional to the transition linewidth $\Delta\nu$, and since $\Delta\nu$ can be quite large, very narrow mode-locked laser pulses can generated. The ratio between the peak and mean intensities is equal to the number of modes M, which can also be quite large (Fig. 16.4-9).

The period of the pulse train is $T_F = 2d/c$ and its pulse repetition rate is $\nu_F = 1/T_F = c/2d$. The period T_F is just the time for a single round trip of reflection within the resonator. Indeed, the light in a mode-locked laser can be regarded as a single narrow pulse of photons that reflects back and forth between the mirrors of the resonator, as portrayed in Fig. 16.4-10. At each encounter with the output mirror, a fraction of the photons exits in the form of a pulse of light. The transmitted pulses are separated by the distance $c(2d/c) = 2d$ and each has a spatial width $d_{\text{pulse}} = c\tau_{\text{pulse}} = 2d/M$.

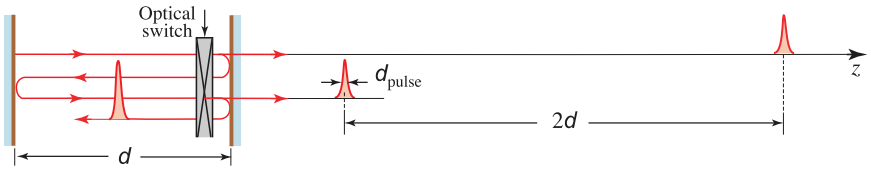


Figure 16.4-10 The mode-locked laser pulse reflects back and forth between the mirrors of the resonator. Each time it reaches the output mirror it transmits a short optical pulse. The transmitted pulses are separated by the distance $2d$ and travel with velocity c. The switch opens only when the pulse reaches the mirror and only for the duration of the pulse. The periodic pulse train is therefore unaffected by the presence of the switch. Other wave patterns, however, suffer losses that preclude oscillation.

The properties of a mode-locked laser pulse train are summarized in Table 16.4-1. Though the formulas presented in the table are applicable for the special case in which the modes have equal amplitudes and phases, calculations based on more realistic behavior provide similar results.

Table 16.4-1 Characteristic properties of a mode-locked pulse train.

Temporal period	$T_F = 2d/c$	Spatial period	$2d$
Frequency spacing	$\nu_F = c/2d$	Pulse repetition rate	$\nu_F = c/2d$
Pulse duration	$\tau_{\text{pulse}} = T_F/M = 1/\Delta\nu$	Pulse length	$d_{\text{pulse}} = 2d/M$
Mean intensity	$\bar{I}$	Peak intensity	$I_p = M\bar{I}$

EXERCISE 16.4-3

Demonstration of Pulsing by Mode Locking. Plot the intensity $I(t) = |\mathcal{A}(t)|^2$ of a wave whose envelope $\mathcal{A}(t)$ is given by the sum in (16.4-28). Assume that the number of modes $M = 11$ and use the following choices for the complex coefficients A_q:

(a) Equal magnitudes and the same phase (this should reproduce the results provided earlier).
(b) Magnitudes that obey the Gaussian spectral profile $|A_q| = \exp[-\frac{1}{2}(q/5)^2]$ and the same phase.
(c) Equal magnitudes and random phases (obtain the phases by using a random number generator to produce a random variable uniformly distributed between 0 and 2π).

Methods of Mode Locking

We have found thus far that if a large number M of modes are locked in phase, they form a giant narrow pulse of photons that reflects back and forth between the mirrors of the resonator. The spatial length of the pulse is a factor of M smaller than twice the resonator length. We now turn to a consideration of methods suitable for locking the phases of the modes together, a task that can be accomplished with the help of an active or passive modulator (switch) placed within the resonator. We consider **active mode locking** and **passive mode locking** in turn.

Suppose that an active optical modulator controlled by an external applied signal (e.g., an acousto-optic or electro-optic switch, as discussed in Secs. 20.2A and 21.1B, respectively) is placed inside the resonator, and that this switch blocks the light at all times except when the optical pulse is about to cross it, whereupon it opens for the duration of the pulse (Fig. 16.4-10). Since the pulse itself is permitted to pass, it is not affected by the presence of the switch and it continues its travels uninterrupted. In the absence of mode locking, the individual modes have different phases that are determined by the random conditions at the onset of their oscillation. If those phases happen, by accident, to take on equal values, the sum of the modes will form a giant pulse and it will not be affected by the presence of the switch. However, any other combination of phases will superpose to form a field distribution that is totally or partially blocked by the switch, which adds to the losses of the system. In the presence of the switch, therefore, lasing can occur only when the modal phases are equal. At turn-on, the laser waits for the "lucky accident" of phase alignment; once oscillation begins the locking continues.

The same argument can be framed in mathematical terms. An optical field must satisfy the wave equation with the boundary conditions imposed by the presence of the switch. The multimode optical field of (16.4-25) does indeed satisfy the wave equation for any combination of phases, including equal phases. Only this latter case, however, also satisfies the boundary conditions imposed by the switch so it is the unique solution.

A passive switch such as a **saturable absorber** may also be used to achieve mode locking. A saturable absorber (see Sec. 15.4A) is a medium whose absorption coefficient decreases as the intensity of the light passing through it increases; it thus transmits intense pulses with relatively little absorption while weak pulses are absorbed. Oscillation can therefore occur only when the phases of the different modes are related to each other in such a way that they form an intense pulse that is able pass through the absorbing medium. **Semiconductor saturable-absorber mirrors** (**SESAMs**), which are saturable absorbers that operate in reflection, are in widespread use; the more intense the light, the greater the reflection provided by these devices. SESAMs can accommodate wavelengths in the range from 800 to 1600 nm, pulse durations from fs to ns, and power levels from mW to hundreds of W. **Graphene saturable absorbers** are effective for achieving broadband laser mode locking since graphene has an absorbance that is nearly constant from $\lambda_o = 0.7$ to 25 μm, and also exhibits strong absorption saturation and fast (ps) recovery time (Sec. 17.1B). Saturable absorbers can also produce Q-switched mode locking, where the laser emits collections of mode-locked pulses within a Q-switching envelope.

Passive mode locking can also be implemented via **Kerr-lens mode locking**, which relies on the optical Kerr effect, a nonlinear optical phenomenon that causes the refractive index of a material to change with optical intensity (Sec. 22.3). A Kerr medium, which may be the gain medium itself or a material placed within the laser resonator, acts as a lens with a focal length that is inversely proportional to the intensity (Exercise 22.3-2). The Kerr lens is designed to reduce the area of the laser mode at a specified location within the resonator when the light intensity is high. An aperture placed at that location will then permit the light to pass through, but only when the phases are aligned so that the pulse intensity is high. Alternatively, the reduced modal

area in the gain medium can be used to increase its overlap with the strongly focused pump beam, thereby increasing the effective gain. The Kerr-lens approach is inherently broadband because of the parametric nature of the process. The rapid response and recovery inherent in passive mode locking generally leads to shorter optical pulses than can be obtained with active mode locking. Passive and active devices are used for the mode locking of inhomogeneously and homogeneously broadened media alike.

Examples of Mode-Locked Lasers

Table 16.4-2 provides a list of pulse durations available using selected mode-locked laser media. They are listed in order of increasing values of the observed pulse duration, which spans a broad range. The observed values depend not only on the medium, but also on the method used to achieve mode locking. Limitations are also imposed by nonlinearities and dispersion in the medium.

Table 16.4-2 Typical pulse durations for a number of mode-locked lasers subject to homogeneous (H) and inhomogeneous (I) broadening.

Laser Medium		Transition Linewidth[a] $\Delta\nu$	Calculated Pulse Duration $\tau_{\text{pulse}} = 1/\Delta\nu$	Observed Pulse Duration
Ti^{3+}:Al_2O_3 (Ti:Sapphire)	H	100 THz	10 fs	10 fs
Cr^{4+}:Mg_2SiO_4 (Forsterite)	H	50 THz	20 fs	20 fs
Rhodamine-6G dye	H/I	40 THz	25 fs	27 fs
Nd^{3+}:Glass (phosphate)	I	7 THz	140 fs	150 fs
Er^{3+}:Silica fiber	H/I	5 THz	200 fs	200 fs
Yb^{3+}:Silica fiber	H/I	5 THz	200 fs	5 ps
Nd^{3+}:YAG	H	150 GHz	7 ps	7 ps
Ar^+	I	3.5 GHz	286 ps	150 ps
He–Ne	I	1.5 GHz	667 ps	600 ps
CO_2	I	60 MHz	16 ns	20 ns

[a] The transition linewidths $\Delta\nu$ are drawn from Table 15.3-1.

EXAMPLE 16.4-2. ***Mode Locking in a Neodymium-Doped Glass Laser.*** Consider a Nd^{3+}:glass laser operating at $\lambda_o = 1.05\ \mu\text{m}$. It has a refractive index $n = 1.5$ and a transition linewidth $\Delta\nu = 7$ THz (Tables 15.3-1 and 16.4-2). The pulse duration is thus $\tau_{\text{pulse}} = 1/\Delta\nu \approx 140$ fs. If the resonator has a length $d = 15$ cm, the temporal period is $T_F = 2nd/c_o = 1.5$ ns and the mode separation (and pulse repetition rate) is $\nu_F = c_o/2nd = 0.67$ GHz. This yields $M = \Delta\nu/\nu_F \approx 10\,500$ modes and a pulse length of $d_{\text{pulse}} = 2d/M \approx 28.6\ \mu\text{m}$. The peak intensity I_p is 10 500 times greater than the mean intensity $\bar{I}$. In media with broad transition linewidths, mode locking is generally more advantageous than Q-switching for obtaining short pulses. However, gas lasers generally have narrow atomic linewidths that make it difficult to obtain ultrashort pulses via mode locking.

EXAMPLE 16.4-3. ***Mode Locking in an Ytterbium-Doped Fiber Laser.*** A passively mode-locked Yb^{3+}:silica-fiber laser operated at $\lambda_o = 1070$ nm produces an average power of 10 W in the form of 200-nJ pulses with a peak power of 40 kW, with the help of a SESAM. The pulse repetition rate is 50 MHz and the observed pulse duration is 5 ps; this is substantially longer than the expected value since $\Delta\nu = 5$ THz (Table 16.4-2), which provides $\tau_{\text{pulse}} = 1/\Delta\nu = 200$ fs. The discrepancy arises because of group velocity dispersion, which imparts broadening and also chirps the pulse as it travels through the laser medium (Fig. 5.7-3). The normal dispersion in silica fiber near $\lambda_o = 1\ \mu$m (Fig. 5.7-5) can be compensated by introducing anomalous dispersion via a fiber Bragg grating or a photonic-crystal fiber, which reduces the observed pulse duration to 200 fs. Additional reductions to the pulse duration can be effected by using suitable pulse-compression techniques (Sec. 23.2). Pulses with far greater energies and peak powers can be obtained.

Mode-locked lasers find use in many applications, including time-resolved measurements, imaging, metrology, communications, materials processing, and clinical medicine. The mode-locked laser of choice is often the Ti:sapphire laser, whose center wavelength can be tuned over the range 700–1050 nm and whose individual pulses have durations as short as 10 fs. A commonly available commercial version of this laser makes use of Kerr-lens mode locking and delivers 50-nJ pulses of 10-fs duration and 1-MW peak power, at a repetition rate of 80 MHz, but repetition rates in excess of 10 GHz are available. The spectral bandwidth $\Delta\nu$ of this laser can also be easily constrained to provide ps-duration mode-locked pulses. The intensity available from a mode-locked Ti:sapphire laser, or an amplified version thereof, is also sufficient to support harmonic generation and other nonlinear wavelength-shifting techniques (Chapters 22 and 23), which can provide sources of mode-locked pulses at shorter wavelengths. In particular, second-harmonic generation produces pulses in the range 350–525 nm and third-harmonic generation reaches 230–350 nm.

In the direction of longer wavelengths, mode-locked operation can be extended beyond $\lambda_o = 1$ μm by using a Ti:sapphire mode-locked laser oscillator as the source for a synchronously pumped optical parametric oscillator employing a crystal such as LBO or a periodically poled crystal (Sec. 22.4C). This approach provides mode-locked signal and idler output beams that cover the 1.0–3.3 μm infrared wavelength range. Mode-locked ytterbium-doped and erbium-doped fiber lasers operate in the vicinity of 1.07 and 1.55 μm, respectively. Though these lasers typically have large transition linewidths, achievable ultrafast pulse durations, along with other performance features, are often limited by the fiber dispersion and/or nonlinearities resulting from long fiber lengths and small modal volumes, respectively. Low pulse repetition rates arising from long resonator lengths may be increased by making use of **harmonic mode locking**, wherein multiple, well-spaced pulses circulate within the fiber resonator. High output powers from ultrafast fiber lasers are usually achieved by making use of chirped-pulse amplification.

In the domain of semiconductor lasers, vertical external-cavity surface-emitting lasers (VECSELS), also called **semiconductor disk lasers (SDLs)**, can be operated in mode-locked configurations with pulse durations < 100 fs and pulse repetition rates in the range of 1–50 GHz (Sec. 18.5A). Monolithic mode-locked laser diodes, by virtue of their very small resonator lengths, can exhibit pulse repetition rates that reach hundreds of GHz or even 1 THz (Example 18.3-4). In the mid infrared, quantum cascade lasers (QCLs) operating in the wavelength region $3 \leq \lambda_o \leq 12$ μm can generate mode-locked optical pulses with durations of a few ps (Sec. 18.4D).

*E. Optical Frequency Combs

Since light from a mode-locked laser has a discrete optical spectrum with uniformly spaced frequencies over a broad band, the spectrum is known as an **optical frequency comb** (**OFC**). The qth frequency component of an OFC is $\nu_q = q\nu_F + \nu_i$, where ν_F is the frequency spacing between the adjacent 'teeth' of the comb and $\nu_i < \nu_F$ is an offset frequency. The number of frequency components M contained in the comb is typically large so that the spectral bandwidth is much greater than the frequency spacing ν_F. An OFC is thus completely described by two frequencies, ν_F and ν_i, both of which are readily measured. Whereas the ν_q are optical frequencies, the frequency spacing for mode-locked lasers $\nu_F = c/2d$ typically lies between tens of MHz and hundreds of GHz, as determined by the length of the Fabry–Perot resonator.

By enabling the high-precision measurement of optical and ultraviolet frequencies, OFCs have found myriad uses in physics and astronomy. They have become valuable tools for precision molecular, atomic, and nuclear spectroscopy, as well as for precision optical imaging and metrology. Moreover, OFCs comprising entangled photon pairs

generated via on-chip quantum circuits (Sec. 13.3D) offer parameter estimation with even greater precision.

Measurement of an octave-spanning OFC. If the highest frequency of an OFC is a factor of two greater than the lowest frequency, the OFC is said to span a frequency octave. Since the frequency spacing is established by the fixed laser cavity, the comb is highly uniform and ν_F is highly stable. The OFC may therefore serve as a frequency ruler against which unknown frequencies may be measured. Such measurements usually rely on light beating (Sec. 2.6B), which generates frequency differences that are readily determined with electronic instruments.

The frequencies ν_q of an OFC may be determined if the frequencies ν_F and ν_i, which are far smaller than ν_q, are measured. The measurement of ν_F is straightforward since it is the pulse repetition rate of the mode-locked pulse train, which can be quantified with a fast detector. Alternatively, the value of ν_F can be established with the help of an electronic spectrum analyzer, which will display the beat frequency between adjacent frequencies in the comb.

Establishing the value of ν_i is more difficult, though straightforward. If the comb spans a frequency octave, ν_i may be elicited by beating the comb with a frequency-doubled version of itself. Frequency doubling may be implemented by making use of second-harmonic generation in a nonlinear optical medium, as described in Sec. 22.2A. As illustrated in Fig. 16.4-11, after frequency doubling the component of order q from the low-frequency side of the frequency-doubled comb has approximately the same frequency as the component $2q$ on the high-frequency side of the original comb, i.e., $2\nu_q \approx \nu_{2q}$. Measuring the beat frequency between these combs yields the offset frequency since $2\nu_q - \nu_{2q} = 2(q\nu_F + \nu_i) - (2q\nu_F + \nu_i) = \nu_i$. The outcome of such a measurement may be used to control the source of the OFC, or even to eliminate the offset frequency altogether, thereby providing a perfect frequency ruler in which $\nu_q = q\nu_F$.

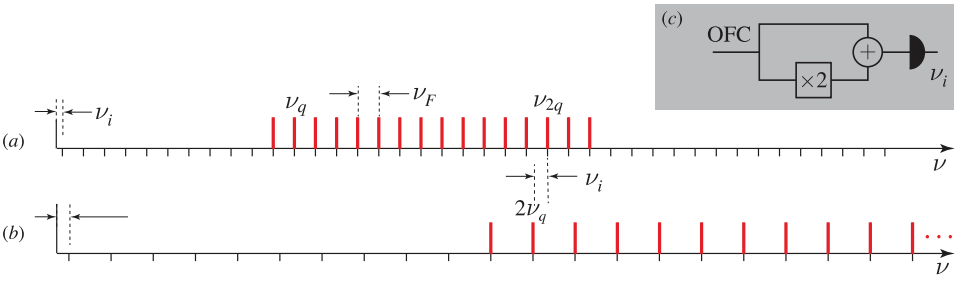


Figure 16.4-11 (a) Spectrum of an octave-spanning optical frequency comb. (b) Frequency-doubled spectrum of the comb in (a). (c) Beating these two spectra yields the offset frequency ν_i.

Precise measurement of optical frequency. An OFC may be used to increase the precision of an optical-frequency measurement made with a conventional wavelength-sensitive spectrum analyzer. If the coarse measurement of the frequency ν_L of a monochromatic light source such as a CW laser is approximately equal to a multiple of the frequency spacing of an OFC, i.e., if $\nu_L \approx q\nu_F$, the precision of the frequency measurement can be enhanced by beating the two optical fields. The smallest beat frequency $\nu_L - (q\nu_F + \nu_i)$ is then used to calculate a more precise value of ν_L. The process is akin to making use of a Vernier scale such as that used to obtain fractional readings from a uniformly divided ruler. In the event that the offset frequency ν_i is not known, then both ν_i and the correction $\nu_L - q\nu_F$ may be determined by beating the OFC once with the CW laser, and again with a frequency-doubled version thereof, as

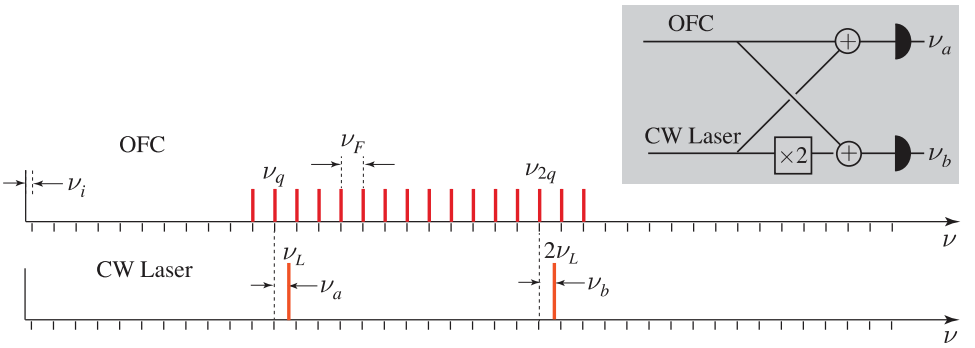


Figure 16.4-12 Precise measurement of the frequency of a CW laser by beating an octave-spanning OFC once with the CW laser, and again with a second-harmonic of the CW laser.

illustrated in Fig. 16.4-12. Since $\nu_L \approx \nu_q$ and $2\nu_L \approx \nu_{2q}$, the resultant beat frequencies are, respectively:

$$\nu_a = \nu_L - (q\nu_F + \nu_i) = (\nu_L - q\nu_F) - \nu_i \tag{16.4-33}$$

$$\nu_b = 2\nu_L - (2q\nu_F + \nu_i) = 2(\nu_L - q\nu_F) - \nu_i. \tag{16.4-34}$$

Solving (16.4-33) and (16.4-34) yields both the offset frequency $\nu_i = \nu_b - 2\nu_a$ and the correction $\nu_L - q\nu_F = \nu_b - \nu_a$. This analysis assumes that the OFC is octave spanning and that $q\nu_F$ lies toward the low-frequency side of the comb.

Extreme-ultraviolet and X-ray OFCs. As indicated earlier, the frequencies ν_q of OFCs generated using mode-locked lasers usually fall in the visible region of the spectrum while the pulse repetition rates $\nu_F = c/2d$ typically range from tens of MHz to hundreds of GHz, depending on the length d of the Fabry–Perot laser resonator (Sec. 11.1A). In recent years, high-harmonic generation (HHG, Sec. 23.5D) has emerged as an alternative technique for generating optical frequency combs. The HHG approach is distinct from mode locking and has made it possible to substantially increase the frequency spacings and frequency reach of OFCs. An optical frequency comb generated via HHG exhibits spacings between adjacent frequency components that lie at twice the frequency of the exciting laser, which typically operates in the near or mid infrared. OFC difference frequencies (and pulse repetition rates) are thus in the vicinity of hundreds of THz, orders of magnitude greater than those obtained using mode-locking techniques. Moreover, at high gas pressures the highest frequency components of OFCs generated via HHG can stretch to the SXR region (Fig. 16.3-7). The pulse trains associated with HHG-generated OFCs exhibit attosecond structure.

READING LIST

Lasers

See also the reading list in Chapter 15.

D. Meschede, *Optics, Light and Lasers: The Practical Approach to Modern Aspects of Photonics and Laser Physics*, Wiley–VCH, 3rd ed. 2017.

K. F. Renk, *Basics of Laser Physics: For Students of Science and Engineering*, Springer-Verlag, 2nd ed. 2017.

M. Prelas, *Nuclear-Pumped Lasers*, Springer-Verlag, 2016.

V. V. Apollonov, *High Energy Molecular Lasers: Self-Controlled Volume-Discharge Lasers and Applications*, Springer-Verlag, 2016.

B. Zohuri, *Directed Energy Weapons: Physics of High Energy Lasers (HEL)*, Springer-Verlag, 2016.

C. C. Davis, *Lasers and Electro-Optics: Fundamentals and Engineering*, Cambridge, 2nd ed. 2014.

F. Träger, ed., *Springer Handbook of Lasers and Optics*, Springer-Verlag, 2nd ed. 2012.

P. W. Milonni and J. H. Eberly, *Laser Physics*, Wiley, 2nd ed. 2010.

O. Svelto, *Principles of Lasers*, Springer-Verlag, paperback 5th ed. 2010.

W. T. Silfvast, *Laser Fundamentals*, Cambridge University Press, paperback 2nd ed. 2008.

R. Paschotta, *Field Guide to Lasers*, SPIE Optical Engineering Press, 2008.

A. Yariv and P. Yeh, *Photonics: Optical Electronics in Modern Communications*, Oxford University Press, 6th ed. 2006.

J.-M. Liu, *Photonic Devices*, Cambridge University Press, 2005, paperback ed. 2009.

N. G. Basov, A. S. Bashkin, V. I. Igoshin, A. N. Oraevsky, and V. A. Shcheglov, *Chemical Lasers*, Springer-Verlag, 1990, paperback ed. 2011.

A. E. Siegman, *Lasers*, University Science, 1986.

Solid-State Lasers

S. B. Mirov, V. V. Fedorov, D. Martyshkin, I. S. Moskalev, M. Mirov, and S. Vasilyev, Progress in Mid-IR Lasers Based on Cr and Fe-Doped II–VI Chalcogenides, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 21, 1601719, 2015.

A. Ikesue, Y. L. Aung, and V. Lupei, *Ceramic Lasers*, Cambridge University Press, 2013.

W. Koechner, *Solid-State Laser Engineering*, Springer-Verlag, 6th ed. 2006.

V. V. Ter-Mikirtychev, ed., *Selected Papers on Tunable Solid-State Lasers*, SPIE Optical Engineering Press (Milestone Series Volume 173), 2002.

L. D. DeLoach, R. H. Page, G. D. Wilke, S. A. Payne, and W. F. Krupke, Transition Metal-Doped Zinc Chalcogenides: Spectroscopy and Laser Demonstration of a New Class of Gain Media, *IEEE Journal of Quantum Electronics*, vol. 32, pp. 885–895, 1996.

P. F. Moulton, Spectroscopic and Laser Characteristics of Ti:$A1_2O_3$, *Journal of the Optical Society of America B*, vol. 3, pp. 125–133, 1986.

J. E. Geusic, H. M. Marcos, and L. G. Van Uitert, Laser Oscillations in Nd-Doped Yttrium Aluminum, Yttrium Gallium and Gadolinium Garnets, *Applied Physics Letters*, vol. 4, pp. 182–184, 1964.

P. P. Sorokin and M. J. Stevenson, Stimulated Infrared Emission from Trivalent Uranium, *Physical Review Letters*, vol. 5, pp. 557–559, 1960.

T. H. Maiman, Stimulated Optical Radiation in Ruby, *Nature*, vol. 187, pp. 493–494, 1960.

Fiber, Stimulated-Raman, and Stimulated-Brillouin Lasers

J. Hecht, High-Power Fiber Lasers, *Optics & Photonics News*, vol. 29, no. 10, pp. 30–37, 2018.

L. Dong and B. Samson, *Fiber Lasers: Basics, Technology, and Applications*, CRC Press, 2017.

S. Fu, W. Shi, Y. Feng, L. Zhang, Z. Yang, S. Xu, X. Zhu, R. A. Norwood, and N. Peyghambarian, Review of Recent Progress on Single-Frequency Fiber Lasers, *Journal of the Optical Society of America B*, vol. 34, pp. A49–A62, 2017.

P. Zhou, H. Xiao, J. Leng, J. Xu, Z. Chen, H. Zhang, and Z. Liu, High-Power Fiber Lasers Based on Tandem Pumping, *Journal of the Optical Society of America B*, vol. 34, pp. A29–A36, 2017.

V. R. Supradeepa, Y. Feng, and J. W. Nicholson, Raman Fiber Lasers, *Journal of Optics*, vol. 19, 023001, 2017.

N. K. Dutta, *Fiber Amplifiers and Fiber Lasers*, World Scientific, 2015.

M. N. Zervas and C. A. Codemard, High Power Fiber Lasers: A Review, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, 0904123, 2014.

H. Rong, A. Liu, R. Jones, O. Cohen, D. Hak, R. Nicolaescu, A. Fang, and M. Paniccia, An All-Silicon Raman Laser, *Nature*, vol. 433, pp. 292–294, 2005.

O. Boyraz and B. Jalali, Demonstration of a Silicon Raman Laser, *Optics Express*, vol. 12, pp. 5269–5273, 2004.

T. J. Kippenberg, S. M. Spillane, B. Min, and K. J. Vahala, Theoretical and Experimental Study of Stimulated and Cascaded Raman Scattering in Ultrahigh-Q Optical Microcavities, *IEEE Journal of Selected Topics in Quantum Electronics*, vol 10, pp. 1219–1228, 2004.

M. J. F. Digonnet, ed., *Rare-Earth-Doped Fiber Lasers and Amplifiers*, CRC Press, 2nd ed. 2001.

D. C. Hanna, R. M. Percival, I. R. Perry, R. G. Smart, P. J. Suni, J. E. Townsend, and A. C. Tropper, Continuous-Wave Oscillation of a Monomode Ytterbium-Doped Fibre Laser, *Electronics Letters*, vol. 24, pp. 1111–1113, 1988.

M. E. Fermann, D. C. Hanna, D. P. Shepherd, P. J. Suni, and J. E. Townsend, Efficient Operation of an Yb-Sensitised Er Fibre Laser at 1.56 μm, *Electronics Letters*, vol. 24, pp. 1135–1136, 1988.

K. O. Hill, B. S. Kawasaki, and D. C. Johnson, CW Brillouin Laser, *Applied Physics Letters*, vol. 28, pp. 608–609, 1976.

E. Snitzer, Laser Emission at 1.06 μ from Nd^{3+}–Yb^{3+} Glass, *IEEE Journal of Quantum Electronics*, vol. QE-2, pp. 562–566, 1966.

E. Snitzer, Proposed Fiber Cavities for Optical Masers, *Journal of Applied Physics*, vol. 32, pp. 36–39, 1961.

E. Snitzer, Optical Maser Action of Nd^{3+} in a Barium Crown Glass, *Physical Review Letters*, vol. 7, pp. 444–446, 1961.

Random Lasers

J. Mysliwiec, K. Cyprych, L. Sznitko and A. Miniewicz, Biomaterials in Light Amplification, *Journal of Optics*, vol. 19, 033003, 2017.

S. F. Yu, Electrically Pumped Random Lasers, *Journal of Physics D*, vol. 48, 483001, 2015.

M. C. Gather and S. H. Yun, Single-Cell Biological Lasers, *Nature Photonics*, vol. 5, pp. 406–410, 2011.

V. Letokhov and S. Johansson, *Astrophysical Lasers*, Oxford University Press, 2009.

D. S. Wiersma, The Physics and Applications of Random Lasers, *Nature Physics*, vol. 4, pp. 359–367, 2008.

M. A. Noginov, *Solid-State Random Lasers*, Springer-Verlag, 2005.

H. Cao, Lasing in Disordered Media, in E. Wolf, ed., *Progress in Optics*, North-Holland, 2003, vol. 45, pp. 317–370.

H. Cao, Y. G. Zhao, S. T. Ho, E. W. Seelig, Q. H. Wang, and R. P. H. Chang, Random Laser Action in Semiconductor Powder, *Physical Review Letters*, vol. 82, pp. 2278–2281, 1999.

N. M. Lawandy, R. M. Balachandran, A. S. L. Gomes, and E. Sauvain, Laser Action in Strongly Scattering Media, *Nature*, vol. 368, pp. 436–438, 1994.

M. Elitzur, *Astronomical Masers*, Springer-Verlag/Kluwer, 1992.

M. J. Mumma, D. Buhl, G. Chin, D. Deming, F. Espenak, and T. Kostiuk, Discovery of Natural Gain Amplification in the 10 μm CO_2 Laser Bands on Mars: A Natural Laser, *Science*, vol. 212, pp. 45–49, 1981.

F. Varsanyi, Surface Lasers, *Applied Physics Letters*, vol. 19, pp. 169–171, 1971.

V. S. Letokhov, Generation of Light by a Scattering Medium with Negative Resonance Absorption, *Soviet Physics–JETP*, vol. 26, pp. 835–840, 1968 [*Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki*, vol. 53, pp. 1442–1447, 1967].

X-Ray and Free-Electron Lasers

H. P. Freund and T. M. Antonsen, Jr., *Principles of Free-Electron Lasers*, Springer, 3rd ed. 2018.

K.-J. Kim, Z. Huang, and R. Lindberg, *Synchrotron Radiation and Free-Electron Lasers: Principles of Coherent X-Ray Generation*, Cambridge University Press, 2017.

C. A. MacDonald, An Introduction to X-Ray Physics, Optics, and Applications, Princeton, 2017.

D. Attwood and A. Sakdinawat, *X-Rays and Extreme Ultraviolet Radiation: Principles and Applications*, Cambridge University Press, 2nd ed. 2016.

C. Pellegrini, A. Marinelli, and S. Reiche, The Physics of X-Ray Free-Electron Lasers, *Reviews of Modern Physics*, vol. 88, 015006, 2016.

C. Bostedt, S. Boutet, D. M. Fritz, Z. Huang, H. J. Lee, H. T. Lemke, A. Robert, W. F. Schlotter, J. J. Turner, and G. J. Williams, Linac Coherent Light Source: The First Five Years, *Reviews of Modern Physics*, vol. 88, 015007, 2016.

F. Canova and L. Poletto, eds., *Optical Technologies for Extreme-Ultraviolet and Soft X-Ray Coherent Sources*, Springer-Verlag, 2015.

P. H. Bucksbaum and N. Berrah, Brighter and Faster: The Promise and Challenge of the X-Ray Free-Electron Laser, *Physics Today*, vol. 68, no. 7, pp. 26–32, 2015.

P. Schmüser, M. Dohlus, J. Rossbach, and C. Behrens, *Free-Electron Lasers in the Ultraviolet and X-Ray Regime*, Springer-Verlag, 2nd ed. 2014.

C. Pellegrini, The History of X-ray Free-Electron Lasers, *The European Physical Journal H*, vol. 37, pp. 659–708, 2012.

J. N. Galayda, J. Arthur, D. F. Ratner, and W. E. White, X-ray Free-Electron Lasers — Present and Future Capabilities, *Journal of the Optical Society of America B*, vol. 27, pp. B106–B118, 2010.

S. Suckewer and P. Jaeglé, X-Ray Laser: Past, Present, and Future, *Laser Physics Letters*, vol. 6, pp. 411–436, 2009.

J. Hecht, The History of the X-ray Laser, *Optics & Photonics News*, vol. 19, no. 5, pp. 26–33, 2008.

P. Jaeglé, *Coherent Sources of XUV Radiation: Soft X-Ray Lasers and High-Order Harmonic Generation*, Springer-Verlag, 2006, paperback ed. 2010.

D. M. Paganin, *Coherent X-Ray Optics*, Oxford University Press, 2006, paperback ed. 2013.

E. L. Saldin, E. A. Schneidmiller, and M. V. Yurkov, *The Physics of Free Electron Lasers*, Springer-Verlag, 2000.

R. W. Waynant and M. N. Ediger, eds., *Selected Papers on UV, VUV, and X-Ray Lasers*, SPIE Optical Engineering Press (Milestone Series Volume 71), 1993.

P. Luchini and H. Motz, *Undulators and Free-Electron Lasers*, Oxford University Press, 1990.

R. C. Elton, *X-Ray Lasers*, Academic Press, 1990.

T. C. Marshall, *Free-Electron Lasers*, Macmillan, 1985.

D. L. Matthews, P. L. Hagelstein, M. D. Rosen, M. J. Eckart, N. M. Ceglio, A. U. Hazi, H. Medecki, B. J. MacGowan, J. E. Trebes, B. L. Whitten, E. M. Campbell, C. W. Hatcher, A. M. Hawryluk, R. L. Kauffman, L. D. Pleasance, G. Rambach, J. H. Scofield, G. Stone, and T. A. Weaver, Demonstration of a Soft X-Ray Amplifier, *Physical Review Letters*, vol. 54, pp. 110–113, 1985.

P. Jaeglé, G. Jamelot, A. Carillon, A. Sureau, and P. Dhez, Superradiant Line in the Soft-X-Ray Range, *Physical Review Letters*, vol. 33, pp. 1070–1073, 1974.

J. M. J. Madey, Stimulated Emission of Bremsstrahlung in a Periodic Magnetic Field, *Journal of Applied Physics*, vol. 42, pp. 1906–1913, 1971.

M. A. Duguay and P. M. Rentzepis, Some Approaches to Vacuum UV and X-Ray Lasers, *Applied Physics Letters*, vol. 10, pp. 350–352, 1967.

Pulsed Lasers and Optical Frequency Combs

I. Coddington, N. Newbury, and W. Swann, Dual-Comb Spectroscopy, *Optica*, vol. 3, pp. 414–426, 2016.

P. H. Bucksbaum, Sources and Science of Attosecond Light, *Optics & Photonics News*, vol. 26, no. 5, pp. 28–35, 2015.

W. H. Renninger and F. W. Wise, Fundamental Limits to Mode-Locked Lasers: Toward Terawatt Peak Powers, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 21, 1100208, 2015.

M. E. Fermann and I. Hartl, Ultrafast Fibre Lasers, *Nature Photonics*, vol. 7, pp. 868–874, 2013.

P. G. Kryukov, Continuous-Wave Femtosecond Lasers, *Physics–Uspekhi*, vol. 56, pp. 849–867, 2013.

L. N. Binh and N. Q. Ngo, *Ultra-Fast Fiber Lasers*, CRC Press/Taylor & Francis, 2011.

S. A. Diddams, The Evolving Optical Frequency Comb, *Journal of the Optical Society of America B*, vol. 27, pp. B51–B62, 2010.

R. Paschotta, *Field Guide to Laser Pulse Generation*, SPIE Optical Engineering Press, 2008.

T. W. Hänsch, Nobel Lecture: Passion for Precision, *Reviews of Modern Physics*, vol. 78, pp. 1297–1309, 2006.

J. L. Hall, Nobel Lecture: Defining and Measuring Optical Frequencies, *Reviews of Modern Physics*, vol. 78, pp. 1279–1295, 2006.

J. Ye and S. T. Cundiff, eds., *Femtosecond Optical Frequency Comb Technology: Principle, Operation and Application*, Springer-Verlag, 2005, paperback ed. 2010.

L. E. Hargrove, R. L. Fork, and M. A. Pollack, Locking of He–Ne Laser Modes Induced by Synchronous Intracavity Modulation, *Applied Physics Letters*, vol. 5, pp. 4–5, 1964.

Popular and Review

T. H. Maiman, *The Laser Inventor: Memoirs of Theodore H. Maiman*, Springer-Verlag, 2018.

M. Bertolotti, *Masers and Lasers: An Historical Approach*, CRC Press, 2nd ed. 2015.

Special Issue, The Laser at 50, *Physics World*, vol. 23, no. 5, May 2010.

J. Hecht, *Understanding Lasers: An Entry Level Guide*, Wiley–IEEE, paperback 3rd ed. 2008.

J. Hecht, *Beam: The Race to Make the Laser*, Oxford University Press, 2005.

R. L. Walsworth, The Maser at 50, *Science*, vol. 306, pp. 236–237, 2004.

Millennium issue, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 6, no. 6, 2000.

C. H. Townes, *How the Laser Happened: Adventures of a Scientist*, Oxford University Press, 1999.

J. L. Bromberg, The Birth of the Laser, *Physics Today*, vol. 41, no. 10, pp. 26–33, 1988.

J. Hecht, ed., *Laser Pioneer Interviews*, High Tech Publications, 1985.

A. Kastler, Birth of the Maser and Laser, *Nature*, vol. 316, pp. 307–309, 1985.

Special issue, Twenty-Five Years of the Laser, *Optica Acta (Journal of Modern Optics)*, vol. 32, no. 9/10, 1985.

C. H. Townes, Science, Technology, and Invention: Their Progress and Interactions, *Proceedings of the National Academy of Sciences (USA)*, vol. 80, pp. 7679–7683, 1983.

A. L. Schawlow, Maser and Laser, *IEEE Transactions on Electron Devices*, vol. ED-23, pp. 773–779, 1976.

W. E. Lamb, Jr., Physical Concepts in the Development of the Maser and Laser, in B. Kursunoglu and A. Perlmutter, eds., *Impact of Basic Research on Technology*, Plenum, 1973.

C. Cohen-Tannoudji and A. Kastler, Optical Pumping, in E. Wolf, ed., *Progress in Optics*, North-Holland, 1966, vol. 5, pp. 1–81.

A. Yariv and J. P. Gordon, The Laser, *Proceedings of the IEEE*, vol. 51, pp. 4–29, 1963.

Historical and Reprint Collections

See also the reading list on historical articles in Chapter 18.

W. T. Silfvast, ed., *Selected Papers on Fundamentals of Lasers*, SPIE Optical Engineering Press (Milestone Series Volume 70), 1993.

D. O'Shea and D. C. Peckham, eds., *Lasers: Selected Reprints*, American Association of Physics Teachers, 1982.

A. L. Schawlow, Spectroscopy in a New Light (Nobel Lecture in Physics, 1981), *Reviews of Modern Physics*, vol. 54, pp. 697–707, 1982.

D. C. O'Shea and D. C. Peckham, eds., Resource Letter L-1: Lasers, *American Journal of Physics*, vol. 49, pp. 915–925, 1981.

F. S. Barnes, ed., *Laser Theory*, IEEE Press Reprint Series, IEEE Press, 1972.

J. Weber, ed., *Lasers: Selected Reprints with Editorial Comment*, CRC Press, 1967.

A. Kastler, Optical Methods for Studying Hertzian Resonances (Nobel Lecture in Physics, 1966), in *Nobel Lectures in Physics, 1963–1970*, World Scientific, 1998, pp. 186–204.

C. H. Townes, Production of Coherent Radiation by Atoms and Molecules (Nobel Lecture in Physics, 1964), in *Nobel Lectures in Physics, 1963–1970*, World Scientific, 1998, pp. 58–86.

N. G. Basov, Semiconductor Lasers (Nobel Lecture in Physics, 1964), in *Nobel Lectures in Physics, 1963–1970*, World Scientific, 1998, pp. 89–105.

A. M. Prokhorov, Quantum Electronics (Nobel Lecture in Physics, 1964), in *Nobel Lectures in Physics, 1963–1970*, World Scientific, 1998, pp. 110–116.

W. E. Lamb, Jr., Theory of an Optical Maser, *Physical Review*, vol. 134, pp. A1429–A1450, 1964.

A. Javan, W. R. Bennett, Jr., and D. R. Herriott, Population Inversion and Continuous Optical Maser Oscillation in a Gas Discharge Containing a He–Ne Mixture, *Physical Review Letters*, vol. 6, pp. 106–110, 1961.

T. H. Maiman, Stimulated Optical Radiation in Ruby, *Nature*, vol. 187, pp. 493–494, 1960.

R. H. Dicke, Molecular Amplification and Generation Systems and Methods, *U.S. Patent 2,851,652*, Patented September 9, 1958.

A. L. Schawlow and C. H. Townes, Infrared and Optical Masers, *Physical Review*, vol. 112, pp. 1940–1949, 1958.

J. P. Gordon, H. J. Zeiger, and C. H. Townes, The Maser—New Type of Microwave Amplifier, Frequency Standard, and Spectrometer, *Physical Review*, vol. 99, pp. 1264–1274, 1955.

N. G. Basov and A. M. Prokhorov, Possible Methods of Obtaining Active Molecules for a Molecular Oscillator, *Soviet Physics–JETP*, vol. 1, pp. 184–185, 1955 [*Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki*, vol. 28, pp. 249–250, 1955].

J. P. Gordon, H. J. Zeiger, and C. H. Townes, Molecular Microwave Oscillator and New Hyperfine Structure in the Microwave Spectrum of NH_3, *Physical Review*, vol. 95, pp. 282–284, 1954.

V. A. Fabrikant, The Emission Mechanism of Gas Discharges, *Trudi Vsyesoyuznogo Elektrotekhnicheskogo Instituta* (Reports of the All-Union Electrotechnical Institute, Moscow), vol. 41, *Elektronnie i Ionnie Pribori* (Electron and Ion Devices), pp. 236–296, 1940.

R. Ladenburg, Dispersion in Electrically Excited Gases, *Reviews of Modern Physics*, vol. 5, pp. 243–256, 1933.

H. Kopfermann and R. Ladenburg, Experimental Proof of 'Negative Dispersion,' *Nature*, vol. 122, pp. 438–439, 1928.

A. Einstein, Zur Quantentheorie der Strahlung, *Physikalische Zeitschrift*, vol. 18, pp. 121–128, 1917 [Translation: On the Quantum Theory of Radiation, in D. ter Haar, *The Old Quantum Theory*, Pergamon, 1967].

PROBLEMS

16.2-2 **Number of Longitudinal Modes.** An Ar^+-ion laser has a resonator of length 100 cm. The refractive index $n = 1$.

(a) Determine the frequency spacing ν_F between the resonator modes.

(b) Determine the number of longitudinal modes that the laser can sustain if the FWHM Doppler-broadened linewidth is $\Delta\nu_D = 3.5$ GHz and the loss coefficient is half the peak small-signal gain coefficient.

(c) What would the resonator length d have to be to achieve operation on a single longitudinal mode? What would that length be for a CO_2 laser, which has a much smaller Doppler linewidth $\Delta\nu_D = 60$ MHz under the same conditions?

16.2-3 **Frequency Drift of the Laser Modes.** A He–Ne laser has the following characteristics: (1) A resonator with 97% and 100% mirror reflectances and negligible internal losses; (2) a Doppler-broadened atomic transition with Doppler linewidth $\Delta\nu_D = 1.5$ GHz; and (3) a small-signal peak gain coefficient $\gamma_0(\nu_0) = 2.5\times10^{-3}\ \text{cm}^{-1}$. While the laser is running, the frequencies of its longitudinal modes drift with time as a result of small thermally induced changes in the length of the resonator. Find the allowable range of resonator lengths such that the laser will always oscillate in one or two (but not more) longitudinal modes. The refractive index $n = 1$.

16.2-4 **Mode Control Using an Etalon.** A Doppler-broadened gas laser operates at 515 nm in a resonator with two mirrors separated by a distance of 50 cm. The photon lifetime is 0.33 ns. The spectral window within which oscillation can occur is of width $B = 1.5$ GHz. The refractive index $n = 1$. To select a single mode, the light is passed into an etalon (a passive Fabry–Perot resonator) whose mirrors are separated by the distance d and its finesse is $\mathcal{F}$. The etalon acts as a filter. Suggest suitable values of d and $\mathcal{F}$. Is it better to place the etalon inside or outside the laser resonator?

16.2-5 **Modal Powers in a Multimode Laser.** A He–Ne laser operating at $\lambda_o = 632.8$ nm produces 50 mW of multimode power at its output. It has an inhomogeneously broadened gain profile with a Doppler linewidth $\Delta\nu_D = 1.5$ GHz and the refractive index $n = 1$. The resonator is 30 cm long.

(a) If the maximum small-signal gain coefficient is twice the loss coefficient, determine the number of longitudinal modes of the laser.

(b) If the mirrors are adjusted to maximize the intensity of the strongest mode, estimate its power.

16.2-6 **Output of a Single-Mode Gas Laser.** Consider a 10-cm-long gas laser operating at the center of the 600-nm line in a single longitudinal and single transverse mode. The mirror reflectances are $\mathcal{R}_1 = 99\%$ and $\mathcal{R}_2 = 100\%$. The refractive index $n = 1$ and the effective area of the output beam is $1\ \text{mm}^2$. The small-signal gain coefficient $\gamma_0(\nu_0) = 0.1\ \text{cm}^{-1}$ and the saturation photon-flux density $\phi_s = 1.43 \times 10^{19}\ \text{photons/cm}^2$-s.

(a) Determine the distributed loss coefficients, α_{m1} and α_{m2}, associated with each of the mirrors separately. Assuming that $\alpha_s = 0$, find the resonator loss coefficient α_r.

(b) Find the photon lifetime τ_p.

(c) Determine the output photon-flux density ϕ_o and the output power P_o.

16.2-7 **Transmittance of a Laser Resonator.** Monochromatic light from a tunable optical source is transmitted through the optical resonator of an unpumped gas laser. The observed transmittance, as a function of frequency, is shown in Fig. P16.2-7.

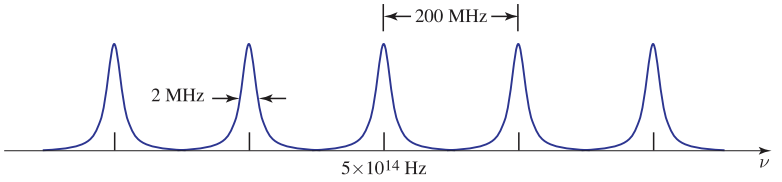


Figure P16.2-7 Transmittance of a laser resonator.

(a) Determine the resonator length, the photon lifetime, and the threshold gain coefficient of the laser. Assume that the refractive index $n = 1$.

(b) Assuming that the central frequency of the laser transition is 5×10^{14} Hz, sketch the transmittance versus frequency if the laser is now pumped but the pumping is not sufficient for laser oscillation to occur.

16.2-8 **Rate Equations for a Four-Level Laser.** Consider a four-level laser with an active volume $V = 1\ \text{cm}^3$. The population densities of the upper and lower laser levels are N_2 and N_1 and $N = N_2 - N_1$. The pumping rate is such that the steady-state population difference N in the absence of the stimulated emission and absorption is N_0. The photon-number density is $\mathfrak{n}$ and the photon lifetime is τ_p. Write the rate equations for N_2, N_1, N, and $\mathfrak{n}$ in terms of N_0, the transition cross section $\sigma(\nu)$, and the times t_{sp}, τ_1, τ_2, τ_{21}, and τ_p. Determine the steady-state values of N and $\mathfrak{n}$.

16.3-1 **Operation of an Ytterbium-Doped YAG Laser.** Yb^{3+}:YAG is a rare-earth-doped dielectric material that lases at $\lambda_o = 1.030\ \mu$m on the ${}^2F_{5/2} \rightarrow {}^2F_{7/2}$ transition (see Tables 14.1-1, 15.3-1, 16.3-1, and Fig. 16.3-2). This three-level laser is usually optically pumped with an array of InGaAs laser diodes.

(a) The pump band (level ③) has a central energy of 1.31915 eV and a width of 0.02475 eV. Determine the free-space wavelength of the desired laser-diode pump and the width of the absorption band in nm.

(b) At the central frequency of the laser transition ν_0, the peak transition cross section $\sigma_0 \equiv \sigma(\nu_0) = 2 \times 10^{-20}\ \text{cm}^2$. Given that the Yb^{3+}-ion doping density is set at $N_a = 1.4 \times 10^{20}\ \text{cm}^{-3}$, determine the absorption and gain coefficients of the material at the center of the line, $\alpha(\nu_0) \equiv -\gamma(\nu_0)$. Assume that the material is in thermal equilibrium at $T = 300°$ K (i.e., there is no pumping).

(c) Consider a laser rod constructed from this material with a length of 6 cm and a diameter of 2 mm. One of its ends is polished to a reflectance of 80% ($\mathcal{R}_1 = 0.8$) while the other is polished to unity reflectance ($\mathcal{R}_2 = 1.0$). Assuming that there is no scattering, and that there are no other extraneous losses, determine the resonator loss coefficient α_r and the resonator photon lifetime τ_p.

(d) As the laser is pumped, the gain coefficient $\gamma(\nu_0)$ increases from its initial negative value at thermal equilibrium and changes sign, thereby providing gain. Determine the threshold population difference N_{t} for laser oscillation.

(e) Why is it advantageous to have the energy of level ③ close to that of level ②?

(f) How might the operation of the laser change if yttrium vanadate (YVO_4) were substituted for YAG ($Y_3Al_5O_{12}$) as the host material?

16.3-2 **Threshold Population Difference for an Ar^+-Ion Laser.** An Ar^+-ion laser has a 1-m-long resonator with 98% and 100% mirror reflectances. Other loss mechanisms are negligible. The atomic transition has a central wavelength $\lambda_o = 515$ nm, spontaneous lifetime $t_{\text{sp}} = 10$ ns, and linewidth $\Delta\lambda = 0.003$ nm. The lower energy level has a very short lifetime and hence zero population. The diameter of the oscillating mode is 1 mm. Determine

(a) the photon lifetime.

(b) the threshold population difference for laser action.

16.3-3 **Spontaneous Lifetime of an EUV Transition.** A visible laser transition at $\lambda_o = 500$ nm has a spontaneous lifetime $t_{\rm sp} = 10$ ns. Estimate the spontaneous lifetime for an EUV laser transition at $\lambda_o = 18.2$ nm, assuming that the transition strength S is the same in both cases. Compare your result with that provided in Table 15.3-1.

*16.4-4 **Transients in a Gain-Switched Laser.**

(a) Introduce the new variables $X = \mathfrak{n}/\tau_p$, $Y = N/N_{\rm t}$, and the normalized time $s = t/\tau_p$, to demonstrate that the rate equations (16.4-3) and (16.4-6) take the form

$$dX/ds = -X + XY \qquad \text{and} \qquad dY/ds = a(Y_0 - Y) - 2XY,$$

where $a = \tau_p/t_{\rm sp}$ and $Y_0 = N_0/N_{\rm t}$.

(b) Solve these two equations for both switching on and switching off. Assume that Y_0 is switched from 0 to 2 to turn the laser on, and from 2 to 0 to turn it off. Assume further that an initially very small photon flux corresponding to $X = 10^{-5}$ starts the oscillation at $t = 0$. Speculate on the possible origin of this flux. Determine the switching transient times for $a = 10^{-3}$, 1, and 10^3. Comment on the significance of your results.

*16.4-5 **Q-Switched Ruby Laser Power.** A Q-switched ruby laser makes use of a 15-cm-long rod of cross-sectional area 1 $\rm cm^2$ placed in a resonator of length 20 cm. The mirrors have reflectances $\mathcal{R}_1 = 0.95$ and $\mathcal{R} = 0.7$. The Cr^{3+} density is 1.58×10^{19} $\rm atoms/cm^3$, and the transition cross section $\sigma(\nu_0) = 2 \times 10^{-20}$ $\rm cm^2$. The laser is pumped to an initial population of 10^{19} $\rm atoms/cm^3$ in the upper state with negligible population in the lower state. The pump band (level ③) is centered at ≈ 450 nm and the decay from level ③ to level ② is fast. The lifetime of level ② is ≈ 3 ms.

(a) How much pump power is required to maintain the population in level ② at 10^{19} $\rm cm^{-3}$?

(b) How much power is spontaneously radiated before the Q-switch is operated?

(c) Determine the peak power, energy, and duration of the Q-switched pulse.

*16.4-6 **Operation of a Cavity-Dumped Laser.** Sketch the variation of the threshold population difference $N_{\rm t}$ (which is proportional to the loss), the population difference $N(t)$, the internal photon number density $\mathfrak{n}(t)$, and the external photon-flux density $\phi_o(t)$, during two cycles of operation of a pulsed cavity-dumped laser.

16.4-7 **Mode Locking with Lorentzian Amplitudes.** Assume that the envelopes of the modes of a mode-locked laser are given by

$$A_q = \sqrt{P}\,\frac{(\Delta\nu/2)^2}{(q\nu_F)^2 + (\Delta\nu/2)^2}, \qquad q = -\infty, \ldots, \infty,$$

and the phases are all equal. Determine expressions for the following parameters of the generated pulse train:

(a) Mean power

(b) Peak power

(c) Pulse duration (FWHM)

16.4-8 **Second-Harmonic Generation.** Crystals with nonlinear optical properties are often used for second-harmonic generation, as explained in Sec. 22.2A. In this process, two photons of frequency ν are converted into a single photon of frequency 2ν. Assume that such a crystal is placed inside a laser resonator with an active medium providing gain at frequency ν. The frequencies ν and 2ν correspond to two modes of the resonator. If the rate of second-harmonic conversion is $\zeta\mathfrak{n}$ (s^{-1}-m^{-3}) and the rate of photon production by the laser process (net effect of stimulated emission and absorption) is $\xi\mathfrak{n}$ ($\rm s^{-1}$-$\rm m^{-3}$), where ζ and ξ are constants, write the rate equations for the photon number densities $\mathfrak{n}$ and $\mathfrak{n}_2$ at the frequencies ν and 2ν. Assume that the photon lifetimes at ν and 2ν are τ_p and τ_{p2}, respectively. Determine the steady-state values of $\mathfrak{n}$ and $\mathfrak{n}_2$.

CHAPTER

17

SEMICONDUCTOR OPTICS

17.1 SEMICONDUCTORS 733

A. Energy Bands and Charge Carriers
B. Semiconductor Materials
C. Carrier Concentrations
D. Generation, Recombination, and Injection
E. Junctions
F. Heterojunctions
G. Quantum-Confined Structures

17.2 INTERACTIONS OF PHOTONS WITH CHARGE CARRIERS 766

A. Photon Interactions in Bulk Semiconductors
B. Interband Transitions in Bulk Semiconductors
C. Absorption, Emission, and Gain in Bulk Semiconductors
D. Photon Interactions in Quantum-Confined Structures
E. Quantum-Dot Single-Photon Emitters
F. Refractive Index

William B. Shockley (1910–1989), seated, **John Bardeen (1908–1991)**, center, and **Walter H. Brattain (1902–1987)**, right, shared the Nobel Prize in 1956 for demonstrating that semiconductor devices could be used to achieve amplification.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Photonics is the technology of controlling the flow of photons, much as electronics is the technology of controlling the flow of charge carriers (electrons and holes). These two technologies join together in semiconductor optics: photons generate mobile charge carriers, and charge carriers generate and control the flow of photons. Semiconductor devices serve as photon sources (light-emitting diodes and laser diodes), amplifiers, photodetectors, waveguides, modulators, multiplexers, sensors, and nonlinear optical elements. The compatibility of semiconductor optical devices and semiconductor electronic devices has fostered the development of both.

Semiconductor materials absorb and emit photons by undergoing transitions among allowed energy levels. Though the basic rules that govern these interactions are the same as those set forth for photons and atoms in Sec. 14.3, semiconductors have a number of unique features, as outlined in Sec. 14.1D:

- Because of the proximity of atoms in a crystal lattice, a semiconductor material should not be viewed as a collection of noninteracting atoms, each with its own individual energy levels. Rather, the energy levels belong to the system as a whole.
- Collections of closely spaced energy levels form energy bands. In the absence of external excitation, these bands are either fully occupied by electrons or totally unoccupied at $T = 0^\circ$ K. The highest-lying fully occupied energy band is known as the *valence band* while the lowest-lying unoccupied energy band is called the *conduction band*. These two bands are separated by a *forbidden band*, with bandgap energy E_g.
- An external energy source (whether thermal, optical, or electronic) can impart energy to an electron in the valence band, causing it to jump across the forbidden band and enter the conduction band. This transition leaves behind a vacancy (hole) in the valence band. In the inverse process, electron–hole recombination, an electron decays from the conduction band to fill an empty state in the valence band (provided that one is accessible), generating a photon and/or phonons in the process. Thus, photons interact with both types of charge carriers, electrons and holes.

Two processes are fundamental to the operation of most semiconductor optical devices:

1. *The absorption of a photon can create an electron–hole pair.* Mobile charge carriers resulting from the absorption of a photon alter the electrical properties of the semiconductor. This process is the basis of operation of photoconductive photodetectors.
2. *The recombination of an electron and a hole can result in the emission of a photon.* This process is responsible for the operation of semiconductor photon sources. Spontaneous radiative electron–hole recombination gives rise to photon generation in the light-emitting diode. Stimulated electron–hole recombination generates photons in a laser diode.

This Chapter

The reader is expected to be familiar with the basic principles of semiconductor physics. In Sec. 17.1 we offer a review of semiconductors and their properties. Section 17.2 provides an introduction to the optical properties of bulk and quantum-confined semiconductors. We present a simplified theory of absorption, spontaneous emission, and stimulated emission patterned on the approach to the interaction of photons with atoms provided in Sec. 14.3.

This and the following two chapters form a unit. Chapter 18 deals with the operation of semiconductor sources such as light-emitting diodes and laser diodes. Chapter 19 is devoted to semiconductor photodetectors.

17.1 SEMICONDUCTORS

As discussed in Sec. 14.1D, a semiconductor is a crystalline or amorphous solid whose electrical conductivity is typically intermediate between that of a metal and that of an insulator. Its conductivity can be significantly altered by modifying the temperature or doping concentration of the material, or by illuminating it with light. The band structure of semiconductors, and the ability to form junctions and heterostructures, offer unique properties. Quantum-confined semiconductor structures further extend the range of available features. As will be elucidated in Sec. 17.1B, semiconductor optical devices often rely on III–V ternary or quaternary compounds (e.g., InGaAsP, AlInGaP, or AlInGaN), but also make use of organic semiconductors and, increasingly, compounds forged from elements residing in group-IV of the periodic table (e.g., C, Si, Ge, and Sn). Electronic semiconductor devices are principally fabricated from Si.

A. Energy Bands and Charge Carriers

Energy Bands in Bulk Semiconductors

The atoms comprising solid-state materials have sufficiently strong interatomic interactions that they cannot be treated as individual entities (Sec. 14.1D). Their conduction electrons are not bound to individual atoms; rather, they belong to the collection of atoms as a whole. The solution of the Schrödinger equation for the electron energy, in the periodic potential created by the collection of atoms in the crystal lattice, results in a splitting of the atomic energy levels and the formation of energy bands. Each band contains a large number of densely packed discrete energy levels that is well approximated as a continuum. As illustrated in Fig. 17.1-1, the valence and conduction bands are separated by a forbidden band or bandgap. The **bandgap energy** E_g plays an important role in determining the electrical and optical properties of the material.

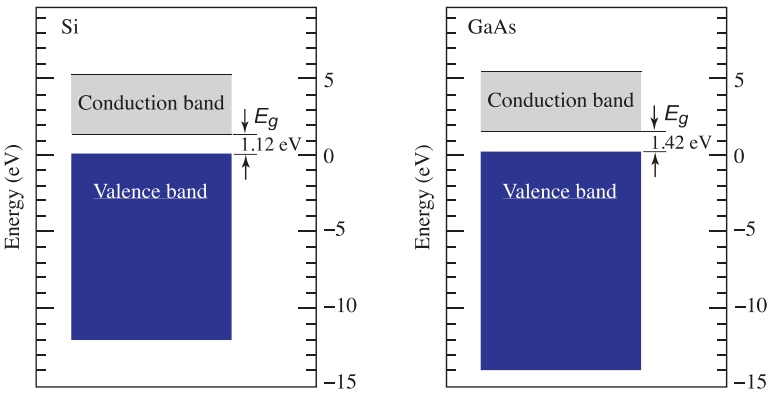


Figure 17.1-1 Energy bands in Si and GaAs. The bandgap energy E_g, which separates the valence and conduction bands, is 1.12 eV for Si and 1.42 eV for GaAs at room temperature.

The origin of the bandgap may be illustrated by means of the **Kronig–Penney model**. In this simple theory the crystal-lattice potential, a one-dimensional version of which is depicted in Fig. 17.1-2(a), is approximated by a 1D periodic rectangular-barrier potential, as shown in Fig. 17.1-2(b). The solution of the associated Schrödinger equation (14.1-3) for this potential yields allowed energy bands with traveling-wave solutions, separated by forbidden bands with exponentially decaying solutions. It can be shown that the results are general and apply to three dimensions. This approach is

similar to that used for analyzing the optics of one-dimensional periodic media, as set forth in Sec. 7.2 and discussed in Appendix C. The traveling-wave eigenfunctions are **Bloch modes** with the periodicity of the crystal lattice [see (7.2-4)].

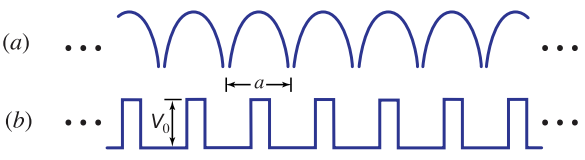


Figure 17.1-2 (a) Crystal-lattice potential associated with an infinite one-dimensional collection of atoms with lattice constant a. (b) Idealized rectangular-barrier potential (height V_0) for the Kronig–Penney model.

Electrons and Holes

As is understood from Sec. 14.1, the wavefunctions of the electrons in a semiconductor overlap so that the **Pauli exclusion principle** applies. This principle dictates that no two electrons may occupy the same quantum state and that the lowest available energy levels fill first. Elemental semiconductors, such as Si and Ge, have four valence electrons per atom that form covalent bonds. At $T = 0^\circ$ K, the number of quantum states that can be accommodated in the valence band is such that it is completely filled while the conduction band is completely empty. The material cannot conduct electricity under these conditions.

As the temperature increases, however, some electrons can be thermally excited from the valence band into the empty conduction band, where unoccupied states are abundant (Fig. 17.1-3). These electrons can then act as mobile carriers, drifting through the crystal lattice under the effect of an applied electric field, and thereby contributing to the electric current. Moreover, an electron departing from the valence band leaves behind an unoccupied quantum state, which in turn allows the remaining electrons in the valence band to exchange places with each other under the influence of an external field. The collection of electrons remaining in the valence band thus undergoes motion. This can equivalently be regarded as motion, in the opposite direction, of the hole left behind by the departed electron. The hole therefore behaves as a particle with positive charge $+e$.

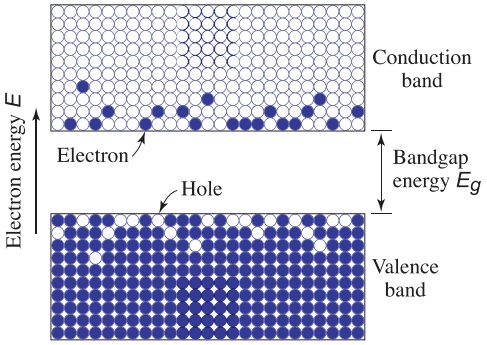


Figure 17.1-3 Electrons in the conduction band and holes in the valence band at $T > 0^\circ$ K.

The net result is that each electron excitation creates a free electron in the conduction band and a free hole in the valence band. The two charge carriers are free to drift under the effect of the applied electric field and thereby to generate an electric current. The material behaves as a *semi*conductor whose conductivity increases sharply with increasing temperature, as an increasing number of mobile carriers are thermally generated.

Energy–Momentum Relations

In accordance with Schrödinger wave mechanics, the energy E and momentum $\mathbf{p}$ of an electron in a region of constant potential, such as free space, are related by $E = p^2/2m_0 = \hbar^2 k^2/2m_0$, where p is the magnitude of the momentum, k is the magnitude of the wavevector $\mathbf{k} = \mathbf{p}/\hbar$, and m_0 is the electron mass (9.1×10^{-31} kg). The E–k relation for a free electron is thus a simple parabola.

EXERCISE 17.1-1

Energy–Momentum Relation for a Free Electron.

(a) Consider a one-dimensional version of the time-independent Schrödinger equation set forth in (14.1-3) for a free electron ($V = 0$) of mass m_0. Use a trial solution of the form $\psi(x) \propto \exp(-jkx)$ to show that the energy–momentum relation assumes the *quadratic* form

$$E = \frac{\hbar^2 k^2}{2m_0}, \tag{17.1-1}$$

so that the electron energy is not quantized in this example.

(b) The free photon, in contrast, has the *linear* energy–momentum relation provided in (13.1-11),

$$E = pc = c\hbar k, \tag{17.1-2}$$

where c is the speed of light in the medium. What is the origin and significance of this distinction?

The motion of an electron in a semiconductor material is similarly governed by the Schrödinger equation, but with a potential generated by the charges in the periodic crystal lattice of the material. As discussed earlier, this construct results in allowed energy bands separated by forbidden bands, as exemplified by the Kronig–Penney model. The ensuing E–k relations for electrons and holes, in the conduction and valence bands respectively, are illustrated in Fig. 17.1-4 for Si and GaAs. The energy E is a periodic function of the components (k_1, k_2, k_3) of the wavevector $\mathbf{k}$, with periodicities $(\pi/a_1, \pi/a_2, \pi/a_3)$, where a_1, a_2, a_3 are the crystal lattice constants. Figure 17.1-4 displays cross sections of this relation along two particular directions of the wavevector $\mathbf{k}$. The range of k values in the interval $[-\pi/a, \pi/a]$ defines the first **Brillouin zone**. The energy of an electron in the conduction band thus depends not only on the magnitude of its momentum, but also on the direction in which it is traveling in the crystal. The semiconductor E–k diagram bears some resemblance to the photonic-crystal ω-K diagram (Fig. 7.3-5).

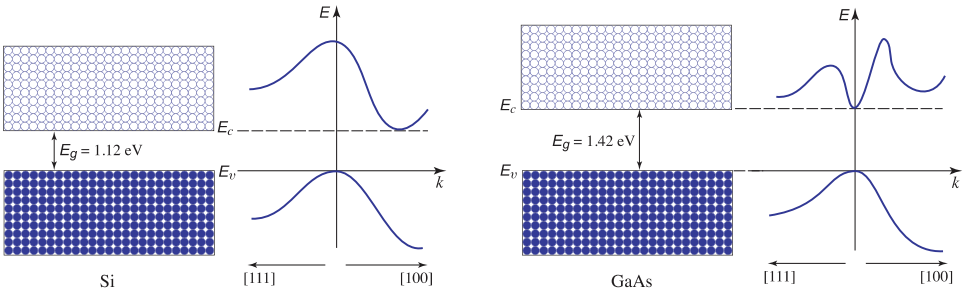


Figure 17.1-4 Cross sections of the E–k relations for Si and GaAs along two crystal directions: [111] toward the left and [100] toward the right.

Effective Mass

It is apparent from Fig. 17.1-4 that near the bottom of the conduction band, for both Si and GaAs, the E–k relation may be approximated by a parabola

$$E = E_c + \frac{\hbar^2 k^2}{2m_c}, \tag{17.1-3}$$

where E_c is the energy at the bottom of the conduction band and k is measured from the value of the wavevector where the minimum occurs. This parabolic relation suggests that a conduction-band electron behaves in a manner analogous to that of a free electron, but with a mass m_c, known as the conduction-band effective mass or the **electron effective mass**, in place of the free-electron mass m_0. The effective mass m_c embodies the influence of the ions of the lattice on the motion of a conduction-band electron. This behavior is highlighted in Fig. 17.1-5.

Similarly, near the top of the valence band, we may write

$$E = E_v - \frac{\hbar^2 k^2}{2m_v}, \tag{17.1-4}$$

where $E_v = E_c - E_g$ is the energy at the top of the valence band and m_v is the valence-band effective mass or the **hole effective mass**, as illustrated in Fig. 17.1-5. The influence of the lattice ions on the motion of a valence-band hole is captured by its effective mass m_v. The effective mass also depends on the particular band under consideration. Indeed, several parabolas of different curvature often coexist near the top of the valence band, corresponding to so-called *heavy holes*, *light holes*, and *split-off-band holes*.

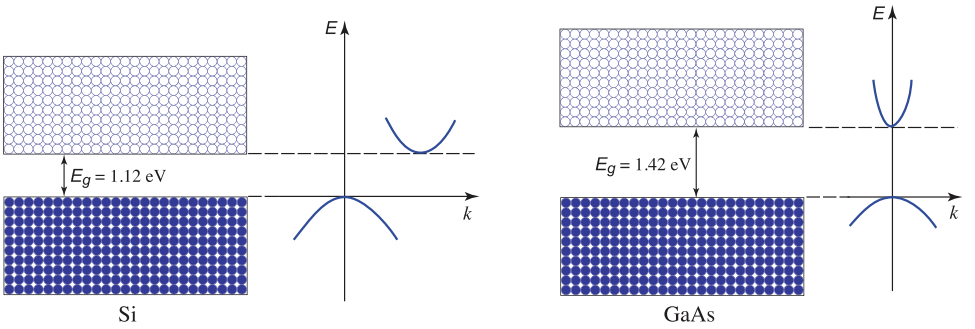


Figure 17.1-5 The E–k relation is well-approximated by parabolas at the bottom of the conduction band and at the top of the valence band, for both Si and GaAs.

The effective mass depends on the crystal structure of the material and on the direction of travel with respect to the lattice since the interatomic spacing varies with crystallographic direction. Typical averaged effective masses, normalized to the free-electron mass m_0, are provided in Table 17.1-1 for Si, GaAs, and GaN.

Table 17.1-1 Typical averaged values of the normalized electron and hole effective masses in selected semiconductor materials.

	m_c/m_0	m_v/m_0
Si	0.98	0.49
GaAs	0.07	0.50
GaN	0.20	0.80

Direct- and Indirect-Bandgap Semiconductors

Semiconductors for which the conduction-band minimum energy and the valence-band maximum energy correspond to the same value of the wavenumber k (same momentum) are called **direct-bandgap** materials. Semiconductors for which this is not the case are known as **indirect-bandgap** materials. As is evident from Fig. 17.1-5, GaAs is a direct-bandgap semiconductor whereas Si is an indirect-bandgap semiconductor. The distinction is important because a transition between the bottom of the conduction band and the top of the valence band in an indirect-bandgap semiconductor must accommodate a substantial change in the momentum of the electron, requiring the participation of a third body such as a phonon. As a consequence, under ordinary circumstances indirect-bandgap semiconductors such as Si cannot serve as efficient light emitters whereas direct-bandgap semiconductors such as GaAs can, as will be elucidated subsequently.

B. Semiconductor Materials

Figure 17.1-6 reproduces the section of the periodic table that comprises the elements important in semiconductor electronics and photonics. Both elemental and compound semiconductors play crucial roles in these technologies. We discuss several classes of these materials in turn and then consider doped semiconductors.

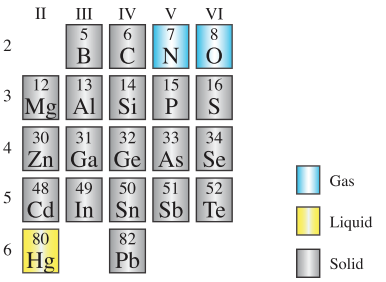
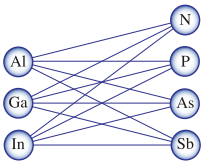


Figure 17.1-6 Section of the periodic table relating to semiconductors. Each column designation should, strictly speaking, have "A" appended to it, so that II represents IIA, etc. The full periodic table is displayed in Fig. 14.1-3. Elements indicated in blue, yellow, and silver take the form of gases, liquids, and solids, respectively, at room temperature.

Elemental Semiconductors

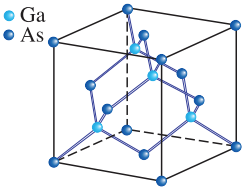
Silicon (Si) and germanium (Ge) are important elemental semiconductors in group IV of the periodic table. Both find widespread use in photonics, although they have traditionally not been used as light emitters because of their indirect bandgaps (their basic properties are provided in Table 17.1-2). However, Ge has been shown to behave as a direct-bandgap material and to emit light under special circumstances. Silicon is used for fabricating virtually all commercial electronic integrated circuits and serves a wide variety of functions in photonics under the rubric *silicon photonics*. Group-IV elements can be alloyed to form compound semiconductors with a broad array of uses. For example, the indirect-bandgap, binary alloy silicon carbide (SiC), also known as carborundum, is useful for fabricating ultraviolet photodetectors and as a template for III–nitride photon emitters. Silicon-germanium (Si_xGe_{1-x}) finds application in both photonics and electronics, including use as an infrared photodetector material. Germanium-tin ($Ge_{1-y}Sn_y$) is suitable for fabricating photodetectors, as well as laser diodes and LEDs. Useful ternary and quaternary group-IV semiconductor compounds include $Si_xGe_{1-x-y}Sn_y$ and $Si_xGe_{1-x-y-z}Sn_yC_z$, respectively. The use of alloys and combinations of group-IV elements is an emerging area of photonics that has come to be called *group-IV photonics*.

Binary III–V Semiconductors



Compounds formed by combining an element in column III, such as aluminum (Al), gallium (Ga), or indium (In), with an element in column V, such as nitrogen (N), phosphorus (P), arsenic (As), or antimony (Sb), are important semiconductors in photonics. These twelve III–V compounds are listed in Table 17.1-2, along with their crystal structure (zincblende or wurtzite), bandgap type (direct or indirect), bandgap energy E_g, and bandgap wavelength $\lambda_g = hc_o/E_g$ (the free-space wavelength of a photon of energy E_g). The bandgap energies and lattice constants of these compounds are also displayed in Fig. 17.1-7. Photon sources (light-emitting diodes and lasers) and photodetectors can be readily fabricated from many of these binary compounds. The first of the binary semiconductors to find use in photonics was gallium arsenide (GaAs), which is also sometimes used as an alternative to Si for fast electronic devices and circuits. Gallium nitride (GaN) plays a central role in photonics by virtue of its near-ultraviolet bandgap wavelength; it is also important in electronics because of its ability to withstand high temperatures. AlN, an insulator by virtue of its large bandgap energy, emits photons in the vicinity of $\lambda_o = 210$ nm in the mid ultraviolet.

Zincblende and Diamond



Wurtzite

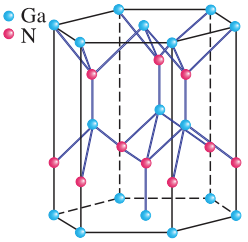


Table 17.1-2 Selected elemental and binary III–V semiconductors along with their crystal structures, bandgap types, bandgap energies, and bandgap wavelengths.

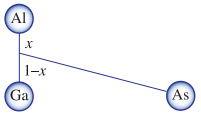
Material	Crystal Structure[a] (D/Z/W)	Bandgap Type[b] (I/D)	Bandgap Energy[c] E_g (eV)	Bandgap Wavelength[d] λ_g (μm)
Si	D	I	1.12	1.11
Ge	D	I	0.66	1.88
AlN	W	D	6.02	0.206
AlP	Z	I	2.45	0.506
AlAs	Z	I	2.16	0.574
AlSb	Z	I	1.58	0.785
GaN	W	D	3.39	0.366
GaP	Z	I	2.26	0.549
GaAs	Z	D	1.42	0.873
GaSb	Z	D	0.73	1.70
InN	W	D	0.65	1.91
InP	Z	D	1.35	0.919
InAs	Z	D	0.36	3.44
InSb	Z	D	0.17	7.29

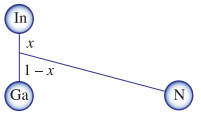
[a]The crystal structure listed above indicates the most commonly used form of the material: D = Diamond, Z = Zincblende, W = Wurtzite. The zincblende structure comprises two interpenetrating face-centered-cubic lattices, one for each element, displaced from each other by $^1/_4$ of the body diagonal. The diamond lattice is the same as zincblende except that all atoms are identical. The Brillouin zone for these structures is illustrated in Fig. 7.3-4. The wurtzite structure consists of two hexagonal close-packed lattices, one for each element, displaced from each other along the three-fold c axis by $^3/_8$ of its length. All atoms are tetrahedrally bonded with their neighbors.

[b]I = Indirect bandgap; D = Direct bandgap. [c]At $T = 300^\circ$ K.

[d]The bandgap wavelength λ_g is related to the bandgap energy E_g by $\lambda_g = hc_o/E_g$; when the bandgap energy is expressed in eV and the bandgap wavelength is expressed in μm, this relation can be expressed as $\lambda_g \approx 1.24/E_g$.

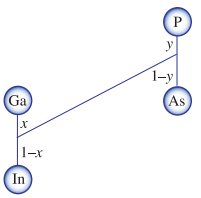
Ternary III–V Semiconductors





Compounds formed from two elements of column III with one element from column V (or one from column III with two from column V) are important ternary semiconductors. $Al_xGa_{1-x}As$, for example, is a compound with properties that interpolate between those of AlAs and GaAs, depending on the compositional mixing ratio x (the fraction of Ga atoms in GaAs that are replaced by Al atoms). The bandgap energy E_g for this material varies between 1.42 eV for GaAs and 2.16 eV for AlAs, as x varies between 0 and 1 along the line connecting GaAs and AlAs in Fig. 17.1-7(a). Because this line is essentially vertical, $Al_xGa_{1-x}As$ is lattice matched to GaAs; a layer of arbitrary composition of this material can therefore be grown on a layer of different composition without straining the lattice. Other useful III–V ternary compounds, such as $GaAs_{1-x}P_x$, are also represented in the bandgap-energy versus lattice-constant diagram displayed in Fig. 17.1-7(a). $In_xGa_{1-x}As$ is widely used for photon sources and detectors in the near-infrared region of the spectrum. Similarly, $Al_xGa_{1-x}N$ and $In_xGa_{1-x}N$ are important ternary semiconductors for photonic devices that operate in the ultraviolet, violet, blue, and green regions of the spectrum, as can be deduced from Fig. 17.1-7(b). In the domain of electronics, $In_xGa_{1-x}As$/InP heterojunction bipolar transistors can both emit light and be switched at high speeds.

Quaternary III–V Semiconductors



These compounds are formed by mixing two elements from column III with two elements from column V (or three from column III with one from column V). Quaternary semiconductors offer more flexibility for fabricating materials with desired properties than do ternary semiconductors by virtue of their additional degree of freedom. An example is provided by $In_{1-x}Ga_xAs_{1-y}P_y$, whose bandgap energy varies between 0.36 eV (InAs) and 2.26 eV (GaP) as the compositional mixing ratios x and y vary between 0 and 1. The lattice constant usually varies linearly with the mixing ratio (Vegard's law). The stippled area in Fig. 17.1-7(a) indicates the range of bandgap energies and lattice constants spanned by this compound. For mixing ratios x and y that satisfy $y = 2.16(1 - x)$, $In_{1-x}Ga_xAs_{1-y}P_y$ can be lattice matched to InP, which can therefore serve as a convenient template (substrate). This quaternary compound is used for fabricating light-emitting diodes, laser diodes, and photodetectors, particularly in the vicinity of the 1550-nm optical fiber communications wavelength (Chapters 18, 19, and 25). Another example is provided by $Al_xIn_yGa_{1-x-y}P$, for which GaAs serves as a template; this compound offers high-brightness emission in the red, orange, and yellow spectral regions [shaded region in Fig. 17.1-7(a)]. Yet another important quaternary material is the III–nitride compound $Al_xIn_yGa_{1-x-y}N$, which serves the green, blue, violet, and ultraviolet spectral regions in the same way [Fig. 17.1-7(b)]. Convenient templates for the III–nitrides are sapphire, SiC, and Si.

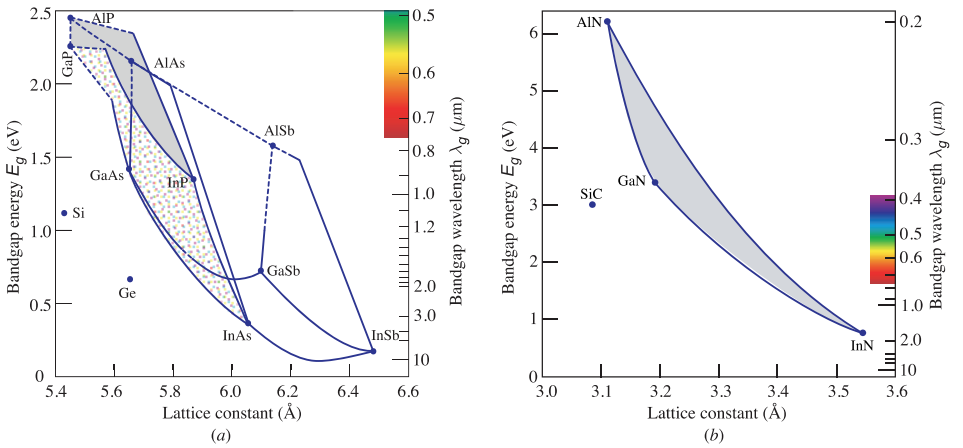


Figure 17.1-7 Dots represent bandgap energies, bandgap wavelengths, and lattice constants for Si, Ge, SiC, and 12 binary III–V compounds. Solid and dashed curves represent direct-bandgap and indirect-bandgap compositions, respectively. A material may have a direct bandgap for one mixing ratio and an indirect bandgap for a different mixing ratio. Ternary materials are represented along the line that joins two binary compounds. A quaternary compound is represented by the area formed by its binary components. (*a*) $In_{1-x}Ga_xAs_{1-y}P_y$ is represented by the stippled area with vertices at InP, InAs, GaAs, and GaP, while $(Al_xGa_{1-x})_yIn_{1-y}P$ is represented by the shaded area with vertices at AlP, InP, and GaP. Both are important quaternary compounds, the former in the near infrared and the latter in the visible. $Al_xGa_{1-x}As$ is represented by points along the line connecting GaAs and AlAs. As x varies from 0 to 1, the point moves along the line from GaAs and AlAs. Since this line is nearly vertical, $Al_xGa_{1-x}As$ is lattice matched to GaAs. (*b*) Though the III–nitride compound $In_xGa_{1-x}N$ can, in principle, be compositionally tuned to accommodate the entire visible spectrum, this material becomes increasingly difficult to grow as the In composition becomes appreciable. $In_xGa_{1-x}N$ is principally used in the green, blue, and violet spectral regions, while $Al_xGa_{1-x}N$ and $Al_xIn_yGa_{1-x-y}N$ serve the ultraviolet region. All compositions of these III–nitride compounds are direct-bandgap semiconductors.

Binary and Ternary II–VI Semiconductors

Binary II–VI materials, i.e., compounds formed from elements in column II (e.g., Zn, Cd, Hg) and chalcogenide elements in column VI (e.g., S, Se, Te) of the periodic table are also useful semiconductors. This family includes ZnS, ZnSe, ZnTe, CdS, CdSe, CdTe, HgS, HgSe, and HgTe, as displayed in Fig. 17.1-8. Unlike the III–V alloys, the II–VI compounds are widely found in nature. All of these materials have a zincblende structure and are direct-bandgap semiconductors, with the exception of HgSe and HgTe, which are semimetals with small negative bandgaps. A particular merit of ZnSe is that it can be deposited on a GaAs substrate with a relatively low defect density since the lattice constants of the two materials are similar. The ternary II–VI semiconductor $Hg_xCd_{1-x}Te$ can be grown without strain on a CdTe substrate since HgTe and CdTe are nearly lattice matched. This material system is widely used for photodetectors, as are other II–VI compounds (Chapter 19). However, photon sources fabricated from these materials are rarely used since they suffer from limited lifetimes. Notwithstanding, binary II–VI semiconductor materials such as CdSe are readily fashioned into quantum dots with tunable photoluminescence emission wavelengths (see, e.g., Fig. 14.1-13).

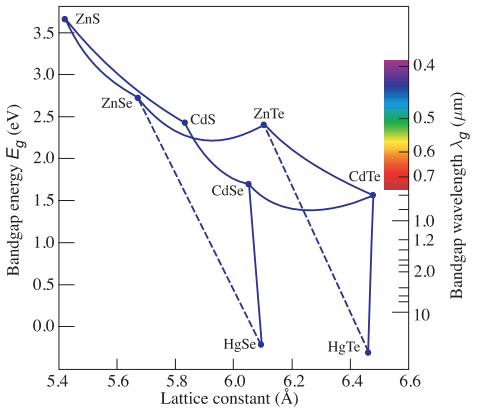


Figure 17.1-8 Bandgap energies, bandgap wavelengths, and lattice constants for various II–VI semiconductors (HgSe and HgTe are semimetals with small negative bandgaps). HgTe and CdTe are nearly lattice matched, as evidenced by the vertical line connecting them, so that the ternary semiconductor $Hg_xCd_{1-x}Te$ can be grown without strain on a CdTe template. It is an important mid-infrared photodetector material.

Ternary IV–VI Semiconductors

Ternary IV–VI compounds such as $Pb_xSn_{1-x}Te$ and $Pb_xSn_{1-x}Se$ have also been used as laser diodes and infrared photodetectors. As photodetectors, however, these alloys have long RC response times because of their large relative permittivities. Moreover, cycling between room and cryogenic temperatures can also be problematic because of their high thermal-expansion coefficients.

Doped Semiconductors

The electrical and optical properties of semiconductors can be modified substantially by the controlled introduction into the material of small amounts of specially chosen impurities called **dopants**. The introduction of these impurities can alter the concentration of mobile charge carriers by many orders of magnitude. Dopants with excess valence electrons, called **donors**, replacing a small proportion of the normal atoms in the crystal lattice, create a predominance of mobile electrons. The material is then said to be an ***n*-type** semiconductor. Thus, atoms from column V (e.g., P or As) replacing a fraction of the column-IV atoms in an elemental semiconductor (e.g., Si or Ge), or atoms from column VI (e.g., Se or Te) replacing a small fraction of the column-V atoms in a III–V binary semiconductor (e.g., As or Sb), produce an n-type material.

Similarly, a ***p*-type** semiconductor is made by using dopants with a deficiency of valence electrons, called **acceptors**. The result is then a predominance of mobile holes. Column IV atoms in an elemental semiconductor replaced with a small proportion of column-III atoms (e.g., B or In), or column-III atoms in a III–V binary semiconductor replaced with a small proportion of column-II atoms (e.g., Zn or Cd), yield p-type material. Column-IV atoms act as donors for column III and as acceptors for column V, and therefore can be used to produce an excess of both electrons and holes in III–V materials. Of course, the charge neutrality of the material is not altered by the introduction of dopants.

Semiconductors can also be doped with impurities that have the same valence as a constituent of the crystal lattice. Rather than introducing excess carriers, such substitutional doping can create a material that acts as a solid-state laser medium. For example, a transition-ion-doped zinc-chalcogenide laser can be formed by introducing Cr^{2+} ions into ZnS to substitute for a fraction of the Zn^{2+} ions comprising the lattice (Sec. 16.3A).

Undoped semiconductors (i.e., semiconductors devoid of intentional doping) are referred to as **intrinsic** materials, whereas doped semiconductors are called **extrinsic** materials. The concentrations of mobile electrons and holes are equal in an intrinsic

semiconductor, $\mathrm{n} = \mathrm{p} = \mathrm{n_i}$, where the intrinsic concentration $\mathrm{n_i}$ grows with increasing temperature at an exponential rate. On the other hand, the concentration of mobile electrons in an n-type semiconductor (**majority carriers**) is far greater than the concentration of holes (**minority carriers**), i.e., $\mathrm{n} \gg \mathrm{p}$. The opposite is true in a p-type semiconductor, where holes are the majority carriers, and $\mathrm{p} \gg \mathrm{n}$. A doped semiconductor at room temperature typically has a majority-carrier concentration that is approximately equal to the doping concentration.

As semiconductor devices shrink in scale, their characteristics are determined by ever smaller numbers of dopant atoms that are randomly distributed in position. At the nanoscale, the average number of dopants shrinks to a handful. However, techniques such as single-ion implantation can be used to fabricate semiconductor materials in which the number of dopant atoms, and their positions, are precisely determined, thereby offering improved control over device behavior. Nowadays, semiconductor materials such as Si and Ge can be grown with sufficient purity that a nanodevice can be totally devoid of impurities, thereby permitting a solitary dopant to be inserted at a specified position.

EXAMPLE 17.1-1. ***Donor-Electron Ionization Energy.*** Consider a germanium crystal of relative permittivity $\epsilon/\epsilon_o = 16$ (Table 17.2-1) doped with arsenic donor atoms. The electron effective mass $m_c = 0.2\,m_0$, where m_0 is the free electron mass. The donor electron moves in the field of the singly charged arsenic ion (As^+), and has energy levels similar to those of an electron in the hydrogen atom. Choosing $\mathrm{n} = 1$ and $Z = 1$ in (14.1-4), and replacing ϵ_o by ϵ, and M_r by m_c, to accommodate the polarization density and crystal lattice of the semiconductor material, respectively, the energy of the donor electron is given by

$$E_D = -\left(\frac{1}{4\pi\epsilon}\right)^2 \frac{m_c e^4}{2\hbar^2}. \tag{17.1-5}$$

Since the energy of the electron in the ground state of hydrogen is ≈ -13.6 eV with respect to the vacuum level (i.e., it is 13.6 eV below the ionization energy), the energy of the arsenic donor electron is $E_D = -(m_c/m_0)(\epsilon_o/\epsilon)^2 \times 13.6$ eV ≈ -0.01 eV. The donor electron thus resides in the forbidden band, at a level ≈ 0.01 eV below the conduction band edge. Since the thermal energy $kT \approx 0.026$ eV at $T = 300°$ K, however, essentially all of the donors are ionized at room temperature and the donor electrons are elevated to the conduction band. The material thus has a conduction-band donor concentration that matches the impurity concentration.

Organic Semiconductors

Organic semiconductors are increasingly employed in electronics and photonics, where they are used in the form of photovoltaic devices, light-emitting diodes, and high-quality organic light-emitting displays. Though they generally do not offer the speed of inorganic semiconductor structures, they can be inexpensively fabricated in the form of thin sheets, making low-cost, mechanically flexible optoelectronic components available. These materials can be engineered to suit specific requirements and can sometimes be printed on a suitable substrate, such as plastic, using inkjet technology.

Organic semiconductors are available in two principal varieties, as illustrated schematically in Fig. 17.1-9:

1. Small organic molecules such as pentacene, which consists of five linearly joined benzene rings [Fig. 17.1-9(a)].
2. Conjugated polymer chains such as polyacetylene, comprising hundreds or thousands of carbon atoms [Fig. 17.1-9(b)].

A hallmark of these amorphous materials, termed conjugation, is their alternating single and double carbon–carbon bonds. Though the double-bond electrons shown in Figs. 17.1-9(a) and (b) are portrayed as belonging to particular atoms, these electrons

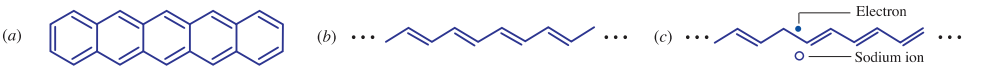


Figure 17.1-9 Organic semiconductors are available in two principal varieties: (a) small organic molecules such as pentacene, and (b) conjugated polymer chains such as polyacetylene. (c) Doping polyacetylene with sodium donors yields an n-type material, whereas doping with iodine acceptors yields a p-type material. Each vertex represents a carbon atom and each line represents a bond between two carbon atoms; double lines represent double bonds. Hydrogen bonds are omitted for simplicity. A wide variety of organic molecules and polymers are used in photonics and electronics.

are actually delocalized and shared among multiple atoms, or along a segment of polymer comprising roughly ten repeat units. The molecule, or polymer segment, behaves as a single system in which the allowed electron states form bands.

In its undoped state, the valence band of a conjugated polymer chain is typically full, and its conduction band empty, so that it behaves as an insulator. However, as illustrated in Fig. 17.1-9(c), dopants such as sodium and iodine act as donors and acceptors, respectively, providing n-type and p-type variants. Small organic molecules are often conductive in their pure state.

A number of fundamental features distinguish organic semiconductors from their inorganic cousins:

- The constituent molecules are bound by weak van der Waals forces (bond energy ≈ 0.01 eV) whereas the atoms in inorganic semiconductors are bound by strong covalent bonds (bond energy ≈ 3 eV).
- Weak intermolecular bonds offer mechanical flexibility whereas inorganic semiconductors are rigid.
- The energy bands derive from localized behavior at the molecular level whereas in inorganic semiconductors they derive from the collection of atoms as a whole.
- The two energy levels that play key roles are the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), whereas in inorganic semiconductors the conduction and valence bands are of paramount importance.
- Charge carriers have high effective mass ($m_c/m_0 \approx 500$) and low mobility ($\mu \approx 10^{-3}$ cm^2/V·s) whereas carriers in inorganic semiconductors have low effective mass ($m_c/m_0 < 1$) and high mobility ($\mu \approx 10^3$ cm^2/V·s).
- Intermolecular electronic transfer occurs via phonon-assisted tunneling (hopping) whereas electronic transport in inorganic semiconductors is characterized by drift and diffusion.
- Organic semiconductors generally exhibit low electrical conductivity and are sensitive to moisture whereas the reverse is true for inorganic semiconductors.

Graphene and 2D Materials

As indicated at the beginning of this section, the semiconductor elements residing in group-IV of the periodic table are of substantial interest in photonics. Most important among these are carbon (C), silicon (Si), germanium (Ge), and tin (Sn). Group-IV elements exist in various structural forms, known as *allotropes*, that have different properties and applications. Among the most widely known allotropes of carbon, for example, are diamond, graphite, carbon nanotubes, carbon dots, and graphene (there are others). Graphene, a material with a one-atom-thick carbon honeycomb lattice, has come to the fore in recent years because of its unique properties and because it can be fashioned into various photonic devices. Elemental group-IV analogs of graphene include silicene, germanene, and stanene. These hexagonal-lattice 2D atomic sheets are often denoted h-C, h-Si, h-Ge, and h-Sn, respectively, where the designation 'h' represents 'hexagonal.' The nascent fields of **graphene photonics** and **2D-material photonics** lie under the rubric of **group-IV photonics**.

Graphene. Graphene is a 2D material comprising a single 0.33-nm-thick layer of graphite with atoms arranged in a hexagonal honeycomb structure (Fig. 17.1-10). Graphene is endowed with a collection of exceptional properties that make it useful in many photonics applications:

- It is an excellent conductor of electricity and has an optical transmittance near unity so it can be used as a transparent electrode. Its optical absorbance is nearly constant at $\mathcal{A} = \pi e^2/\hbar c \approx 2.3\%$ over a broad wavelength band that stretches from 0.7 to 25 μm; its reflectance is a negligible $\mathcal{R} \approx 1.3 \times 10^{-4}$; and its transmittance at normal incidence is $\mathcal{T} \approx 97.7\%$. Moreover, its current-carrying capacity is substantial ($\approx 10^8$ A/cm^2 on SiO_2).
- It is a semimetal with zero bandgap that can interact with radiation over a broad spectral range stretching from the THz to the ultraviolet. Its absorption coefficient $\alpha \approx 7 \times 10^5$ cm^{-1} is an order of magnitude greater than that of Si or GaAs. It is readily doped, so that its electronic properties can be altered.
- It has an unusually high electron mobility. When deposited on SiO_2, its mobility is $\approx 1.5 \times 10^4$ cm^2/V·s so that the drift velocity of carriers is an order of magnitude greater than that in Si, as indicated in (19.1-9). It therefore has an inordinately fast response and is suitable for use in ultrafast photodetectors. Its high area-to-volume ratio makes it highly effective for applications involving sensing.
- It is chemically stable, refractory to high temperatures, and resilient in high humidity. It has high thermal conductivity and excellent mechanical strength, yet is elastic and therefore bendable.
- It exhibits fast and strong absorption saturation, rendering it suitable for use as a saturable absorber for mode-locked lasers and as a broadband modulator.

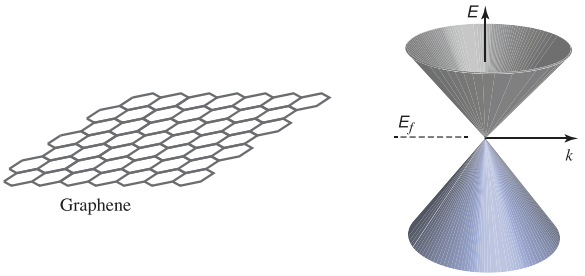


Figure 17.1-10 Graphene, also referred to as h-C, is a single layer of carbon atoms arranged in a hexagonal honeycomb lattice. Its E–k diagram is conical rather than parabolic (compare with Fig. 17.1-5). Graphene behaves as a semimetal with zero bandgap since its conduction- and valence-band cones meet at points that define the Fermi level E_f.

Because of its particular 2D symmetry, the band structure for carriers in graphene takes the form of cones (Fig. 17.1-10), rather than the parabolas that are characteristic of traditional semiconductors (Fig. 17.1-5). The E–k diagram is therefore linear rather than parabolic; it is similar to that for photons and is characterized by (17.1-2) rather than by (17.1-1). As with photons, the electronic excitations (called Dirac fermions) behave as if they were massless; this leads to an unusually large Fermi velocity, $v \approx c/300$, that underlies graphene's fast response. Furthermore, the conduction- and valence-band cones meet at single points (called Dirac points) that define the Fermi level, so that graphene behaves as a semimetal with zero bandgap. Several other 2D materials also host massless Dirac fermions and behave as semimetals (e.g., silicene, germanene, stanene, and β_{12}-borophene), but most 2D materials have approximately parabolic, rather than conical, band structures. Though Dirac fermions have been most widely studied in 2D materials, they are also hosted by 3D materials such as compressively strained α-Sn (gray tin) and Na_3Bi.

Though the interaction of light with graphene is strong on a per-unit-distance basis ($\alpha \approx 7 \times 10^5$ cm^{-1}), devices that rely on single-pass operation encounter an insignificant thickness of material (0.33 nm). Building an effective 2D-based device thus generally requires that the interaction be enhanced, which may be achieved by specialized doping or siting, by coupling to a photonic waveguide or cavity, or by coupling to plasmons, phonons, or excitons. Significant enhancement of the light–matter interaction can be attained by making use of traveling surface plasmon polaritons.

Other 2D materials. By virtue of its semimetallic nature, graphene is a poor emitter of light. However, a number of other 2D materials, including various transition-metal dichalcogenides (TMDs) such as molybdenum disulfide, behave as direct-bandgap semiconductors with bandgap energies E_g that lie between 0.5 and 3 eV. As with 3D semiconductors, the bandgap energy can be tuned via chemistry, composition, and/or quantum confinement. These materials can serve as light emitters or reflectors dominated by excitonic transitions.

Single-layer TMDs such as MoS_2 and WSe_2 consist of a sublayer of transition metal sandwiched between two sublayers of chalcogen. MoS_2, for example, has an overall layer thickness of 0.65 nm and a bandgap energy of 1.8 eV. In their 3D configurations, some of these materials (e.g., graphite, MoS_2) serve as industrial lubricants. This is because consecutive atomic layers are bound only by weak van der Waals forces and easily slide over each other, a property that has made it relatively easy to peel off individual 2D layers. Indeed, such 2D materials are often called **van der Waals materials**. Other 3D precursors (e.g., silicon, germanium) form tight bonds in all three dimensions so that their 2D versions, when extracted, tend to buckle. The number of possible TMDs that can be formed is substantial since there are tens of transition metals and, as is apparent in column VIA of the periodic table (Fig. 14.1-3), at least three chalcogens (S, Se, and Te; the elements O, Po, and Lv are sometimes also included in this category). Some single-layer materials behave as insulators (e.g., hexagonal BN, with $E_g \approx 6$ eV) and others behave as metals (e.g., TiS_2). 2D materials can be used in isolation, or combined in layers of various compositions, to create atomically thin heterostructures that serve as planar photonic devices.

C. Carrier Concentrations

Determining the concentration of carriers (electrons and holes) as a function of energy requires knowledge of two features that we consider in turn:

- The density of allowed energy levels (density of states)
- The probability that each of these levels is occupied

Density of States

The quantum state of an electron in a semiconductor material is characterized by its energy E, its wavevector $\mathbf{k}$ [the magnitude of which is approximately related to E by (17.1-3) or (17.1-4)], and its spin. The state is described by a wavefunction that satisfies certain boundary conditions.

An electron near the conduction band edge may be approximately described as a particle of mass m_c confined to a three-dimensional cubic box (of dimension d) with perfectly reflecting walls, i.e., a three-dimensional infinite rectangular potential well. The standing-wave solutions require that the components of the vector $\mathbf{k} = (k_x, k_y, k_z)$ assume the discrete values $\mathbf{k} = (q_1\pi/d,\ q_2\pi/d,\ q_3\pi/d)$, where the respective mode numbers (q_1, q_2, q_3) are positive integers. This result is a three-dimensional generalization of the one-dimensional infinite square well (Exercise 17.1-5). The tip of the vector $\mathbf{k}$ must lie on the points of a lattice whose cubic unit cell has dimension π/d. There are therefore $(d/\pi)^3$ points per unit volume in $\mathbf{k}$-space. The number of states whose vectors $\mathbf{k}$ have magnitudes between 0 and k is determined by counting

the number of points lying within the positive octant of a sphere of radius k [with volume $\approx (\frac{1}{8})4\pi k^3/3 = \pi k^3/6$]. Because of the two possible values of the electron spin, each point in $\mathbf{k}$-space corresponds to two states. There are therefore approximately $2(\pi k^3/6)/(\pi/d)^3 = (k^3/3\pi^2)d^3$ such points in the volume d^3 and $(k^3/3\pi^2)$ points per unit volume. It follows that the number of states with electron wavenumbers between k and $k + \Delta k$, per unit volume, is $\varrho(k)\Delta k = [(d/dk)(k^3/3\pi^2)]\Delta k = (k^2/\pi^2)\Delta k$, so that the density of states is

$$\boxed{\varrho(k) = \frac{k^2}{\pi^2}\,.} \qquad \begin{matrix}(17.1\text{-}6)\\ \text{Density of States}\end{matrix}$$

This derivation is identical to that used for counting the number of modes that can be supported in a three-dimensional electromagnetic resonator (Sec. 11.3C). In the case of electromagnetic modes there are two degrees of freedom associated with the field polarization (i.e., two photon spin values), whereas in the semiconductor case there are two spin values associated with the electron state. In resonator optics the allowed electromagnetic solutions for $\mathbf{k}$ were converted into allowed frequencies via the linear frequency–wavenumber relation $\nu = ck/2\pi$. In semiconductor physics, on the other hand, the allowed solutions for $\mathbf{k}$ are converted into allowed energies via the quadratic energy–wavenumber relations given in (17.1-3) and (17.1-4).

If $\varrho_c(E)\,\Delta E$ represents the number of conduction-band energy levels (per unit volume) lying between E and $E + \Delta E$, then, because of the one-to-one correspondence between E and k governed by (17.1-3), the densities $\varrho_c(E)$ and $\varrho(k)$ must be related by $\varrho_c(E)\,dE = \varrho(k)\,dk$. Thus, the density of allowed energies in the conduction band is $\varrho_c(E) = \varrho(k)/(dE/dk)$. Similarly, the density of allowed energies in the valence band is $\varrho_v(E) = \varrho(k)/(dE/dk)$, where E is given by (17.1-4). The approximate quadratic E–k relations (17.1-3) and (17.1-4), which are valid near the edges of the conduction band and valence band, respectively, are used to evaluate the derivative dE/dk for each band. The result that obtains is

$$\varrho_c(E) = \frac{(2m_c)^{3/2}}{2\pi^2\hbar^3}\sqrt{E - E_c}\,, \quad E \geq E_c \qquad (17.1\text{-}7)$$

$$\varrho_v(E) = \frac{(2m_v)^{3/2}}{2\pi^2\hbar^3}\sqrt{E_v - E}\,, \quad E \leq E_v. \qquad \begin{matrix}(17.1\text{-}8)\\ \text{Density of States}\\ \text{Near Band Edges}\end{matrix}$$

The square-root relation is a result of the quadratic energy–wavenumber formulas for electrons and holes near the band edges. The dependence of the density of states on energy is illustrated in Fig. 17.1-11(c). It is zero at the band edge, and increases away from it at a rate that depends on the effective masses of the electrons and holes. The values of m_c and m_v provided in Table 17.1-1 are averaged values suitable for calculating the density of states.

Probability of Occupancy

In the absence of thermal excitation (at $T = 0°$ K), all electrons occupy the lowest possible energy levels, subject to the Pauli exclusion principle. The valence band is then completely filled (there are no holes) and the conduction band is completely empty (it contains no electrons). When the temperature is raised, thermal excitations raise some electrons from the valence band to the conduction band, leaving behind empty states in the valence band (holes). The laws of statistical mechanics dictate that under conditions of thermal equilibrium at temperature T, the probability that a given state of energy E is occupied by an electron is determined by the **Fermi function**

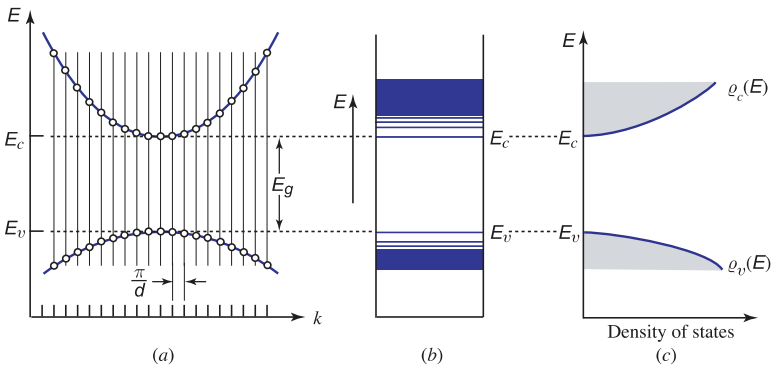


Figure 17.1-11 (a) Cross section of the E–k diagram (e.g., in the direction of the k_1 component, with k_2 and k_3 fixed). (b) Allowed energy levels (at all $\mathbf{k}$). (c) Density of states near the edges of the conduction and valence bands. The quantity $\varrho_c(E)\,dE$ is the number of quantum states with energy between E and $E+dE$, per unit volume, in the conduction band. The quantity $\varrho_v(E)$ has an analogous interpretation for the valence band.

$$f(E) = \frac{1}{\exp\left[(E - E_f)/kT\right] + 1}, \tag{17.1-9}$$

Fermi Function

where k is Boltzmann's constant (at $T = 300°$ K, $kT = 0.026$ eV) and E_f is a constant known as the **Fermi energy** or **Fermi level**. This function, plotted in Fig. 17.1-12, is also known as the **Fermi–Dirac distribution**. Each energy level E is either occupied [with probability $f(E)$], or empty [with probability $1 - f(E)$]. The probabilities $f(E)$ and $1 - f(E)$ depend on the energy E in accordance with (17.1-9). The function $f(E)$ is not itself a probability distribution, and it does not integrate to unity; rather, it is a sequence of occupation probabilities for successive energy levels.

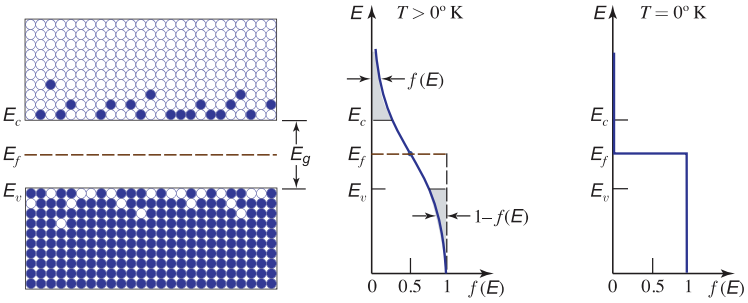


Figure 17.1-12 The Fermi function $f(E)$ is the probability that an energy level E is filled with an electron; $1 - f(E)$ is the probability that it is empty. In the valence band, $1 - f(E)$ is the probability that energy level E is occupied by a hole. At $T = 0°$ K, $f(E) = 1$ for $E \le E_f$, and $f(E) = 0$ for $E > E_f$; there are then no electrons in the conduction band and no holes in the valence band.

Because $f(E_f) = 1/2$, whatever the temperature T, the Fermi level is that energy for which the probability of occupancy (if there were an allowed state there) would be $1/2$. The Fermi function is a monotonically decreasing function of E. At $T = 0°$ K, $f(E)$ is 0 for $E > E_f$ and 1 for $E \le E_f$. This establishes the significance of E_f; it is the division between the occupied and unoccupied energy levels at $T = 0°$ K. Since $f(E)$ is the probability that the energy level E is occupied, $1 - f(E)$ is the probability

that it is empty, i.e., that it is occupied by a hole if E lies in the valence band. Thus, for energy level E:

$$f(E) = \text{probability of occupancy by an electron}$$
$$1 - f(E) = \text{probability of occupancy by a hole (valence band).}$$

When $E - E_f \gg kT$, $f(E) \approx \exp[-(E-E_f)/kT]$, so that the high-energy tail of the Fermi function in the conduction band decreases exponentially with increasing energy. The Fermi function is then proportional to the Boltzmann distribution, which describes the exponential energy dependence of the fraction of a population of atoms excited to a given energy level (Sec. 14.2). By symmetry, when $E < E_f$ and $E_f - E \gg kT$, $1 - f(E) \approx \exp[-(E_f - E)/kT]$; the probability of occupancy by holes in the valence band then decreases exponentially as the energy decreases well below the Fermi level.

Thermal-Equilibrium Carrier Concentrations

Let $\mathrm{n}(E)\,\Delta E$ and $\mathrm{p}(E)\,\Delta E$ be the number of electrons and holes per unit volume, respectively, with energy lying between E and $E + \Delta E$. The densities $\mathrm{n}(E)$ and $\mathrm{p}(E)$ can be obtained by multiplying the densities of states at energy level E by the probabilities of occupancy of the level by electrons or holes, so that

$$\mathrm{n}(E) = \varrho_c(E) f(E), \qquad \mathrm{p}(E) = \varrho_v(E)[1 - f(E)]. \tag{17.1-10}$$

The concentrations (populations per unit volume) of electrons and holes, n and p, are then obtained from the integrals

$$\mathrm{n} = \int_{E_c}^{\infty} \mathrm{n}(E)\,dE, \qquad \mathrm{p} = \int_{-\infty}^{E_v} \mathrm{p}(E)\,dE. \tag{17.1-11}$$

In an intrinsic (pure) semiconductor at any temperature, $\mathrm{n} = \mathrm{p}$ because thermal excitations always create electrons and holes in pairs. The Fermi level must therefore be placed at an energy value such that $\mathrm{n} = \mathrm{p}$. In materials for which $m_v = m_c$, the functions $\mathrm{n}(E)$ and $\mathrm{p}(E)$ are also symmetric, so that E_f must lie precisely in the middle of the bandgap (Fig. 17.1-13). In most intrinsic semiconductors, the Fermi level does indeed lie near the middle of the bandgap.

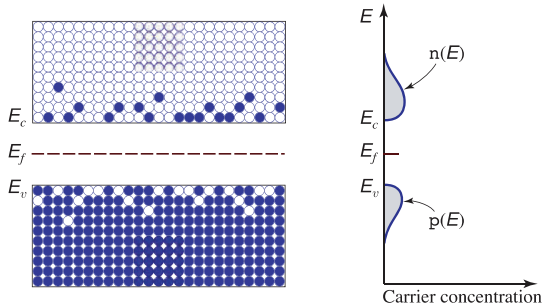


Figure 17.1-13 The concentrations of electrons and holes, $\mathrm{n}(E)$ and $\mathrm{p}(E)$, as a function of energy E, for an intrinsic semiconductor. The total concentrations of electrons and holes are n and p, respectively.

The energy-band diagrams, Fermi functions, and equilibrium concentrations of electrons and holes for n-type and p-type doped semiconductors are illustrated in Figs. 17.1-14 and 17.1-15, respectively. Donor electrons occupy an energy E_D slightly below the conduction-band edge so that they are easily raised to it. If $E_D = 0.01$ eV, for example, at room temperature ($kT = 0.026$ eV) most donor electrons will be thermally excited into the conduction band (Example 17.1-1). As a result, the Fermi

level [the energy at which $f(E_f) = 1/2$] will lie above the middle of the bandgap. For a p-type semiconductor, the acceptor energy level lies at an energy E_A just above the valence-band edge so that the Fermi level will lie below the middle of the bandgap. Our attention has been directed to the mobile carriers in doped semiconductors. These materials are, of course, electrically neutral, as assured by the fixed donor and acceptor ions, so that $\mathrm{n} + N_A = \mathrm{p} + N_D$, where N_A and N_D are, respectively, the number of ionized acceptors and donors per unit volume.

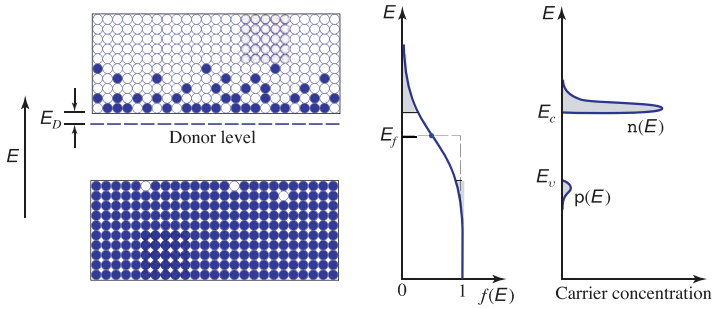


Figure 17.1-14 Energy-band diagram, Fermi function $f(E)$, and concentrations of mobile electrons and holes, $\mathrm{n}(E)$ and $\mathrm{p}(E)$, respectively, in an n-type semiconductor.

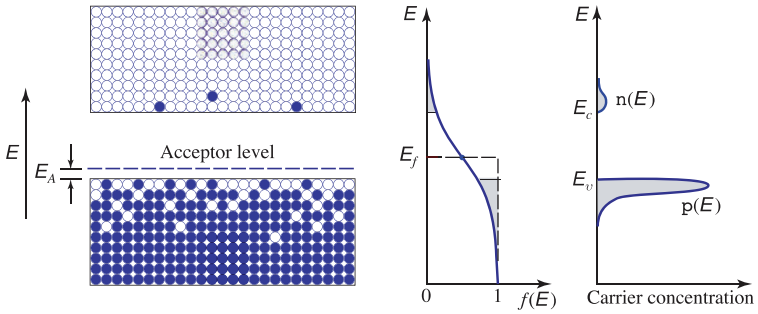


Figure 17.1-15 Energy-band diagram, Fermi function $f(E)$, and concentrations of mobile electrons and holes, $\mathrm{n}(E)$ and $\mathrm{p}(E)$, respectively, in a p-type semiconductor.

EXERCISE 17.1-2

Exponential Approximation of the Fermi Function. When $E - E_f \gg kT$, the Fermi function $f(E)$ may be approximated by an exponential function. Similarly, when $E_f - E \gg kT$, $1 - f(E)$ may be approximated by an exponential function. These conditions apply when the Fermi level lies within the bandgap, but away from its edges by an energy of at least several times kT (at room temperature $kT \approx 0.026$ eV whereas $E_g = 1.12$ eV in Si and 1.42 eV in GaAs). Using these approximations, which apply for both intrinsic and doped semiconductors, show that (17.1-11) gives

$$\mathrm{n} = N_c \exp\left(-\frac{E_c - E_f}{kT}\right) \tag{17.1-12}$$

$$\mathrm{p} = N_v \exp\left(-\frac{E_f - E_v}{kT}\right) \tag{17.1-13}$$

$$\mathrm{np} = N_c N_v \exp\left(-\frac{E_g}{kT}\right), \tag{17.1-14}$$

where $N_c = 2(2\pi m_c kT/h^2)^{3/2}$ and $N_v = 2(2\pi m_v kT/h^2)^{3/2}$. Verify that if E_f is closer to the conduction band and $m_v = m_c$, then $\mathrm{n} > \mathrm{p}$, whereas if it is closer to the valence band, then $\mathrm{p} > \mathrm{n}$.

Law of Mass Action

Equation (17.1-14) reveals that, in thermal equilibrium, the product

$$\mathrm{n}\mathrm{p} = 4\left(\frac{2\pi kT}{h^2}\right)^3 (m_c m_v)^{3/2} \exp\left(-\frac{E_g}{kT}\right) \tag{17.1-15}$$

is independent of the location of the Fermi level E_f within the bandgap and the semiconductor doping level, provided that the exponential approximation to the Fermi function is valid. The constancy of the concentration product is called the **law of mass action**. For an intrinsic semiconductor, $\mathrm{n} = \mathrm{p} \equiv \mathrm{n_i}$. Combining this latter relation with (17.1-14) then leads to

$$\boxed{\mathrm{n_i} \approx \sqrt{N_c N_v}\, \exp\left(-\frac{E_g}{2kT}\right),} \tag{17.1-16}$$

Intrinsic Carrier Concentration

revealing that the intrinsic concentration of electrons and holes increases with temperature T at an exponential rate. The law of mass action may therefore be written in the form

$$\boxed{\mathrm{n}\mathrm{p} = \mathrm{n_i^2}.} \tag{17.1-17}$$

Law of Mass Action

The values of $\mathrm{n_i}$ for different materials vary because of differences in the bandgap energies and effective masses. The room-temperature intrinsic carrier concentrations for Si, GaAs, and GaN are provided in Table 17.1-3.

Table 17.1-3 Intrinsic carrier concentrations at $T = 300°$ K.[a]

Material	$\mathrm{n_i}$ (cm^{-3})
Si	1.5×10^{10}
GaAs	1.8×10^{6}
GaN	1.9×10^{-10}

[a] Substitution of the values of m_c and m_v provided in Table 17.1-1, and the value for E_g given in Table 17.1-2, into (17.1-16), does not yield the listed values of $\mathrm{n_i}$ because of the sensitivity of the formula to the precise values of the parameters.

The law of mass action is useful for determining the concentrations of electrons and holes in doped semiconductors. A moderately doped n-type material, for example, has a concentration of electrons n that is essentially equal to the donor concentration N_D. Using the law of mass action, the hole concentration is then $\mathrm{p} = \mathrm{n_i^2}/N_D$. Knowledge of n and p allows the Fermi level to be determined via (17.1-11). As long as the Fermi level lies within the bandgap, at an energy greater than several times kT from its edges, the approximate relations in (17.1-12) and (17.1-13) can be used to determine it directly.

If the Fermi level lies inside the conduction (or valence) band, the material is referred to as a **degenerate semiconductor**. In that case, the exponential approximation of the Fermi function cannot be used, so that $\mathrm{n}\mathrm{p} \neq \mathrm{n_i^2}$. The carrier concentrations must then be obtained by numerical solution. Under conditions of very heavy doping, the donor (acceptor) impurity band actually merges with the conduction (valence) band to become what is known as the **band tail**. This results in an effective decrease of the bandgap.

Quasi-Equilibrium Carrier Concentrations

The occupancy probabilities and carrier concentrations considered above are applicable only for a semiconductor in thermal equilibrium. They are not valid when thermal equilibrium is disturbed. There are, nevertheless, situations in which the conduction-band electrons are in thermal equilibrium among themselves, as are the valence-band holes, but the electrons and holes are not in mutual thermal equilibrium. This can occur, for example, when an external electric current or photon flux induces band-to-band transitions at too high a rate for interband equilibrium to be achieved. This situation, which is known as **quasi-equilibrium**, arises when the relaxation (decay) times for transitions within each of the bands are much shorter than the relaxation time between the two bands. Typically, the intraband relaxation time $< 10^{-12}$ s, whereas the radiative electron–hole recombination time $\approx 10^{-9}$ s.

Under these circumstances, it is appropriate to use a separate Fermi function for each band; the two associated Fermi levels, denoted E_{fc} and E_{fv}, are known as **quasi-Fermi levels** (Fig. 17.1-16). When E_{fc} and E_{fv} lie well inside the conduction and valence bands, respectively, the concentration of *both* electrons and holes can be quite large.

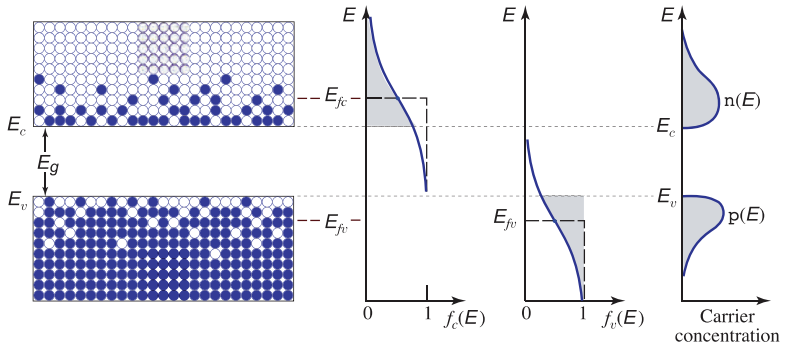


Figure 17.1-16 A semiconductor in quasi-equilibrium. The probability that a particular conduction-band energy level E is occupied by an electron is $f_c(E)$, a Fermi function with Fermi level E_{fc}. The probability that a valence-band energy level E is occupied by a hole is $1 - f_v(E)$, where $f_v(E)$ is a Fermi function with Fermi level E_{fv}. The concentrations of electrons and holes are $\mathrm{n}(E)$ and $\mathrm{p}(E)$, respectively. Both can be large.

EXERCISE 17.1-3

Determination of the Quasi-Fermi Levels Given the Electron and Hole Concentrations.

(a) Given the concentrations of electrons n and holes p in a semiconductor at $T = 0°$ K, use (17.1-10) and (17.1-11) to show that the quasi-Fermi levels are

$$E_{fc} = E_c + (3\pi^2)^{2/3} \frac{\hbar^2}{2m_c} \mathrm{n}^{2/3} \tag{17.1-18a}$$

$$E_{fv} = E_v - (3\pi^2)^{2/3} \frac{\hbar^2}{2m_v} \mathrm{p}^{2/3}. \tag{17.1-18b}$$

(b) Show that these equations are approximately applicable for an arbitrary temperature T if n and p are sufficiently large so that $E_{fc} - E_c \gg kT$ and $E_v - E_{fv} \gg kT$, i.e., if the quasi-Fermi levels lie deep within the conduction and valence bands.

D. Generation, Recombination, and Injection

Generation and Recombination in Thermal Equilibrium

The thermal excitation of electrons from the valence band into the conduction band results in the **electron–hole generation** (Fig. 17.1-17). Thermal equilibrium requires that this generation process be accompanied by a simultaneous reverse process of de-excitation. This process, called **electron–hole recombination**, occurs when an electron decays from the conduction band to fill a hole in the valence band (Fig. 17.1-17). The energy released by the electron may take the form of an emitted photon, in which case the process is called **radiative recombination**.

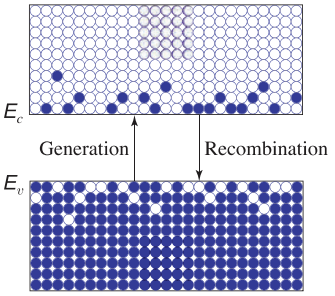


Figure 17.1-17 Electron–hole generation and recombination.

Nonradiative recombination can occur via a number of independent competing processes, including the transfer of energy to lattice vibrations (creating one or more phonons) or to another free electron via Auger recombination (which, as a three-particle interaction, can take place when the carrier density is very high). Recombination may also take place at surfaces and indirectly via traps or defect centers, which are energy levels associated with impurities or defects associated with grain boundaries, dislocations, or other lattice imperfections that lie within the forbidden band. An impurity or defect state can act as a recombination center if it is capable of trapping both an electron and a hole, thereby increasing their probability of recombining (Fig. 17.1-18). Impurity-assisted recombination may be radiative or nonradiative.

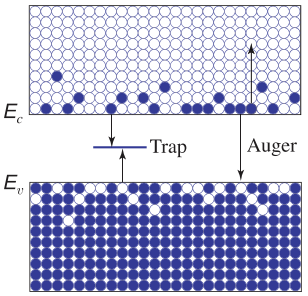


Figure 17.1-18 Electron–hole recombination via a trap; via Auger recombination.

Because it takes both an electron *and* a hole for a recombination to occur, the rate of recombination is proportional to the product of the concentration of electrons and holes, i.e.,

$$\text{rate of recombination} = \mathtt{r}np, \tag{17.1-19}$$

where the **recombination coefficient** $\mathtt{r}$ (cm^3/s) depends on the characteristics of the material, including its composition and defect density, and on temperature; it also depends relatively weakly on the doping level.

The equilibrium concentrations of electrons and holes n_0 and p_0 are established when the generation and recombination rates are in balance. In the steady state, the rate of recombination must equal the rate of generation. If G_0 is the rate of thermal electron–hole generation at a given temperature, then, in thermal equilibrium,

$$G_0 = r n_0 p_0. \tag{17.1-20}$$

The product of the electron and hole concentrations $n_0 p_0 = G_0/r$ is approximately the same whether the material is n-type, p-type, or intrinsic. Thus, $n_i^2 = G_0/r$, which leads directly to the law of mass action $n_0 p_0 = n_i^2$. This law is therefore seen to be a consequence of the balance between generation and recombination in thermal equilibrium.

Electron–Hole Injection

A semiconductor in thermal equilibrium with carrier concentrations n_0 and p_0 has equal rates of generation and recombination, $G_0 = r n_0 p_0$. Now let additional electron–hole pairs be generated at a steady rate R (pairs per unit volume per unit time) by means of an external (nonthermal) injection mechanism, such as light falling on the material. A new steady state will be reached in which the concentrations are $n = n_0 + \Delta n$ and $p = p_0 + \Delta p$. It is clear, however, that $\Delta n = \Delta p$ since the electrons and holes are created in pairs. Equating the new rates of generation and recombination, we obtain

$$G_0 + R = r n p. \tag{17.1-21}$$

Substituting $G_0 = r n_0 p_0$ into (17.1-21) leads to

$$R = r(np - n_0 p_0) = r\left(n_0 \Delta n + p_0 \Delta n + \Delta n^2\right) = r \Delta n (n_0 + p_0 + \Delta n), \tag{17.1-22}$$

which we write in the form

$$R = \frac{\Delta n}{\tau}, \tag{17.1-23}$$

with

$$\tau = \frac{1}{r\left[(n_0 + p_0) + \Delta n\right]}. \tag{17.1-24}$$

For an injection rate such that $\Delta n \ll n_0 + p_0$,

$$\boxed{\tau \approx \frac{1}{r(n_0 + p_0)}.} \tag{17.1-25}$$

Excess-Carrier Recombination Lifetime

In an n-type material, where $n_0 \gg p_0$, the recombination lifetime $\tau \approx 1/r n_0$ is inversely proportional to the electron concentration. Similarly, for a p-type material where $p_0 \gg n_0$, we obtain $\tau \approx 1/r p_0$. This simple formulation is not applicable when traps play an important role in the process.

The parameter τ may be regarded as the **electron–hole recombination lifetime** of the injected excess electron–hole pairs. This is readily understood by noting that the injected-carrier concentration is governed by the rate equation

$$\frac{d(\Delta n)}{dt} = R - \frac{\Delta n}{\tau}, \tag{17.1-26}$$

which is similar to (15.2-3). In the steady state, $d(\Delta \mathrm{n})/dt = 0$ whereupon (17.1-23), which is like (15.2-13), is recovered. If the source of injection is suddenly removed (R becomes 0) at the time t_0, then $\Delta \mathrm{n}$ decays exponentially with time constant τ, i.e., $\Delta \mathrm{n}(t) = \Delta \mathrm{n}(t_0) \exp[-(t - t_0)/\tau]$. In the presence of strong injection, on the other hand, τ is itself a function of $\Delta \mathrm{n}$, as evident from (17.1-24), so that the rate equation is nonlinear and the decay is no longer exponential.

If the injection rate R is known, the steady-state injected concentration may be determined from

$$\boxed{\Delta \mathrm{n} = R\tau,} \tag{17.1-27}$$

permitting the total concentrations $\mathrm{n} = \mathrm{n}_0 + \Delta \mathrm{n}$ and $\mathrm{p} = \mathrm{p}_0 + \Delta \mathrm{n}$ to be determined. Furthermore, if quasi-equilibrium is assumed, (17.1-11) may be used to determine the quasi-Fermi levels. Quasi-equilibrium is not inconsistent with the balance of generation and recombination assumed in the analysis above; it simply requires that the intraband equilibrium time be short in comparison with the recombination time τ.

This type of analysis will prove useful in developing theories of the semiconductor light-emitting diode and the semiconductor laser diode, which are based on enhancing light emission by means of carrier injection, as will become clear in Chapter 18.

EXERCISE 17.1-4

Electron–Hole Pair Injection in GaAs. Assume that electron–hole pairs are injected into n-type GaAs ($E_g = 1.42$ eV, $m_c \approx 0.07\, m_0$, $m_v \approx 0.50\, m_0$) at a rate $R = 10^{23}/\mathrm{cm}^3\text{-s}$. The thermal equilibrium concentration of electrons is $\mathrm{n}_0 = 10^{16}/\mathrm{cm}^3$. If the recombination coefficient $\mathrm{r} = 10^{-11}\ \mathrm{cm}^3/\mathrm{s}$ and $T = 300^\circ$ K, determine:

(a) The equilibrium concentration of holes p_0.
(b) The steady-state excess concentration $\Delta \mathrm{n}$.
(c) The recombination lifetime τ.
(d) The separation between the quasi-Fermi levels $E_{fc} - E_{fv}$, assuming that $T = 0^\circ$ K.

Internal Quantum Efficiency

The **internal quantum efficiency** η_i of a semiconductor material is defined as the ratio of the radiative electron–hole recombination coefficient to the total (radiative and nonradiative) recombination coefficient. This parameter is important because it determines the efficiency of light generation in a semiconductor material. The total rate of recombination is given by (17.1-19). If the recombination coefficient r is split into a sum of radiative and nonradiative parts, $\mathrm{r} = \mathrm{r}_\mathrm{r} + \mathrm{r}_\mathrm{nr}$, the internal quantum efficiency is

$$\eta_i = \frac{\mathrm{r}_\mathrm{r}}{\mathrm{r}} = \frac{\mathrm{r}_\mathrm{r}}{\mathrm{r}_\mathrm{r} + \mathrm{r}_\mathrm{nr}}. \tag{17.1-28}$$

The internal quantum efficiency may also be written in terms of the recombination lifetimes since τ is inversely proportional to r [see (17.1-25)]. Defining the radiative and nonradiative lifetimes τ_r and τ_nr, respectively, leads to

$$\frac{1}{\tau} = \frac{1}{\tau_\mathrm{r}} + \frac{1}{\tau_\mathrm{nr}}. \tag{17.1-29}$$

The internal quantum efficiency is then $\mathrm{r_r}/\mathrm{r} = (1/\tau_\mathrm{r})/(1/\tau)$, or

$$\eta_i = \frac{\tau}{\tau_\mathrm{r}} = \frac{\tau_\mathrm{nr}}{\tau_\mathrm{r} + \tau_\mathrm{nr}} . \tag{17.1-30}$$

Internal Quantum Efficiency

The radiative recombination lifetime τ_r governs the rate of photon absorption and emission, as explained in Sec. 17.2C. Its value depends on the carrier concentrations and the material parameter $\mathrm{r_r}$. For low to moderate injection rates,

$$\tau_\mathrm{r} \approx \frac{1}{\mathrm{r_r}(\mathrm{n}_0 + \mathrm{p}_0)} , \tag{17.1-31}$$

in accordance with (17.1-25). The nonradiative recombination lifetime is governed by a similar equation. However, if nonradiative recombination takes place via defect centers in the forbidden band, τ_nr is more sensitive to the concentration of these centers than to the electron and hole concentrations.

Typical values for recombination coefficients and lifetimes are listed in Table 17.1-4. Order-of-magnitude values are given for the radiative recombination coefficients $\mathrm{r_r}$; the radiative, nonradiative, and overall recombination lifetimes, τ_r, τ_nr, and τ, respectively; and the internal quantum efficiencies η_i.

Table 17.1-4 Representative values for radiative recombination coefficients $\mathrm{r_r}$, recombination lifetimes, and internal quantum efficiencies η_i, for representative semiconductors.[a]

Material	$\mathrm{r_r}$ (cm^3/s)	τ_r	τ_nr	τ	η_i
Si	10^{-15}	10 ms	100 ns	100 ns	10^{-5}
GaAs	10^{-10}	100 ns	100 ns	50 ns	0.5
GaN[b]	10^{-8}	20 ns	0.1 ns	0.1 ns	0.005

[a] Assuming n-type material with a carrier concentration $\mathrm{n}_0 = 10^{17}/\mathrm{cm}^3$ and defect centers with a concentration $10^{15}/\mathrm{cm}^3$, at $T = 300^\circ$ K.
[b] As a matter of practice, InGaN is used; this increases the internal quantum efficiency to $\eta_i \approx 0.3$.

The radiative lifetime for bulk Si is orders of magnitude longer than its overall lifetime, principally because of its indirect bandgap. This results in a small internal quantum efficiency. For GaAs and GaN, on the other hand, the decay is largely via radiative transitions (these materials have a direct bandgap), and consequently the internal quantum efficiency is large. Direct-bandgap materials are therefore useful for fabricating light-emitting structures that operate via interband spontaneous and stimulated emission, whereas indirect-bandgap materials generally are not.

Light emission from indirect-bandgap materials can, nevertheless, be achieved by making use of interactions such as stimulated Raman scattering and intersubband transitions (Sec. 18.1D), which do not rely on the interband transitions discussed in Sec. 17.2B.

E. Junctions

Juxtapositions of differently doped regions of a single semiconductor material are called **homojunctions**. An important example is the p–n junction, which is discussed in this section. Junctions between different semiconductor materials are called **heterojunctions**. These are discussed subsequently.

The p–n Junction

The p–n junction is a homojunction between a p-type and an n-type semiconductor. It acts as a diode, which can serve in electronics as a rectifier, logic gate, voltage regulator (Zener diode), or tuner (varactor diode); and in photonics as a light-emitting diode (LED), laser diode (LD), photodetector, or solar cell.

A p–n junction consists of a p-type and an n-type section of the same semiconductor materials in metallurgical contact. The p-type region has an abundance of holes (majority carriers) and few mobile electrons (minority carriers); the n-type region has an abundance of mobile electrons and few holes (Fig. 17.1-19). Both charge carriers are in continuous random thermal motion in all directions.

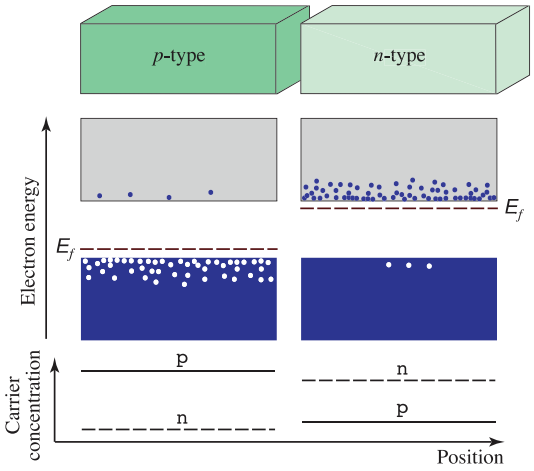


Figure 17.1-19 Energy levels and carrier concentrations for a p-type and an n-type semiconductor before contact.

When the two regions are brought into contact (Fig. 17.1-20), the following sequence of events takes place:

- Electrons and holes diffuse from areas of high concentration toward areas of low concentration. Thus, electrons diffuse from the n-region into the p-region, leaving behind positively charged ionized donor atoms. In the p-region the electrons recombine with the abundant holes. Similarly, holes diffuse from the p-region into the n-region, leaving behind negatively charged ionized acceptor atoms. In the n-region the holes recombine with the abundant mobile electrons. This diffusion process does not continue indefinitely, however, because it causes a disruption of the charge balance in the two regions.
- As a result, a narrow region on both sides of the junction becomes nearly depleted of *mobile* charge carriers. This region is called the **depletion layer**. It contains only the *fixed* charges (positive ions on the n-side and negative ions on the p-side). The thickness of the depletion layer in each region is inversely proportional to the concentration of dopants in the region.
- The fixed charges create an electric field in the depletion layer that points from the n-side toward the p-side of the junction. This **built-in field** obstructs the diffusion of further mobile carriers through the junction region.
- An equilibrium condition is established that results in a net built-in potential difference V_0 between the two sides of the depletion layer, with the n-side exhibiting a higher potential than the p-side.
- The built-in potential provides a lower potential energy for an electron on the n-side relative to the p-side. As a result, the energy bands bend, as illustrated in

Fig. 17.1-20. In thermal equilibrium there is only a single Fermi function for the entire structure so that the Fermi levels in the p- and n-regions must align.

- No *net* current flows across the junction. The currents associated with diffusion and built-in field (drift current) cancel for both the electrons and holes.

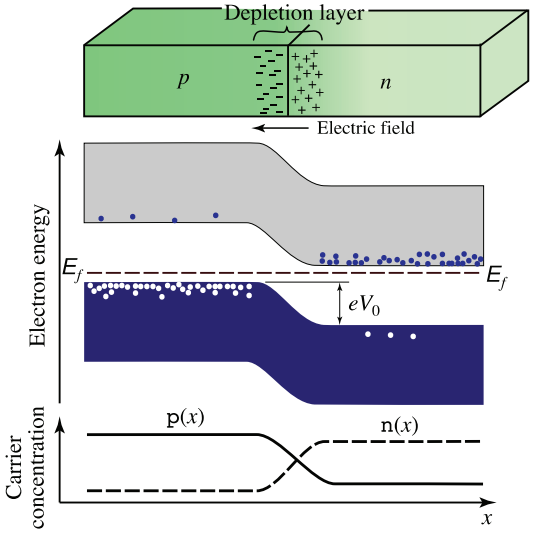


Figure 17.1-20 A p–n junction in thermal equilibrium at $T > 0°$ K. The depletion-layer, energy-band diagram, and concentrations (on a logarithmic scale) of mobile electrons $\mathrm{n}(x)$ and holes $\mathrm{p}(x)$ are shown as functions of the position x. The built-in potential difference V_0 corresponds to an energy eV_0, where e is the magnitude of the electron charge.

The Biased p–n Junction

An externally applied potential will alter the potential difference between the p- and n-regions. This in turn will modify the flow of majority carriers, so that the junction can be used as a "gate." If the junction is **forward biased** by applying a positive voltage V to the p-region (Fig. 17.1-21), its potential is increased with respect to the n-region, so that an electric field is produced in a direction opposite to that of the built-in field. The presence of the external bias voltage causes a departure from equilibrium and a misalignment of the Fermi levels in the p- and n-regions, as well as in the depletion layer. The presence of two Fermi levels in the depletion layer, E_{fc} and E_{fv}, represents a state of quasi-equilibrium.

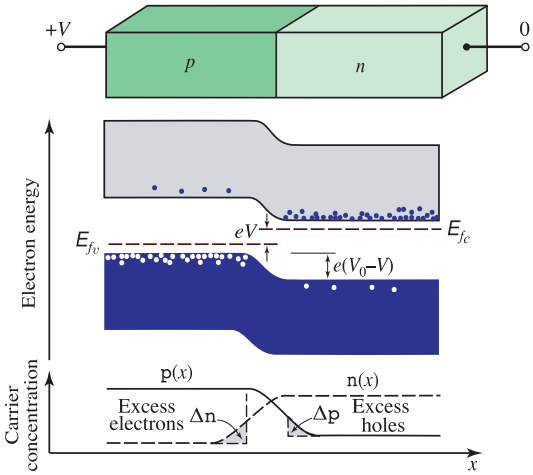


Figure 17.1-21 Energy-band diagram and carrier concentrations for a forward-biased p–n junction.

The net effect of the forward bias is to reduce the height of the potential-energy hill by an amount eV. The majority carrier current turns out to increase by an exponential factor $\exp(eV/kT)$ so that the net current becomes $i = i_s \exp(eV/kT) - i_s$, where i_s is a constant. The excess majority carrier holes and electrons that enter the n- and p-regions, respectively, become minority carriers and recombine with the local majority carriers. Their concentration therefore decreases with distance from the junction as shown in Fig. 17.1-21. This process is known as **minority carrier injection**.

If the junction is **reverse biased** by applying a negative voltage V to the p-region, the height of the potential-energy hill is augmented by eV. This impedes the flow of majority carriers. The corresponding current is multiplied by the exponential factor $\exp(eV/kT)$, where V is negative; i.e., it is reduced. The net result for the current is $i = i_s \exp(eV/kT) - i_s$, so that a small current of magnitude $\approx i_s$ flows in the reverse direction when $|V| \gg kT/e$.

A p–n junction therefore acts as a diode with a current–voltage (i–V) characteristic

$$i = i_s \left[\exp\left(\frac{eV}{kT}\right) - 1 \right], \tag{17.1-32}$$

Ideal Diode Characteristic

as illustrated in Fig. 17.1-22. The ideal diode characteristic in (17.1-32) is known as the **Shockley equation**.

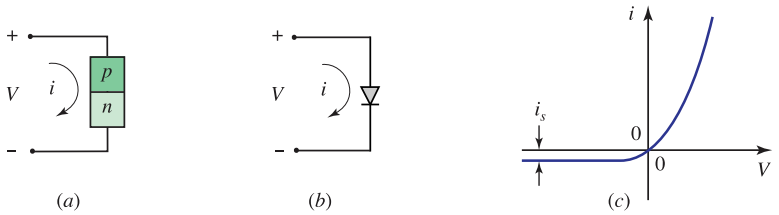


Figure 17.1-22 (a) Voltage and current in a p–n junction. (b) Circuit representation of the p–n junction diode. (c) Current–voltage characteristic of the ideal p–n junction diode.

The response of a p–n junction to a dynamic (AC) applied voltage is determined by solving the set of differential equations governing the processes of electron and hole diffusion, drift (under the influence of the built-in and external electric fields), and recombination. These effects are important for determining the speed at which the diode can be operated. They may be conveniently modeled by two capacitances, a junction capacitance and diffusion capacitance, in parallel with an ideal diode. The **junction capacitance** accounts for the time necessary to change the fixed positive and negative charges stored in the depletion layer when the applied voltage changes. The thickness l of the depletion layer turns out to be proportional to $\sqrt{V_0 - V}$; it therefore increases under reverse-bias conditions (negative V) and decreases under forward-bias conditions (positive V). The junction capacitance $C = \epsilon A/l$ (where A is the area of the junction) is therefore inversely proportional to $\sqrt{V_0 - V}$. The junction capacitance of a reverse-biased diode is smaller (and the RC response time is therefore shorter) than that of a forward-biased diode. The dependence of C on V is used to make voltage-variable capacitors (varactors).

Minority carrier injection in a forward-biased diode is described by the **diffusion capacitance**, which depends on the minority carrier lifetime and the operating current.

The p–i–n Junction Diode

A p–i–n (**PIN**) junction diode is made by inserting a layer of intrinsic (or lightly doped) semiconductor material between a p-type region and an n-type region (Fig. 17.1-23). Because the depletion layer extends into each side of a junction by a distance inversely proportional to the doping concentration, the depletion layer of the p–i junction penetrates deeply into the i-region. Similarly, the depletion layer of the i–n junction extends well into the i-region. As a result, the p–i–n diode can behave like a p–n junction with a depletion layer that encompasses the entire intrinsic region. The electron energy, density of fixed charges, and the electric field in a p–i–n junction diode in thermal equilibrium are illustrated in Fig. 17.1-23. One advantage of using a diode with a large depletion layer is its small junction capacitance and its consequent fast response. For this reason, p–i–n diodes are often favored over p–n diodes for use as semiconductor photodetectors. The large depletion layer also permits an increased fraction of the incident light to be captured, thereby increasing the photodetection efficiency (Sec. 19.3B).

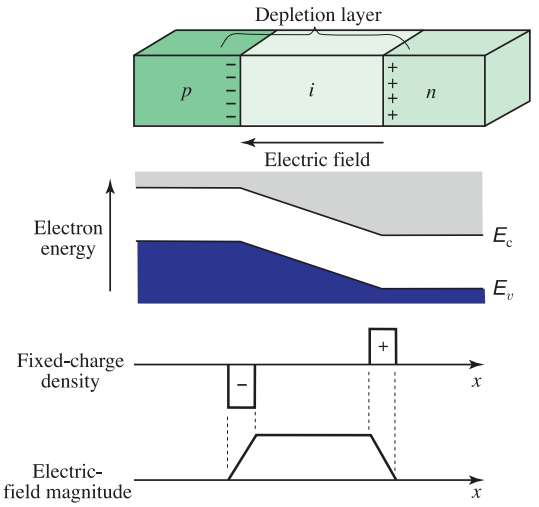


Figure 17.1-23 Electron energy, fixed-charge density, and electric field magnitude for a p–i–n junction diode in thermal equilibrium.

F. Heterojunctions

Junctions between different semiconductor materials are known as heterojunctions. Optical sources and detectors make extensive use of heterojunctions in their designs; they are used not only as active regions but also as contact layers and waveguiding regions. The electron affinities of the materials determine the alignments of the conduction- and valence-band edges. It is often advantageous to lattice match the semiconductor materials and to make use of graded junctions rather than abrupt ones. The juxtaposition of different semiconductors can have manifold advantages in photonics:

- Junctions between materials of different bandgap create localized jumps in the energy-band diagram, as portrayed in Fig. 17.1-24. A potential-energy discontinuity provides a barrier that can be useful in preventing selected charge carriers from entering regions where they are undesired. This property may be used in a p–n junction, for example, to reduce the proportion of current carried by minority carriers, and thus to increase injection efficiency.

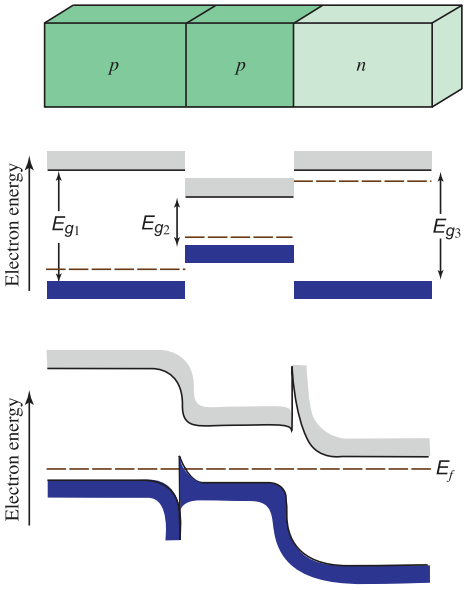


Figure 17.1-24 The p–p–n double heterojunction structure. The middle layer is of narrower bandgap than the outer layers. In equilibrium, the Fermi levels align so that the edge of the conduction band drops sharply at the p–p junction and the edge of the valence band drops sharply at the p–n junction. The conduction- and valence-band discontinuities are known as **band offsets**. When the device is forward biased, these jumps act as barriers that confine the injected minority carriers to the region of lower bandgap. Electrons injected from the n-region, for example, are prevented from diffusing beyond the barrier at the p–p junction. Similarly, holes injected from the p-region are not permitted to diffuse beyond the energy barrier at the p–n junction. This double-heterostructure configuration therefore forces electrons and holes to occupy a narrow common region. This substantially increases the efficiency of light-emitting diodes, semiconductor optical amplifiers, and laser diodes, as discussed in Chapter 18.

- Discontinuities in the energy-band diagram created by two heterojunctions can be useful for confining charge carriers to a desired region of space. For example, a layer of narrow-bandgap material can be sandwiched between two layers of a wider bandgap material, as shown in the p–p–n structure illustrated in Fig. 17.1-24 (which consists of a p–p heterojunction and a p–n heterojunction). This **double-heterostructure** (DH) configuration is used effectively in the fabrication of LEDs, semiconductor optical amplifiers, and laser diodes, as explained in Chapter 18.
- Heterojunctions are useful for creating energy-band discontinuities that accelerate carriers at specific locations. The additional kinetic energy suddenly imparted to a carrier can be useful for selectively enhancing the probability of impact ionization in a multilayer avalanche photodiode (Sec. 19.4B).
- Semiconductors of different bandgap type (direct and indirect) can be used in the same device to select regions of the structure where light is emitted. Semiconductors of the direct-bandgap type emit light efficiently (Sec. 17.2B).
- Semiconductors of different bandgaps can be used in the same device to select regions of the structure where light is absorbed. A semiconductor material whose bandgap energy is larger than the photon energy of light incident on it will be transparent, acting as a **window layer**.
- Heterojunctions of materials with different refractive indices can be used to create photonic structures and optical waveguides that confine and direct photons, as discussed in Chapters 7 and 9.

G. Quantum-Confined Structures

Heterostructures of thin layers of semiconductor materials can be grown epitaxially, i.e., as layers of one semiconductor material over another, by using techniques such as molecular-beam epitaxy (MBE); liquid-phase epitaxy (LPE); and vapor-phase epitaxy (VPE), of which common variants are metalorganic chemical vapor deposition (MOCVD) and hydride vapor-phase epitaxy (HVPE). **Homoepitaxy** is the growth of

materials that have the same composition as the substrate whereas **heteroepitaxy** is the growth of materials on a substrate of different composition, whether lattice-matched or not. MBE makes use of molecular beams of the constituent elements that are caused to impinge on an appropriately prepared substrate in a high-vacuum environment, LPE uses the cooling of a saturated solution containing the constituents in contact with the substrate, and VPE uses gases in a reactor. The compositions and dopings of the individual layers, which can be made as thin as monolayers, are determined by manipulating the arrival rates of the molecules and the temperature of the substrate surface.

When the layer thickness is comparable to, or smaller than, the de Broglie wavelength of a thermalized electron, the quantized energy of an electron resident in the layer must be accommodated, in which case the energy–momentum relation for a bulk semiconductor material is no longer applicable. The de Broglie wavelength is expressed as $\lambda_{dB} = h/p$, where h is Planck's constant and p is the electron momentum ($\lambda_{dB} \approx 50$ nm for GaAs). Three structures offer substantial advantages for use in photonics: quantum wells, quantum wires, and quantum dots (Sec. 14.1D). The appropriate energy–momentum relations for these structures are derived below. The use of quantum-confined structures in photonic devices is considered in Chapters 18 and 19.

Quantum Wells

A **quantum-well** structure, displayed in Fig. 17.1-25, is a double heterostructure consisting of an ultrathin ($\lesssim 50$ nm) layer of semiconductor material whose bandgap is smaller than that of the surrounding material. An example is provided by a thin layer of GaAs surrounded by AlGaAs (Fig. 14.1-12). The sandwich forms 1D conduction- and valence-band rectangular potential wells within which electrons and holes are confined: electrons in the conduction-band well and holes in the valence-band well.

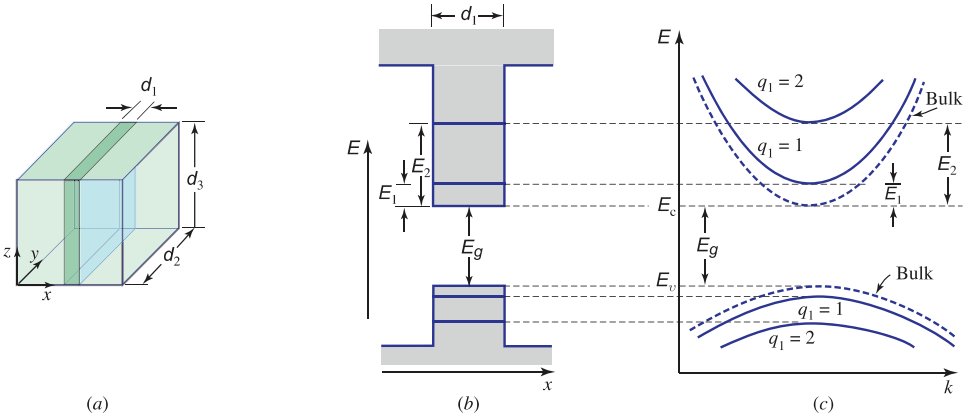


Figure 17.1-25 (a) Geometry of the quantum-well structure. (b) Energy-level diagram for electrons and holes in a quantum well. (c) Cross section of the E–k relation in the direction of k_2 or k_3. The energy subbands are labeled by their quantum number $q_1 = 1, 2, \ldots$. The E–k relation for bulk semiconductor is indicated by the dashed curves.

A sufficiently deep potential well can be approximated as an infinite rectangular potential well (Fig. 17.1-26). The energy levels E_q of a particle of mass m (m_c for electrons and m_v for holes) confined to a one-dimensional infinite rectangular well

of full width d are determined by solving the time-independent Schrödinger equation (14.1-3). As shown in Exercise 17.1-5, the energy levels turn out to be

$$E_q = \frac{\hbar^2(q\pi/d)^2}{2m}, \qquad q = 1, 2, 3, \ldots . \tag{17.1-33}$$

As an example, the first three allowed energy levels of an electron in an infinitely deep GaAs well ($m_c = 0.07\,m_0$) of width $d = 10$ nm are $E_q = 54$, 216, and 486 meV, respectively (recall that $kT = 26$ meV at $T = 300°$ K). The smaller the width of the well, the larger the separation between adjacent energy levels.

EXERCISE 17.1-5

Energy Levels of a Quantum Well. Solve the Schrödinger equation (14.1-3) to determine the allowed energies of an electron of mass m in an infinitely deep one-dimensional rectangular potential well [$V(x) = 0$ for $0 < x < d$ and $V(x) = \infty$ otherwise], confirming that $E_q = \hbar^2(q\pi/d)^2/2m$, $q = 1, 2, 3, \ldots$, as illustrated in Fig. 17.1-26(a). Compare these energies with those for the particular finite square quantum well shown in Fig. 17.1-26(b).

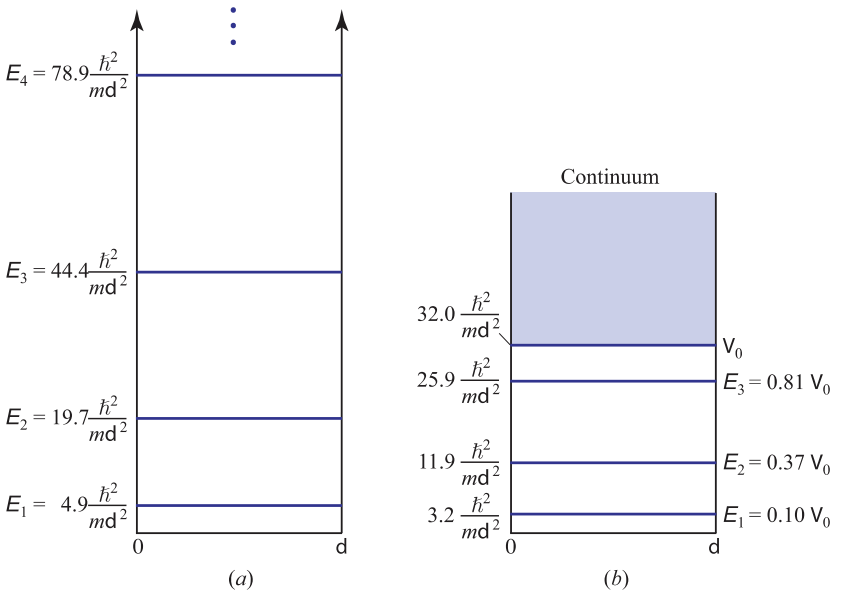


Figure 17.1-26 Energy levels of (a) a one-dimensional infinite rectangular potential well, and (b) a finite square quantum well with an energy depth $V_0 = 32\hbar^2/md^2$.

However, semiconductor quantum wells are actually three-dimensional constructs. In the quantum-well structure shown in Fig. 17.1-25, electrons (and holes) are confined in the x direction to within a distance d_1 (the well thickness), but they extend over much larger dimensions ($d_2, d_3 \gg d_1$) in the plane of the confining layer. Thus, in the y–z plane, they behave as if they were in bulk semiconductor.

The electron energy–momentum relation is

$$E = E_c + \frac{\hbar^2 k_1^2}{2m_c} + \frac{\hbar^2 k_2^2}{2m_c} + \frac{\hbar^2 k_3^2}{2m_c}, \tag{17.1-34}$$

where $k_1 = q_1\pi/d_1$, $k_2 = q_2\pi/d_2$, $k_3 = q_3\pi/d_3$, and $q_1, q_2, q_3 = 1, 2, 3, \ldots$. Since $d_1 \ll d_2, d_3$, the parameter k_1 takes on well-separated discrete values, whereas k_2 and k_3 have finely spaced discrete values that may be approximated as a continuum. It follows that the energy–momentum relation for electrons in the conduction band of a quantum well is given by

$$E = E_c + E_{q1} + \frac{\hbar^2 k^2}{2m_c}, \qquad q_1 = 1, 2, 3, \ldots, \tag{17.1-35}$$

where k is the magnitude of a two-dimensional $\mathbf{k} = (k_2, k_3)$ vector in the y–z plane. Each quantum number q_1 corresponds to a **subband** whose lowest energy is $E_c + E_{q1}$. Similar relations apply for the valence band.

The energy–momentum relation for a bulk semiconductor is given by (17.1-3), where k is the magnitude of a three-dimensional vector $\mathbf{k} = (k_1, k_2, k_3)$. The key distinction is that for the quantum well, k_1 takes on well-separated, discrete values. As a result, the density of states associated with a quantum-well structure differs from that associated with bulk material, for which the density of states is determined from the magnitude of the three-dimensional vector with components $k_1 = q_1\pi/d$, $k_2 = q_2\pi/d$, and $k_3 = q_3\pi/d$ for $d_1 = d_2 = d_3 = d$. The result is $\varrho(k) = k^2/\pi^2$ per unit volume [see (17.1-6)], which yields the density of conduction-band states [see (17.1-7) and Fig. 17.1-11]

$$\varrho_c(E) = \frac{\sqrt{2}\, m_c^{3/2}}{\pi^2\hbar^3}\sqrt{E - E_c}, \qquad E > 0. \tag{17.1-36}$$

In a quantum-well structure the density of states is obtained from the magnitude of the *two*-dimensional vector (k_2, k_3). For each quantum number q_1 the density of states is therefore $\varrho(k) = k/\pi$ states per unit area in the y–z plane, and therefore $k/\pi d_1$ per unit volume. The densities $\varrho_c(E)$ and $\varrho(k)$ are related by $\varrho_c(E)\, dE = \varrho(k)\, dk = (k/\pi d_1)\, dk$. Finally, using the E–k relation (17.1-35) we obtain $dE/dk = \hbar^2 k/m_c$, from which

$$\varrho_c(E) = \begin{cases} \dfrac{m_c}{\pi\hbar^2 d_1}, & E > E_c + E_{q1} \\ 0, & E < E_c + E_{q1}, \end{cases} \qquad q_1 = 1, 2, 3, \ldots. \tag{17.1-37}$$

Thus, for each quantum number q_1, the density of states per unit volume is constant when $E > E_c + E_{q1}$. The overall density of states is the sum of the densities for all values of q_1, so that it exhibits the staircase distribution shown in Fig. 17.1-27. Each step of the staircase corresponds to a different quantum number q_1 and may be regarded as a subband within the conduction band (Fig. 17.1-25). The bottoms of these subbands move progressively higher for higher quantum numbers. It can be shown by substituting $E = E_c + E_{q1}$ in (17.1-36), and by using (17.1-33), that at $E = E_c + E_{q1}$ the quantum-well density of states is the same as that for the bulk material. The density of states in the valence band has a similar staircase distribution.

In contrast with bulk semiconductor, the quantum-well structure exhibits a substantial density of states at its lowest allowed conduction-band energy level and at its highest allowed valence-band energy level. This property has an important effect on the optical characteristics of the material, as discussed in Sec. 18.2D.

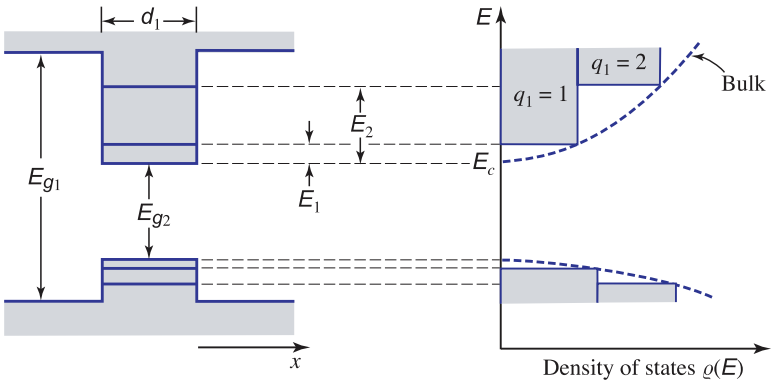


Figure 17.1-27 Density of states for a quantum-well structure (solid curve) and for a bulk semiconductor (dashed curve).

Multiquantum Wells and Superlattices

Multilayered structures comprising alternating semiconductor materials are known as **multiquantum-well** (MQW) structures (Fig. 17.1-28). They can be fabricated so that the energy bandgap varies with position in any desired way (see, e.g., Fig. 14.1-12). A MQW structure can have any number of layers, from just a few to hundreds. As an example, a MQW structure with 100 layers, each of thickness ≈ 10 nm and containing some 40 atomic planes, has an overall thickness ≈ 1 μm. As discussed in Sec. 14.1D, if the energy barriers between adjacent wells are sufficiently thin so that electrons can readily tunnel through them, the discrete energy levels broaden into minibands, in which case the multiquantum-well structure is referred to as a **superlattice structure**. The transition from MQW subbands to superlattice minibands is analogous to the transition from discrete energy levels in an atom to energy bands in a solid as the atoms are brought into closer proximity and permitted to interact (see Figs. 14.1-9 and 14.1-10). Quantum wells and superlattices can also be created by spatially varying the doping of a material, thereby creating space-charge fields that form potential barriers.

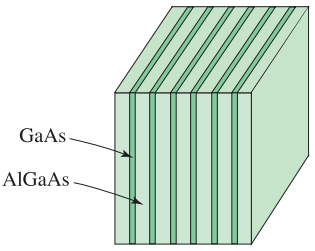


Figure 17.1-28 A MQW structure fabricated from alternating layers of materials of different bandgaps, such as AlGaAs and GaAs. These particular materials are often used to illustrate multiquantum-well structures because they can be lattice matched over a broad range of compositions [see Fig. 17.1-7(a)], which minimizes the strain between the two lattices, and because of their large difference in bandgap energies [see Table 17.1-2], which provides substantial carrier confinement. Other combinations of MQW materials commonly used in photonics include AlInAsSb/GaSb, AlInAs/InGaAs, AlInGaP/InGaP, GaN/InGaN, and $Al_xGa_{1-x}N/Al_yGa_{1-y}N$.

Biased Multiquantum-Well Structures

The energy-band diagrams of unbiased and biased multiquantum-well and superlattice structures are schematized in Fig. 17.1-29. The electric field causes the wells to become canted and alters the energy levels. In superlattice structures, the discrete energy levels smear into minibands. Multiquantum-well structures find use in a wide variety of photonic devices, such as active regions in light-emitting diodes, semiconductor optical amplifiers, and laser diodes (Secs. 18.1C, 18.2D, and 18.4, respectively). They also serve as photodetectors (Sec. 19.2C) and modulators (Sec. 21.5).

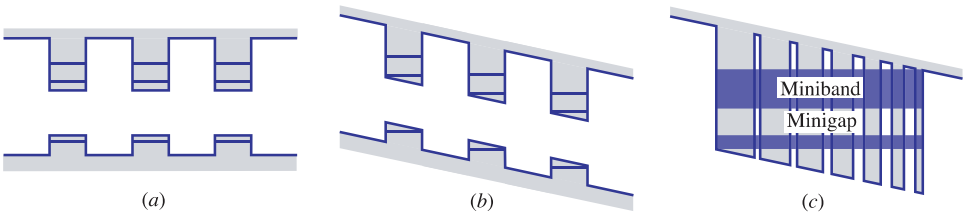


Figure 17.1-29 Energy-band diagrams of MQW and superlattice structures fabricated from alternating layers of materials with different bandgaps, such as AlGaAs and GaAs. (*a*) Unbiased MQW structure. (*b*) Biased MQW structure. (*c*) Biased superlattice structure with minibands and minigap.

Quantum Wires

As discussed in Sec. 14.1D, a semiconductor material that takes the form of a thin wire surrounded by a material of wider bandgap is called a **quantum-wire** structure (Fig. 17.1-30). The wire acts as a potential well that narrowly confines electrons and holes in two directions, x and y. Assuming that the wire has a rectangular cross section of area $d_1 d_2$, the energy–momentum relation in the conduction band is

$$E = E_c + E_{q1} + E_{q2} + \frac{\hbar^2 k^2}{2m_c}, \tag{17.1-38}$$

where

$$E_{q1} = \frac{\hbar^2 (q_1 \pi / d_1)^2}{2m_c}, \qquad E_{q2} = \frac{\hbar^2 (q_2 \pi / d_2)^2}{2m_c}, \qquad q_1, q_2 = 1, 2, 3, \ldots, \tag{17.1-39}$$

and k is the vector component in the z direction (along the axis of the wire).

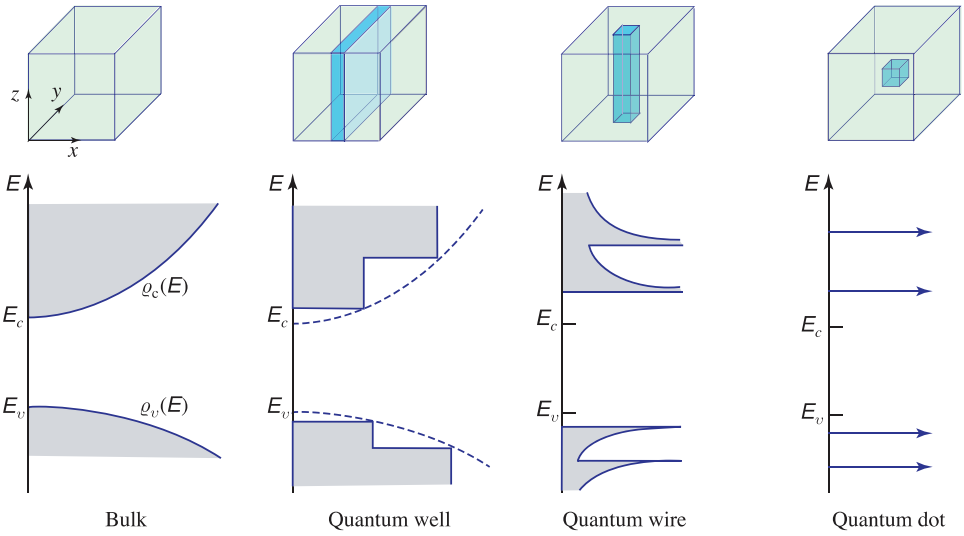


Figure 17.1-30 The density of states in different confinement configurations. The conduction and valence bands split into overlapping subbands that become successively narrower as the electron motion is restricted in a greater number of dimensions.

Each pair of quantum numbers (q_1, q_2) is associated with an energy subband that has a density of states $\varrho(k) = 1/\pi$ per unit length of the wire and therefore $1/\pi d_1 d_2$ per unit volume. The corresponding quantum-wire density of states (per unit volume), as a function of energy, is

$$\varrho_c(E) = \begin{cases} \dfrac{(1/d_1 d_2)\left(\sqrt{m_c}/\sqrt{2}\,\pi\hbar\right)}{\sqrt{E - E_c - E_{q1} - E_{q2}}}, & E > E_c + E_{q1} + E_{q2} \\ 0, & \text{otherwise,} \end{cases}$$

$$q_1, q_2 = 1, 2, 3, \ldots. \quad (17.1\text{-}40)$$

These are decreasing functions of energy, as illustrated in Fig. 17.1-30. The incorporation of quantum wires in devices will be discussed in Sec. 18.4B.

Quantum Dots

A **quantum-dot** structure narrowly confines the electrons in all three directions within a region that can be modeled as a box of volume $d_1 d_2 d_3$ (Sec. 14.1D). The energy is therefore quantized to

$$E = E_c + E_{q1} + E_{q2} + E_{q3}, \quad (17.1\text{-}41)$$

where

$$E_{q1} = \frac{\hbar^2(q_1\pi/d_1)^2}{2m_c}, \qquad E_{q2} = \frac{\hbar^2(q_2\pi/d_2)^2}{2m_c}, \qquad E_{q3} = \frac{\hbar^2(q_3\pi/d_3)^2}{2m_c},$$

$$q_1, q_2, q_3 = 1, 2, 3, \ldots. \quad (17.1\text{-}42)$$

The allowed energy levels are discrete and well separated so that the density of states is represented by a sequence of delta functions at the allowed energies, as illustrated in Fig. 17.1-30. Quantum dots are thus often called artificial atoms. Even though they contain enormous numbers of strongly interacting natural atoms, the discrete energy levels of the quantum dot can, in principle, be chosen at will by proper design. Quantum-dot-based devices are considered in Chapters 18 and 19.

17.2 INTERACTIONS OF PHOTONS WITH CHARGE CARRIERS

We proceed to consider some of the basic optical properties of semiconductors, with an emphasis on the processes of absorption and emission important in the operation of photonic devices. This domain of study is known as **semiconductor optics**.

A. Photon Interactions in Bulk Semiconductors

A number of mechanisms can lead to the absorption and emission of photons in bulk semiconductors. The most important of these are:

- *Band-to-Band (Interband) Transitions.* An absorbed photon can result in an electron in the valence band making an upward transition to the conduction band, thereby creating an electron–hole pair [Fig. 17.2-1(*a*)]. Electron–hole recombination can result in the emission of a photon. Band-to-band transitions may be assisted by one or more phonons. A **phonon** is a quantum of the lattice vibrations associated with molecular or acoustic vibrations of the atoms in a material.

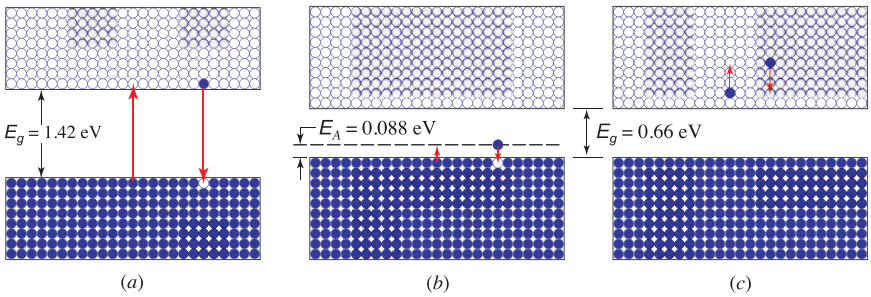


Figure 17.2-1 Examples of absorption and emission of photons in bulk semiconductors. (a) Band-to-band transitions in GaAs can result in the absorption or emission of photons of wavelength $\lambda_o < \lambda_g = hc_o/E_g = 0.87$ μm. (b) The absorption of a photon of wavelength $\lambda_A = hc_o/E_A = 14$ μm results in a valence-band to acceptor-level transition in Hg-doped Ge (Ge:Hg). (c) Free-carrier transitions within the conduction band of Ge.

- *Impurity-to-Band Transitions.* An absorbed photon can result in a transition between a donor (or acceptor) level and a band in a doped semiconductor. In a p-type material, for example, a low-energy photon can lift an electron from the valence band to the acceptor level, where it becomes trapped by an acceptor atom [Fig. 17.2-1(b)]. A hole is created in the valence band and the acceptor atom is ionized. Or a hole may be trapped by an ionized acceptor atom; the result is that the electron decays from its acceptor level to recombine with the hole. The energy may be released radiatively (in the form of an emitted photon) or nonradiatively (in the form of phonons). The transition may also be assisted by traps in defect states, as illustrated in Fig. 17.1-18.
- *Free-Carrier (Intraband) Transitions.* An absorbed photon can impart its energy to an electron in a given band, causing it to move higher within that band. An electron in the conduction band, for example, can absorb a photon and move to a higher energy level within the conduction band [Fig. 17.2-1(c)]. This is followed by thermalization, a process whereby the electron relaxes down to the bottom of the conduction band while releasing its energy in the form of phonons. The strength of free-carrier absorption is proportional to the carrier density; it decreases with photon energy as a power-law function.
- *Phonon Transitions.* Long-wavelength photons can release their energy by directly exciting lattice vibrations, i.e., by creating phonons.
- *Excitonic Transitions.* The absorption of a photon in a semiconductor can result in the formation of a free electron in the conduction band and a hole that rises to the top of the valence band, where its energy is minimized. The hole and electron can be bound together by their mutual Coulomb attraction to form an **exciton**; the attractive potential results in a reduction of the total energy of the electron and hole. This entity is much like a hydrogen atom in which a hole plays the role of the proton. Excitons typically have lifetimes that range from hundreds of picoseconds to nanoseconds. A photon may be emitted as a result of the electron and hole recombining, thereby annihilating the exciton.

These transitions all contribute to the overall absorption coefficient, which is displayed in Fig. 17.2-2 for Si and GaAs, and at greater magnification in Fig. 17.2-3 for a number of semiconductor materials. For photon energies greater than the bandgap energy E_g, the absorption is dominated by band-to-band transitions that form the basis of many photonic devices. The spectral region where the material changes from being relatively transparent ($h\nu < E_g$) to strongly absorbing ($h\nu > E_g$) is known as the **absorption edge**. Direct-bandgap semiconductors have a more abrupt absorption edge

than indirect-bandgap materials, as is apparent in Figs. 17.2-2 and 17.2-3.

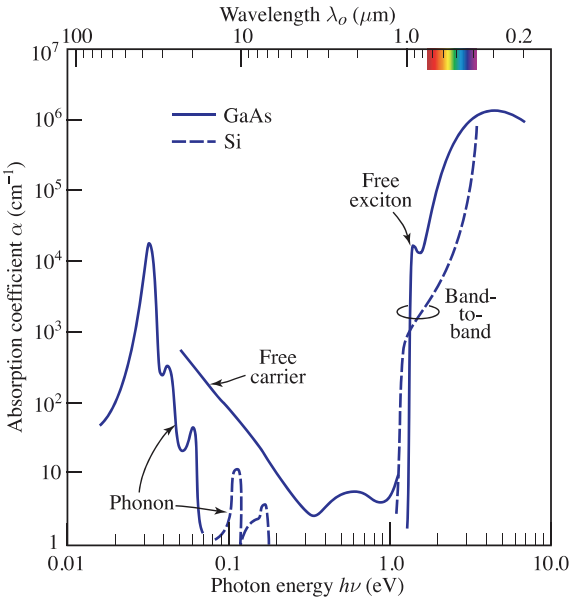


Figure 17.2-2 Observed optical absorption coefficient α versus photon energy and wavelength for Si and GaAs in thermal equilibrium at $T = 300°$ K. The bandgap energy E_g is 1.12 eV for Si and 1.42 eV for GaAs. Silicon is relatively transparent in the band $\lambda_o \approx 1.1$ to 12 μm, whereas intrinsic GaAs is relatively transparent in the band $\lambda_o \approx 0.87$ to 12 μm (Fig. 5.5-1).

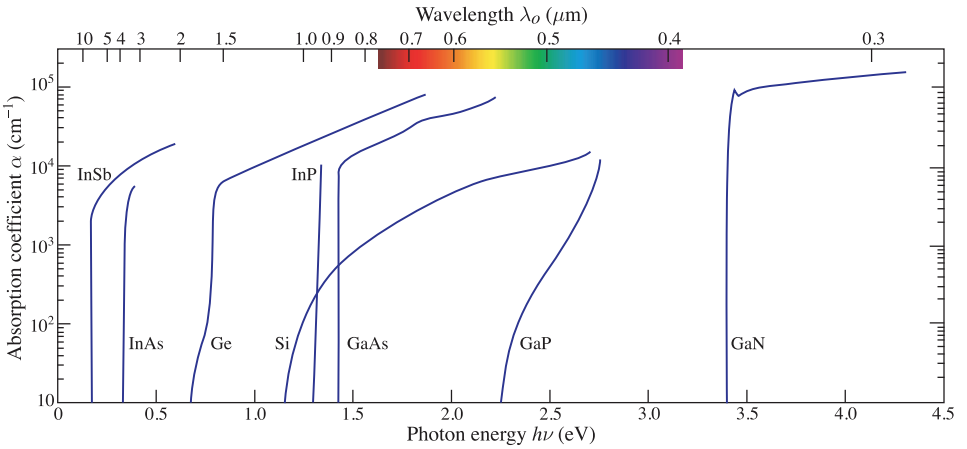


Figure 17.2-3 Absorption coefficient versus photon energy and wavelength for Ge, Si, GaAs, GaN, and several other III–V binary semiconductors at $T = 300°$ K, on an expanded scale. Direct- and indirect-bandgap materials follow different functional forms near the band edge.

B. Interband Transitions in Bulk Semiconductors

We proceed to develop a simple theory of direct interband (band-to-band) photon absorption and emission in bulk semiconductors, ignoring the other types of transitions.

Bandgap Wavelength

Direct interband absorption and emission can take place only at frequencies for which the photon energy $h\nu > E_g$. The minimum frequency ν necessary for this to occur is $\nu_g = E_g/h$, so that the corresponding maximum wavelength is $\lambda_g = c_o/\nu_g = hc_o/E_g$.

If the bandgap energy is given in eV (rather than in J), the bandgap wavelength $\lambda_g = hc_o/eE_g$ in μm turns out to be

$$\lambda_g \approx \frac{1.24}{E_g}. \qquad (17.2\text{-}1)$$

Bandgap Wavelength
λ_g (μm) and E_g (eV)

The quantity λ_g is known as the **bandgap wavelength** (or **cutoff wavelength**).

The bandgap wavelength λ_g, and its associated bandgap energy E_g, are provided in Table 17.1-2, and in Figs. 17.1-7 and 17.1-8, for a number of semiconductor materials of importance in photonics. III–V ternary and quaternary semiconductors of different compositions span a substantial range of bandgap wavelengths, from the mid infrared to the mid-ultraviolet.

Conditions for Photon Absorption and Emission

Electron excitation from the valence to the conduction band may be induced by the absorption of a photon of appropriate energy ($h\nu > E_g$ or $\lambda < \lambda_g$). An electron–hole pair is generated [Fig. 17.2-4(a)]. This adds to the concentration of mobile charge carriers and increases the conductivity of the material. The material behaves as a photoconductor with a conductivity proportional to the photon flux. This effect is used for the photodetection of light (Chapter 19).

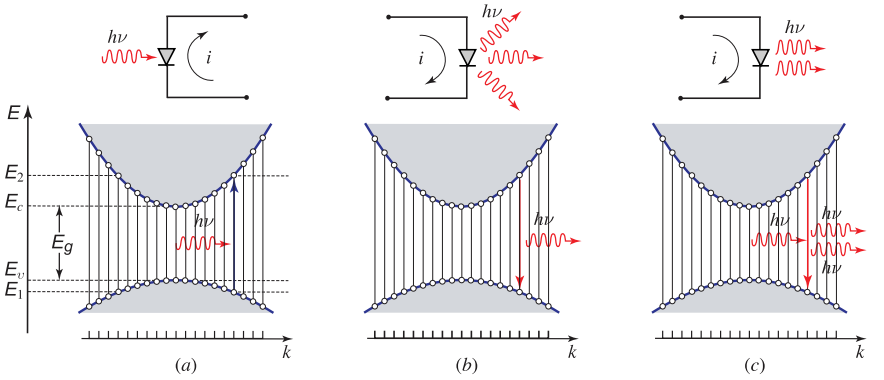


Figure 17.2-4 (a) The absorption of a photon results in the generation of an electron–hole pair. This process is used for the photodetection of light. (b) The recombination of an electron–hole pair results in the spontaneous emission of a photon. Light-emitting diodes (LEDs) operate on this basis. (c) Electron–hole recombination can be induced by a photon. The result is the stimulated emission of an identical photon. This is the underlying process responsible for the operation of semiconductor laser diodes.

Electron de-excitation from the conduction to the valence band (electron–hole recombination) may result in the spontaneous emission of a photon of energy $h\nu > E_g$ [Fig. 17.2-4(b)], or in the stimulated emission of a photon [Fig. 17.2-4(c)] when a photon of energy $h\nu > E_g$ is initially present (Sec. 14.3). Spontaneous emission is the underlying phenomenon on which the light-emitting diode is based (Sec. 18.1). Stimulated emission is responsible for the operation of semiconductor optical amplifiers and laser diodes (Secs. 18.2–18.6).

The conditions under which interband absorption and emission take place are summarized as follows:

- *Conservation of Energy.* The absorption or emission of a photon of energy $h\nu$ requires that the energies of the two states involved in the interaction (say E_1 and E_2 in the valence band and conduction band, respectively, as depicted in Fig. 17.2-4) be separated by $h\nu$. Thus, for photon emission to occur by electron–hole recombination, for example, an electron occupying an energy level E_2 must interact with a hole occupying an energy level E_1, such that energy is conserved:

$$E_2 - E_1 = h\nu. \tag{17.2-2}$$

- *Conservation of Momentum.* Momentum must also be conserved in the process of photon emission/absorption, so that $p_2 - p_1 = h\nu/c = h/\lambda$, or $k_2 - k_1 = 2\pi/\lambda$. The magnitude of the photon momentum h/λ is, however, very small in comparison with the range of momentum values that electrons and holes can assume. The semiconductor E–k diagram extends to values of k of the order $2\pi/a$, where the lattice constant a is much smaller than the wavelength λ, so that $2\pi/\lambda \ll 2\pi/a$. The momenta of the electron and the hole participating in the interaction must therefore be approximately equal. This condition, $k_2 \approx k_1$, is called the **k-selection rule**. Transitions that obey this rule are represented in the E–k diagram (Fig. 17.2-4) by vertical lines, indicating that the change in k is negligible on the scale of the diagram.
- *Energies and Momenta of the Electron and Hole with Which a Photon Interacts.* As is apparent from Fig. 17.2-4, conservation of both energy and momentum requires that a photon of frequency ν interact with electrons and holes of specific energies and momenta determined by the semiconductor E–k relation. Using (17.1-3) and (17.1-4) to approximate this relation for a direct-bandgap semiconductor by two parabolas, and writing $E_c - E_v = E_g$, (17.2-2) may be written in the form

$$E_2 - E_1 = \frac{\hbar^2 k^2}{2m_v} + E_g + \frac{\hbar^2 k^2}{2m_c} = h\nu, \tag{17.2-3}$$

from which

$$k^2 = \frac{2m_r}{\hbar^2}(h\nu - E_g), \tag{17.2-4}$$

where

$$\frac{1}{m_r} = \frac{1}{m_v} + \frac{1}{m_c}. \tag{17.2-5}$$

Substituting (17.2-4) into (17.1-3) provides the energy levels E_1 and E_2 with which the photon interacts:

$$E_2 = E_c + \frac{m_r}{m_c}(h\nu - E_g), \tag{17.2-6}$$

$$E_1 = E_v - \frac{m_r}{m_v}(h\nu - E_g) = E_2 - h\nu. \tag{17.2-7}$$

In the special case when $m_c = m_v$, we obtain $E_2 = E_c + \frac{1}{2}(h\nu - E_g)$, as required by symmetry.

- *Optical Joint Density of States.* We now determine the density of states $\varrho(\nu)$ with which a photon of energy $h\nu$ interacts under conditions of energy and momentum conservation in a direct-bandgap semiconductor. This quantity incorporates the density of states in both the conduction and valence bands and is known as the **optical joint density of states**. The one-to-one correspondence between E_2 and

ν embodied in (17.2-6) permits $\varrho(\nu)$ to be related to the density of states $\varrho_c(E_2)$ in the conduction band by use of the incremental relation $\varrho_c(E_2)\,dE_2 = \varrho(\nu)\,d\nu$, from which $\varrho(\nu) = (dE_2/d\nu)\varrho_c(E_2)$, so that

$$\varrho(\nu) = \frac{hm_r}{m_c}\,\varrho_c(E_2). \tag{17.2-8}$$

Using (17.1-7) and (17.2-6), we finally obtain the number of interacting states per unit volume per unit frequency,

$$\varrho(\nu) = \frac{(2m_r)^{3/2}}{\pi\hbar^2}\sqrt{h\nu - E_g}\,, \qquad h\nu \geq E_g, \tag{17.2-9}$$

(17.2-9) Optical Joint Density of States

which is sketched in Fig. 17.2-5. The one-to-one correspondence between E_1 and ν in (17.2-7), together with $\varrho_v(E_1)$ from (17.1-8), results in an expression for $\varrho(\nu)$ identical to (17.2-9).

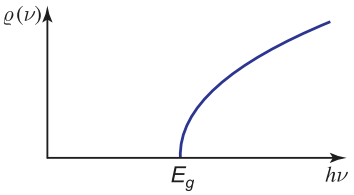


Figure 17.2-5 The density of states with which a photon of energy $h\nu$ interacts increases with $h\nu - E_g$ in accordance with a square-root law.

- *Photon Absorption Is Not Unlikely in an Indirect-Bandgap Semiconductor.* The energy and momentum conservation required for photon absorption in an indirect-bandgap semiconductor is readily accommodated by means of a two-step process (Fig. 17.2-6). The electron is first excited to a high energy level within the conduction band by a k-conserving vertical transition. It then quickly relaxes to the bottom of the conduction band by a process called **thermalization**, in which its momentum is transferred to phonons. The generated hole behaves similarly. Since the process occurs sequentially, it does not require the simultaneous presence of three bodies and is thus not unlikely in indirect- bandgap semiconductors. Indeed, Si and Ge are widely used as photodetector materials (Chapter 19), as are direct-bandgap semiconductors such as AlGaAs and InGaAs.

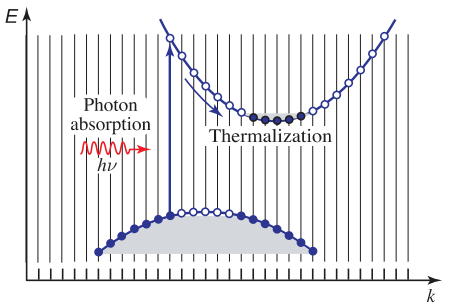


Figure 17.2-6 Photon absorption in an indirect-bandgap semiconductor via a vertical (k-conserving) transition. The photon generates an excited electron in the conduction band, leaving behind a hole in the valence band. The electron and hole then undergo fast transitions — to the lowest and highest available levels in the conduction and valence bands, respectively, releasing their energy in the form of phonons. Since the process is sequential, it is not unlikely.

- *Photon Emission Is Unlikely in an Indirect-Bandgap Semiconductor.* Radiative electron–hole recombination is unlikely in an indirect-bandgap semiconductor. This is because a transition from near the bottom of the conduction band to near the top of the valence band (where electrons and holes are most likely to reside, respectively) requires an exchange of momentum that cannot be accommodated by the emitted photon (Fig. 17.2-7). Momentum may be conserved, however, by the participation of phonons in the interaction. Phonons can carry relatively large momenta but typically have small energies ($\approx$ 0.01–0.1 eV; see Fig. 17.2-2), so that their transitions appear horizontal on the E–k diagram as portrayed in Fig. 17.2-7. The net result is that the k-selection rule is violated but momentum is conserved. However, because phonon-assisted emission involves the simultaneous participation of three bodies (electron, photon, and phonon), the probability of its occurrence is substantially reduced. Thus, Si, which is an indirect-bandgap semiconductor, has a substantially lower radiative recombination coefficient than does GaAs, which is a direct-bandgap semiconductor (Table 17.1-4). Silicon therefore does not emit light efficiently via interband transitions, whereas GaAs does. However, under special circumstances it is sometimes possible to elicit photon emission from an indirect-bandgap semiconductor; a particular situation where this can be achieved is considered in Example 17.2-1.

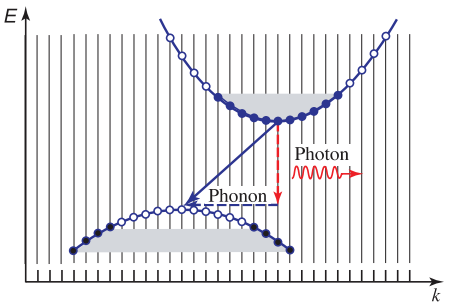


Figure 17.2-7 Photon emission via an interband transition in an indirect-bandgap semiconductor. The recombination of an electron near the bottom of the conduction band with a hole near the top of the valence band requires the exchange of energy and momentum. The energy may be carried off by a photon, but one or more phonons are required to conserve momentum. Such a simultaneous multiparticle interaction has a reduced likelihood of taking place.

EXAMPLE 17.2-1. ***Photon Emission from Indirect-Bandgap Germanium.*** Under some circumstances, the application of an externally applied mechanical stress can result in an attendant strain in a material that results in a crossover from indirect- to direct-bandgap behavior. Germanium is one such material, as evidenced by the operation of the Ge laser diode. An electrically pumped, Fabry–Perot Ge heterostructure uses a combination of mechanical strain and n-type doping to achieve laser action.[†] The tensile strain in the plane of the layer serves to transform the indirect bandgap of Ge to a direct bandgap by equalizing the energies of the direct-bandgap (Γ) valley and the energetically lowest indirect-bandgap (L) valley in its band structure. A biaxial tensile strain of 0.2% is introduced by forming a heterojunction of Ge and Si at high temperature; upon cooling the Ge becomes strained because of the thermal-expansion mismatch between the two materials. The germanium is doped with phosphorus at a concentration of 4×10^{19} cm^{-3}, which provides sufficient gain to overcome the losses associated with an electrically pumped device. The width of the pump current pulses lies between 20 μs and 100 ms, and the device is operated with a duty cycle of 4% to avoid undue heating. Room-temperature operation yields 1 mW of multimode optical output power at a wavelength that is tunable between 1.5 and 1.7 μm. The operation of the Ge heterostructure laser demonstrates that bandgap engineering can, in principle, be used to elicit recombination radiation from an inherently indirect-bandgap material.

[†] R. E. Camacho-Aguilera, Y. Cai, N. Patel, J. T. Bessette, M. Romagnoli, L. C. Kimerling, and J. Michel, An Electrically Pumped Germanium Laser, *Optics Express*, vol. 20, pp. 11316–11320, 2012.

C. Absorption, Emission, and Gain in Bulk Semiconductors

We now proceed to determine the probability densities of a photon of energy $h\nu$ being emitted or absorbed by a bulk semiconductor material in a direct interband transition. Conservation of energy and momentum, in the form of (17.2-4), (17.2-6), and (17.2-7), determines the energies E_1 and E_2, and the momentum $\hbar k$, of the electrons and holes with which the photon may interact. Three factors determine these probability densities, as discussed below:

1. Occupancy probabilities
2. Transition probabilities
3. Optical joint density of states

Occupancy Probabilities

The occupancy conditions for photon emission and absorption by means of transitions between the discrete energy levels E_2 and E_1 are stated as follows:

Emission condition: A conduction-band state of energy E_2 is filled (with an electron) *and* a valence-band state of energy E_1 is empty (i.e., filled with a hole).

Absorption condition: A conduction-band state of energy E_2 is empty *and* a valence-band state of energy E_1 is filled.

The probabilities that these occupancy conditions are satisfied for various values of E_2 and E_1 are determined from the appropriate Fermi functions $f_c(E)$ and $f_v(E)$ associated with the conduction and valence bands of a semiconductor in quasi-equilibrium. Thus, the probability $f_e(\nu)$ that the emission condition is satisfied for a photon of energy $h\nu$ is the product of the probabilities that the upper state is filled and that the lower state is empty since these are independent events, i.e.,

$$\boxed{f_e(\nu) = f_c(E_2)\,[1 - f_v(E_1)].} \tag{17.2-10}$$

The energies E_1 and E_2 are related to ν by (17.2-6) and (17.2-7). Similarly, the probability $f_a(\nu)$ that the absorption condition is satisfied is

$$\boxed{f_a(\nu) = [1 - f_c(E_2)]\,f_v(E_1).} \tag{17.2-11}$$

EXERCISE 17.2-1

Requirement for the Photon Emission Rate to Exceed the Absorption Rate.

(a) For a bulk semiconductor in thermal equilibrium, show that $f_e(\nu)$ is always smaller than $f_a(\nu)$ so that the rate of photon emission cannot exceed the rate of photon absorption.

(b) For a semiconductor in quasi-equilibrium ($E_{fc} \neq E_{fv}$), with radiative transitions occurring between a conduction-band state of energy E_2 and a valence-band state of energy E_1 with the same value of k, show that emission is more likely than absorption if the separation between the quasi-Fermi levels is larger than the photon energy, i.e., if

$$\boxed{E_{fc} - E_{fv} > h\nu.} \tag{17.2-12}$$

Condition for Net Emission

What does this condition imply about the locations of E_{fc} relative to E_c, and E_{fv} relative to E_v?

Transition Probabilities

Satisfying the emission/absorption occupancy condition does not assure that the emission/absorption actually takes place. These processes are governed by the probabilistic laws of interaction between photons and atomic systems examined at length in Secs. 14.3A–14.3C (see also Exercise 14.3-1). As they relate to semiconductors, these laws are generally expressed in terms of emission into (or absorption from) a narrow band of frequencies between ν and $\nu + d\nu$:

Summary

A radiative transition between two discrete energy levels E_1 and E_2 is characterized by a transition cross section $\sigma(\nu) = (\lambda^2/8\pi t_{\rm sp})g(\nu)$, where ν is the frequency, $t_{\rm sp}$ is the effective spontaneous lifetime, and $g(\nu)$ is the lineshape function [centered about the transition frequency $\nu_0 = (E_2 - E_1)/h$, with transition linewidth $\Delta\nu$ and with unity area]. In semiconductors, the radiative electron–hole recombination lifetime $\tau_{\rm r}$, which was discussed in Sec. 17.1D, plays the role of $t_{\rm sp}$ so that

$$\sigma(\nu) = \frac{\lambda^2}{8\pi\tau_{\rm r}}\, g(\nu)\,. \tag{17.2-13}$$

- If the occupancy condition for emission is satisfied, the probability density (per unit time) for the spontaneous emission of a photon into any of the available radiation modes in the narrow frequency band between ν and $\nu+d\nu$ is

$$P_{\rm sp}(\nu)\, d\nu = \frac{1}{\tau_{\rm r}}\, g(\nu)\, d\nu. \tag{17.2-14}$$

- If the occupancy condition for emission is satisfied *and* a mean spectral photon-flux density ϕ_ν (photons per unit time per unit area per unit frequency) at frequency ν is present, the probability density (per unit time) for the stimulated emission of one photon into the narrow frequency band between ν and $\nu + d\nu$ is

$$W_i(\nu)\, d\nu = \phi_\nu\, \sigma(\nu)\, d\nu = \phi_\nu \frac{\lambda^2}{8\pi\tau_{\rm r}}\, g(\nu)\, d\nu. \tag{17.2-15}$$

- If the occupancy condition for absorption is satisfied *and* a mean spectral photon-flux density ϕ_ν at frequency ν is present, the probability density for the absorption of one photon from the narrow frequency band between ν and $\nu + d\nu$ is also given by (17.2-15).

Since each transition has a different central frequency ν_0, and since we are considering a collection of such transitions, we explicitly label the central frequency of the transition by writing $g(\nu)$ as $g_{\nu 0}(\nu)$. In semiconductors the homogeneously broadened lineshape function $g_{\nu 0}(\nu)$ associated with a pair of energy levels generally has its origin in electron–phonon collision broadening. It therefore typically exhibits a Lorentzian lineshape [see (14.3-34) and (14.3-38)] of width $\Delta\nu \approx 1/\pi T_2$, where the electron–phonon collision time T_2 is of the order of picoseconds. If $T_2 = 1$ ps, for example, then $\Delta\nu = 318$ GHz, corresponding to an energy width $h\Delta\nu \approx 1.3$ meV. The radiative lifetime broadening of the levels is negligible in comparison with collisional broadening.

Overall Emission and Absorption Transition Rates

For a pair of energy levels separated by $E_2 - E_1 = h\nu_0$, the rates of spontaneous emission, stimulated emission, and absorption of photons of energy $h\nu$ (in units of photons/s-Hz-cm^3 of the semiconductor material), at the frequency ν, are obtained as follows: The appropriate transition probability density $P_{\rm sp}(\nu)$ or $W_i(\nu)$ [as provided in (17.2-14) or (17.2-15)] is multiplied by the appropriate occupation probability $f_e(\nu_0)$ or $f_a(\nu_0)$ [as given in (17.2-10) or (17.2-11)], and by the density of states that can interact with the photon $\varrho(\nu_0)$ [as set forth in (17.2-9)]. The overall transition rate for all allowed frequencies is then calculated by integrating over ν_0.

The rate of spontaneous emission at frequency ν, for example, is given by

$$r_{\rm sp}(\nu) = \int [(1/\tau_{\rm r}) g_{\nu 0}(\nu)]\, f_e(\nu_0)\, \varrho(\nu_0)\, d\nu_0. \tag{17.2-16}$$

When the collision-broadened width $\Delta\nu$ is substantially less than the width of the product $f_e(\nu_0)\varrho(\nu_0)$, which is the usual situation, $g_{\nu 0}(\nu)$ may be approximated by $\delta(\nu - \nu_0)$, whereupon the transition rate simplifies to $r_{\rm sp}(\nu) = (1/\tau_{\rm r})\varrho(\nu) f_e(\nu)$. The rates of stimulated emission and absorption are obtained in a similar fashion, and the following formulas result:

$$r_{\rm sp}(\nu) = \frac{1}{\tau_{\rm r}}\, \varrho(\nu) f_e(\nu) \tag{17.2-17}$$

$$r_{\rm st}(\nu) = \phi_\nu \frac{\lambda^2}{8\pi\tau_{\rm r}}\, \varrho(\nu) f_e(\nu) \tag{17.2-18}$$

$$r_{\rm ab}(\nu) = \phi_\nu \frac{\lambda^2}{8\pi\tau_{\rm r}}\, \varrho(\nu) f_a(\nu). \tag{17.2-19}$$

Emission and Absorption Rates

These equations, together with (17.2-9)–(17.2-11), permit the rates of spontaneous emission, stimulated emission, and absorption arising from direct interband transitions (photons/s-Hz-cm^3) to be calculated in the presence of a mean spectral photon-flux density ϕ_ν (photons/s-Hz-cm^2). The products $\varrho(\nu) f_e(\nu)$ and $\varrho(\nu) f_a(\nu)$ are analogous to the products of the lineshape function and atomic number densities in the upper and lower levels, $g(\nu)N_2$ and $g(\nu)N_1$, respectively, used in Chapters 14–16 to study emission and absorption in atomic systems.

The determination of the occupancy probabilities $f_e(\nu)$ and $f_a(\nu)$ requires knowledge of the quasi-Fermi levels E_{fc} and E_{fv}. It is via the control of these two parameters (by the application of an external bias to a p–n junction, for example) that the emission and absorption rates are modified to produce semiconductor photonic devices that carry out different functions. Equation (17.2-17) is the basic result that describes the operation of the light-emitting diode (LED), a semiconductor source based on spontaneous emission (Sec. 18.1). Equation (17.2-18) is applicable to semiconductor optical amplifiers and laser diodes, which operate on the basis of stimulated emission (Secs. 18.2–18.6). Equation (17.2-19) is appropriate for semiconductor detectors that function by means of photon absorption (see Sec. 19.1B).

Spontaneous-Emission Spectral Intensity in Thermal Equilibrium

A semiconductor in thermal equilibrium has only a single Fermi function so that (17.2-10) becomes $f_e(\nu) = f(E_2)[1 - f(E_1)]$. If the Fermi level lies within the bandgap, away from the band edges by at least several times kT, use may be made of the

exponential approximations to the Fermi functions, $f(E_2) \approx \exp[-(E_2 - E_f)/kT]$ and $1 - f(E_1) \approx \exp[-(E_f - E_1)/kT]$, whereupon $f_e(\nu) \approx \exp[-(E_2 - E_1)/kT]$, i.e.,

$$f_e(\nu) \approx \exp\left(-\frac{h\nu}{kT}\right). \tag{17.2-20}$$

Substituting (17.2-9) for $\varrho(\nu)$ and (17.2-20) for $f_e(\nu)$ into (17.2-17) therefore provides

$$\boxed{r_{\text{sp}}(\nu) \approx D_0 \sqrt{h\nu - E_g}\, \exp\left(-\frac{h\nu - E_g}{kT}\right), \qquad h\nu \geq E_g,} \tag{17.2-21}$$

where

$$D_0 = \frac{(2m_r)^{3/2}}{\pi\hbar^2\tau_{\text{r}}} \exp\left(-\frac{E_g}{kT}\right) \tag{17.2-22}$$

is a parameter that increases with temperature at an exponential rate.

The spontaneous emission rate (17.2-21), which is plotted versus $h\nu$ in Fig. 17.2-8, comprises two factors: a function associated with the density of states that increases as the square-root of $h\nu - E_g$, and an exponentially decreasing function of $h\nu - E_g$ arising from the Fermi function. The spontaneous emission rate can be increased by augmenting $f_e(\nu)$. In accordance with (17.2-10), this can be achieved by purposely causing the material to depart from thermal equilibrium in such a way that $f_c(E_2)$ is made large and $f_v(E_1)$ is made small. This assures an abundance of *both* electrons and holes, which is the desired condition for the operation of an LED, as discussed in Sec. 18.1.

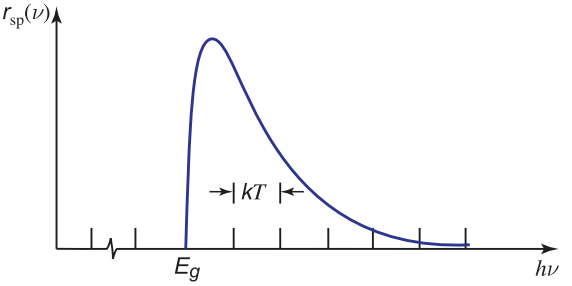


Figure 17.2-8 Spectral intensity of the direct interband spontaneous emission rate $r_{\text{sp}}(\nu)$ (photons/s-Hz-cm^3) from a semiconductor in thermal equilibrium, as a function of $h\nu$. The spectrum has a low-frequency cutoff at $\nu = E_g/h$ and extends over a range of frequencies of approximate width $2kT/h$.

Gain Coefficient in Quasi-Equilibrium

The net gain coefficient $\gamma_0(\nu)$ corresponding to the rates of stimulated emission and absorption in (17.2-18) and (17.2-19) is determined by taking a cylinder of unit area and incremental length dz, and assuming that a mean spectral photon-flux density is directed along its axis (see Fig. 15.1-1). If $\phi_\nu(z)$ and $\phi_\nu(z) + d\phi_\nu(z)$ are the mean spectral photon-flux densities entering and leaving the cylinder, respectively, $d\phi_\nu(z)$ must be the mean spectral photon-flux density emitted from within the cylinder. The

incremental number of photons, per unit time per unit frequency per unit area, is simply the number of photons gained, per unit time per unit frequency per unit volume $[r_{\rm st}(\nu) - r_{\rm ab}(\nu)]$, multiplied by the thickness of the cylinder dz. Hence, $d\phi_\nu(z) = [r_{\rm st}(\nu) - r_{\rm ab}(\nu)]\, dz$. Substituting the rates set forth in (17.2-18) and (17.2-19) leads to

$$\frac{d\phi_\nu(z)}{dz} = \frac{\lambda^2}{8\pi\tau_{\rm r}}\, \varrho(\nu)\, [f_e(\nu) - f_a(\nu)]\, \phi_\nu(z) = \gamma_0(\nu)\, \phi_\nu(z). \tag{17.2-23}$$

The net gain coefficient is therefore

$$\boxed{\gamma_0(\nu) = \frac{\lambda^2}{8\pi\tau_{\rm r}}\, \varrho(\nu)\, f_g(\nu),} \tag{17.2-24}$$

Gain Coefficient

where the **Fermi inversion factor** $f_g(\nu)$ takes the form

$$f_g(\nu) \equiv f_e(\nu) - f_a(\nu) = f_c(E_2) - f_v(E_1), \tag{17.2-25}$$

as may be understood from (17.2-10) and (17.2-11), with E_1 and E_2 related to ν by (17.2-6) and (17.2-7). Comparing (17.2-24) with (15.1-4) reveals that $\varrho(\nu)\, f_g(\nu)$ in the semiconductor system plays the role of $Ng(\nu)$ in the atomic system. Using (17.2-9), the gain coefficient may be cast in the form

$$\gamma_0(\nu) = D_1 \sqrt{h\nu - E_g}\, f_g(\nu), \qquad h\nu > E_g \tag{17.2-26a}$$

with

$$D_1 = \frac{\sqrt{2}\, m_r^{3/2} \lambda^2}{h^2 \tau_{\rm r}}. \tag{17.2-26b}$$

The sign and spectral form of the Fermi inversion factor $f_g(\nu)$ are governed by the quasi-Fermi levels E_{fc} and E_{fv}, which in turn depend on the state of excitation of the carriers in the semiconductor. As shown in Exercise 17.2-1, this factor is positive (corresponding to a population inversion and net gain) only when $E_{fc} - E_{fv} > h\nu$. When the semiconductor is pumped to a sufficiently high level by means of an external source of power, this condition may be satisfied and net gain achieved, as we shall see in Sec. 18.2. This reflects the physics underlying the operation of semiconductor optical amplifiers and laser diodes.

Absorption Coefficient in Thermal Equilibrium

A semiconductor in thermal equilibrium has only a single Fermi level $E_f = E_{fc} = E_{fv}$, so that

$$f_c(E) = f_v(E) = f(E) = \frac{1}{\exp[(E - E_f)/kT] + 1}. \tag{17.2-27}$$

The factor $f_g(\nu) = f_c(E_2) - f_v(E_1) = f(E_2) - f(E_1) < 0$, and therefore the gain coefficient $\gamma_0(\nu)$ is always negative [since $E_2 > E_1$ and $f(E)$ decreases monotonically with E]. This is true whatever the location of the Fermi level E_f. Thus, a semiconductor in thermal equilibrium, whether it be intrinsic or doped, always attenuates light. The attenuation (absorption) coefficient, $\alpha(\nu) = -\gamma_0(\nu)$, is therefore

$$\boxed{\alpha(\nu) = D_1 \sqrt{h\nu - E_g}\, \big[f(E_1) - f(E_2)\big],} \tag{17.2-28}$$

Absorption Coefficient

where E_2 and E_1 are given by (17.2-6) and (17.2-7), respectively, and D_1 is given by (17.2-26b).

If E_f lies within the bandgap but away from the band edges by an energy of at least several times kT, then $f(E_1) \approx 1$ and $f(E_2) \approx 0$ so that $[f(E_1) - f(E_2)] \approx 1$. In that case, the direct interband contribution to the absorption coefficient is

$$\alpha(\nu) \approx \frac{\sqrt{2}\, c^2 m_r^{3/2}}{\tau_r} \frac{1}{(h\nu)^2} \sqrt{h\nu - E_g}\,. \tag{17.2-29}$$

Equation (17.2-29) is plotted in Fig. 17.2-9 for GaAs, using the following parameters: $n = 3.6$, $m_c = 0.07\, m_0$, $m_v = 0.50\, m_0$, $m_0 = 9.1 \times 10^{-31}$ kg, a doping level such that $\tau_r = 0.4$ ns (this differs from that given in Table 17.1-4 because of the difference in doping level), $E_g = 1.42$ eV, and a temperature such that $[f(E_1) - f(E_2)] \approx 1$. As the temperature increases, $f(E_1) - f(E_2)$ decreases below unity and the absorption coefficient set forth in (17.2-28) is reduced.

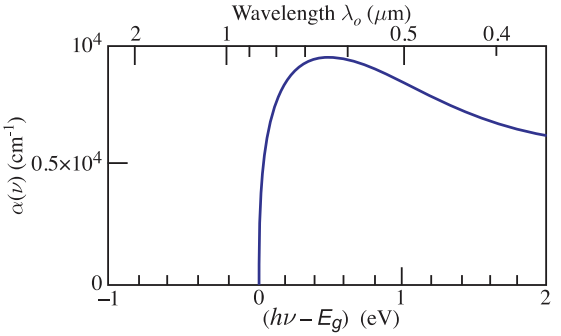


Figure 17.2-9 Calculated absorption coefficient $\alpha(\nu)$ (cm^{-1}) resulting from direct interband transitions, as a function of the photon energy $h\nu$ (eV) and the wavelength λ_o (μm), for GaAs. This curve should be compared with the empirical result displayed in Fig. 17.2-3, which encompasses all absorption mechanisms.

In accordance with (17.2-29), absorption near the band edge in a direct-bandgap semiconductor should follow the functional form $\sqrt{h\nu - E_g}$. However, the sharp onset of absorption at $h\nu = E_g$ is an idealization. As is evident in Fig. 17.2-3, direct-bandgap semiconductors generally exhibit an exponential absorption tail, known as the **Urbach tail**, with a characteristic width $\approx kT$ that extends slightly into the forbidden band. This is associated with thermal and static disorder in the crystal arising from several factors, including phonon-assisted absorption, randomness in the doping distribution, and variations in material composition. Absorption near the band edge in indirect-bandgap semiconductors (e.g., Ge, Si, and GaP in Fig. 17.2-3) generally follows the functional form $(h\nu - E_g)^2$ rather than the square-root relation applicable for direct-bandgap semiconductors.

EXERCISE 17.2-2

Wavelength of Maximum Interband Absorption. Use (17.2-29) to determine the (free-space) wavelength λ_p at which the absorption coefficient of a semiconductor in thermal equilibrium is maximum. Calculate the value of λ_p for GaAs. Note that this result applies only to absorption mediated by direct interband transitions.

D. Photon Interactions in Quantum-Confined Structures

Multiquantum-well and superlattice structures were considered in Sec. 17.1G. The photon interactions in these structures bear a considerable resemblance to those for bulk semiconductors (Sec. 17.2A). Several mechanisms play important roles in absorption and emission in quantum-confined structures:

- Interband (band-to-band) transitions
- Excitonic transitions
- Intersubband transitions
- Miniband transitions

These are illustrated in Fig. 17.2-10 and discussed below.

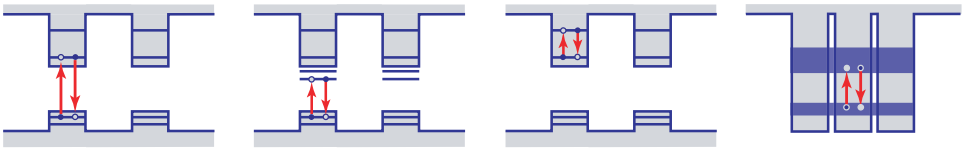


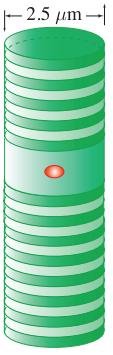
Figure 17.2-10 Photon absorption and emission in multiquantum-well structures. (a) Interband transitions. (b) Excitonic transitions. (c) Intersubband transitions. (d) Miniband transitions in a superlattice structure.

- *Interband Transitions.* Interband emission and absorption takes place between states in the valence and conduction bands [Fig. 17.2-10(a)], much as in bulk semiconductors. Because of quantum confinement, however, the optical joint density of states (17.2-9) must be replaced by (18.2-11). Interband transitions are responsible for the operation of MQW light-emitting diodes, superluminescent diodes, and laser diodes (Figs. 18.1-21, 18.2-11, and 18.4-4, respectively), as well as MQW electroabsorption modulators (Fig. 21.5-2).
- *Excitonic Transitions.* The 1D carrier confinement associated with MQW structures results in an increase in the exciton binding energy. This leads to strong **excitonic transitions**, even at $T = 300°$ K, as schematized in Fig. 17.2-10(b). Excitonic transitions play an important role in many quantum-confined devices, including MQW electroabsorption modulators (Fig. 21.5-2).
- *Intersubband Transitions.* Transitions that take place between energy levels within a single band of a MQW structure [Fig. 17.2-10(c)] are known as **intersubband transitions**. Devices that operate on the basis of these intraband transitions include the quantum-well quantum cascade laser [Fig. 18.4-8(a)] and the quantum-well infrared photodetector (Fig. 19.2-3). In the latter device, the absorption of a photon causes a transition from a bound energy level to the continuum. The picosecond carrier dynamics of intersubband systems offer large bandwidths.
- *Miniband Transitions.* In superlattices, the discrete MQW energy levels broaden into minibands that are separated by minigaps. Such **miniband transitions** [Fig. 17.2-10(d)] play a crucial role in the operation of superlattice quantum cascade lasers [Fig. 18.4-8(b)]. Such transitions, as well as intersubband transitions, exhibit fast relaxation and large nonlinearities, and are therefore appealing for applications such as all-optical switching and demultiplexing.

E. Quantum-Dot Single-Photon Emitters

Quantum dots, which can be excited optically or electrically, emit only one photon at a time; the emitted photons are thus separated from each other in time (antibunched) and the light is generally sub-Poisson (Sec. 13.3C). When embedded in photonic structures such as microcavities, 2D materials, and semiconductor heterostructures, quantum dots are thus useful for fabricating **single-photon emitters (SPEs)**. Efficient, on-demand sources of pure single photons that are highly indistinguishable are important for implementing quantum information processing, communications, computing, and cryptography in the form of scalable systems. Several examples of quantum-dot single-photon emitters are provided below. Other approaches can also be employed for creating SPEs, including the use of diamond defect centers, single-walled carbon nanotubes, and defects in 2D materials.

EXAMPLE 17.2-2. ***Silicon-Photonics Quantum-Dot Emitter.*** The confinement of carriers in a quantum dot results in a reduction of their positional uncertainty Δx. Since $\Delta x\,\Delta k \geq \frac{1}{2}$, in accordance with (A.2-6) of Appendix A, this is accompanied by a concomitant increase in the wavenumber uncertainty Δk. The increase in Δk obviates the need for phonons to take part in radiative recombination. The use of quantum-dot structures in this context is analogous to the incorporation of nitrogen impurities at sharply localized positions in indirect-bandgap GaP to make GaP:N LEDs (Sec. 18.1C). The small size of the quantum dot therefore measurably enhances radiative recombination via interband transitions in an indirect-bandgap semiconductor such as Si. Furthermore, surface passivation enhances the radiative rate via induced surface-localized excitons. As a result, light emission from silicon nanoparticles and porous silicon becomes possible.



EXAMPLE 17.2-3. ***Quantum-Dot/Micropillar Single-Photon Emitter.*** A single InAs/GaAs self-assembled quantum dot embedded in a 2.5-μm-diameter, cryogenically cooled micropillar microcavity (Sec. 11.4B) efficiently generates indistinguishable photons of high purity (one and only one photon is emitted at a time) via resonance fluorescence. By virtue of the micropillar's small cavity volume and high quality factor ($Q = 6124$) the Purcell spontaneous-emission enhancement factor provided in (14.3-48) is $F_P \approx 6$. A single-mode optical fiber efficiently coupled to the micropillar microcavity funnels $\approx 3.7\times10^6$ single photons/s out of the device. Excitation is provided by 25-nW, 3-ps near-infrared pulses at $\lambda_o = 897$ nm, resonant with the microcavity, at a repetition rate of 81 MHz. The photon extraction efficiency is 66% while the overall system efficiency is 4.5%. The source is about a factor of ten brighter than that provided by a two-photon heralded device based on spontaneous parametric downconversion (Sec. 13.3D).

EXAMPLE 17.2-4. ***Single-Photon Emission from 2D Materials.*** The large bandgap ($E_g \approx 6$ eV) of hexagonal boron nitride (h-BN), an insulating 2D material, facilitates its use as a host for quantum-dot single-photon emitters. Advantages of using this 2D host include: 1) room-temperature operation; 2) the ability to localize the emitter; and 3) the elimination of decoherence and background luminescence from a third dimension. Semiconducting 2D materials can also serve as single-photon emitters. Deterministic arrays of hundreds of photoluminescent single-photon emitters can be fabricated by depositing monolayers of transition-metal dichalcogenides such as WSe_2 and WS_2 onto silica substrates patterned with arrays of nanopillars (150-nm diameter $\times$ 100-nm height). The nanopillars create localized material deformations that accommodate the quantum confinement of excitons and serve as the loci of photoluminescent single-photon emissions. When excited by green light, these cryogenic devices emit at wavelengths ranging from the red to the near infrared.

F. Refractive Index

The ability to control the refractive index of a semiconductor is important in the design of many photonic devices, particularly those that make use of optical waveguides, laser diodes, and integrated photonics. Semiconductor materials are dispersive, so that the refractive index is dependent on the wavelength. Indeed, the refractive index is related to the absorption coefficient $\alpha(\nu)$ inasmuch as the real and imaginary parts of the susceptibility must satisfy the Kramers–Kronig relations (Sec. 5.5B and Sec. B.1 of Appendix B). The group index and refractive index for GaAs, calculated from the Sellmeier equation discussed in Sec. 5.5C, are displayed in Fig. 17.2-11. The refractive index depends on temperature and doping level.

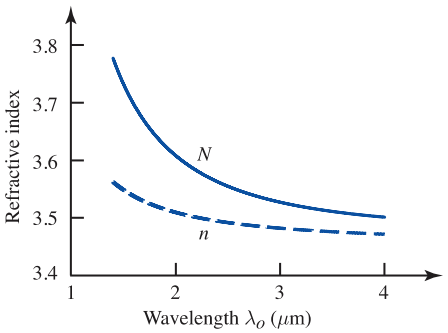


Figure 17.2-11 Refractive index n and group index N for GaAs as a function of the free-space wavelength λ_o. The results are determined from the Sellmeier equation provided in Table 5.5-1.

The refractive indices of selected elemental and binary bulk semiconductors, under specific conditions and near the bandgap wavelength, are provided in Table 17.2-1. The refractive indices of ternary and quaternary semiconductors can be approximated via linear interpolation between the refractive indices of their components.

Table 17.2-1 Refractive indices of selected semiconductor materials.[a]

Material	Refractive Index
Elemental semiconductors	
Ge	4.0
Si	3.5
III–V binary semiconductors	
AlN	2.2
AlP	3.0
AlAs	3.2
AlSb	3.8
GaN	2.5
GaP	3.3
GaAs	3.6
GaSb	4.0
InN	3.0
InP	3.5
InAs	3.8
InSb	4.2

[a]Results reported are for photon energies near the bandgap energy of the material ($h\nu \approx E_g$) and at $T = 300^\circ$ K.

READING LIST

Semiconductor Materials, Growth, and Characterization

J. E. Ayers, T. Kujofsa, P. Rago, and J. Raphael, *Heteroepitaxy of Semiconductors: Theory, Growth, and Characterization*, CRC Press/Taylor & Francis, 2nd ed. 2017.

J. Orton and T. Foxon, *Molecular Beam Epitaxy: A Short History*, Oxford University Press, 2015.

D. K. Schroder, *Semiconductor Material and Device Characterization*, Wiley–IEEE, 3rd ed. 2015.

E. Prati and T. Shinada, eds., *Single-Atom Nanoelectronics*, CRC Press/Taylor & Francis, 2013.

O. Oda, *Compound Semiconductor Bulk Materials and Characterizations*, Volume 2, World Scientific, 2012.

P. M. Koenraad and M. E. Flatté, Single Dopants in Semiconductors, *Nature Materials*, vol. 10, pp. 91–100, 2011.

M. Razeghi, *The MOCVD Challenge: A Survey of GaInAsP-InP and GaInAsP-GaAs for Photonic and Electronic Device Applications*, CRC Press/Taylor & Francis, 2010.

Graphene and 2D-Material Photonics

S. Cahangirov, H. Sahin, G. Le Lay, and A. Rubio, *Introduction to the Physics of Silicene and Other 2D Materials*, Springer-Verlag, 2017.

P. Avouris, T. F. Heinz, and T. Low, eds., *2D Materials: Properties and Devices*, Cambridge University Press, 2017.

P. Ajayan, P. Kim, and K. Banerjee, Two-dimensional van der Waals Materials, *Physics Today*, vol. 69, no. 9, pp. 38–44, 2016.

C. Jagadish and E. R. Weber, eds., *Semiconductors and Semimetals*, F. Iacopi, J. J. Boeckl, and C. Jagadish, eds., Volume 95, *2D Materials*, Academic Press/Elsevier, 2016.

E. L. Wolf, *Graphene: A New Paradigm in Condensed Matter and Device Physics*, Oxford University Press, 2016.

P. A. D. Gonçalves and N. M. R. Peres, *An Introduction to Graphene Plasmonics*, World Scientific, 2016.

Z. Sun, A. Martinez, and F. Wang, Optical Modulators with 2D Layered Materials, *Nature Photonics*, vol. 10, pp. 227–238, 2016.

A. C. Ferrari *et al.*, Science and Technology Roadmap for Graphene, Related Two-Dimensional Crystals, and Hybrid Systems, *Nanoscale*, vol. 7, pp. 4598–4810, 2015.

F. Xia, H. Wang, D. Xiao, M. Dubey, and A. Ramasubramaniam, Two-Dimensional Material Nanophotonics, *Nature Photonics*, vol. 8, pp. 899–907, 2014.

A. K. Geim, Nobel Lecture: Random Walk to Graphene, *Reviews of Modern Physics*, vol. 83, pp. 851–862, 2011.

K. S. Novoselov, Nobel Lecture: Graphene: Materials in the Flatland, *Reviews of Modern Physics*, vol. 83, pp. 837–849, 2011.

Semiconductor Physics, Optics, and Devices

See also the reading lists in Chapters 18 and 19.

C. Jagadish and E. R. Weber, eds., *Semiconductors and Semimetals*, Z. Mi and C. Jagadish, eds., Volume 96, *III-Nitride Semiconductor Optoelectronics*, Academic Press/Elsevier, 2017.

M. Grundmann, *The Physics of Semiconductors: An Introduction Including Nanophysics and Applications*, Springer-Verlag, 3rd ed. 2016.

M. Rudan, *Physics of Semiconductor Devices*, Springer-Verlag, 2015.

B. G. Streetman and S. Banerjee, *Solid State Electronic Devices*, Pearson, 7th ed. 2014.

G. Grosso and G. V. Parravicini, *Solid State Physics*, Academic Press/Elsevier, 2nd ed. 2014.

S. H. Simon, *The Oxford Solid State Basics*, Oxford University Press, paperback ed. 2013.

B. K. Ridley, *Quantum Processes in Semiconductors*, Oxford University Press, 5th ed. 2013.

W. Barford, *Electronic and Optical Properties of Conjugated Polymers*, Oxford University Press, 2nd ed. 2013.

S. M. Sze and M. K. Lee, *Semiconductor Devices: Physics and Technology*, Wiley, 3rd ed. 2012.

C. F. Klingshirn, *Semiconductor Optics*, Springer-Verlag, 4th ed. 2012.

M. Kira and S. W. Koch, *Semiconductor Quantum Optics*, Cambridge University Press, 2012.

S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*, Pearson, 2nd ed. 2012.

S. Adachi, *The Handbook on Optical Constants of Semiconductors: In Tables and Figures*, World Scientific, 2012.

W. Brütting and C. Adachi, eds., *Physics of Organic Semiconductors*, Wiley–VCH, 2nd ed. 2012.

D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles*, McGraw–Hill, 4th ed. 2011.

T. Yoshimura, *Thin-Film Organic Photonics: Molecular Layer Deposition and Applications*, CRC Press/Taylor & Francis, 2011.

J. Chu and A. Sher, *Device Physics of Narrow Gap Semiconductors*, Springer-Verlag, 2010.

M. Fox, *Optical Properties of Solids*, Oxford University Press, paperback 2nd ed. 2010.

P. Yu and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties*, Springer-Verlag, 4th ed. 2010.

S. L. Chuang, *Physics of Photonic Devices*, Wiley, 2nd ed. 2009.

H. Haug and S. W. Koch, *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, World Scientific, 5th ed. 2009.

H. Morkoç, *Handbook of Nitride Semiconductors and Devices*, Volume I: *Materials Properties, Physics and Growth*; Volume II: *Electronic and Optical Processes in Nitrides*; Volume III: *GaN-Based Optical and Electronic Devices*, Wiley–VCH, 2008.

T. Meier, P. Thomas, and S. W. Koch, *Coherent Semiconductor Optics: From Basic Concepts to Nanostructure Applications*, Springer-Verlag, 2007.

S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, Wiley, 3rd ed. 2006.

A. Moliton, *Optoelectronics of Molecules and Polymers*, Springer-Verlag, 2006, paperback ed. 2011.

C. Kittel, *Introduction to Solid State Physics*, Wiley, 8th ed. 2004, paperback ed. 2012.

K. K. Ng, *Complete Guide to Semiconductor Devices*, Wiley–IEEE, 2nd ed. 2002.

W. W. Chow and S. W. Koch, *Semiconductor-Laser Fundamentals: Physics of the Gain Materials*, Springer-Verlag, 1999, paperback ed. 2010.

M. Riordan and L. Hoddeson, The Origins of the *pn* Junction, *IEEE Spectrum*, vol. 34, no. 6, pp. 46–51, 1997.

P. T. Landsberg, *Recombination in Semiconductors*, Cambridge University Press, 1991, paperback ed. 2003.

L. Esaki, Long Journey into Tunneling (Nobel Lecture in Physics, 1973), in S. Lundqvist, ed., *Nobel Lectures in Physics 1971–1980*, World Scientific, 1992.

J. I. Pankove, *Optical Processes in Semiconductors*, Prentice Hall, 1971; Dover, paperback ed. 2010.

W. B. Shockley, Transistor Technology Evokes New Physics (Nobel Lecture in Physics, 1956), in *Nobel Lectures in Physics, 1942–1962*, World Scientific, 1998.

J. Bardeen, Semiconductor Research Leading to the Point Contact Transistor (Nobel Lecture in Physics, 1956), in *Nobel Lectures in Physics, 1942–1962*, World Scientific, 1998.

W. H. Brattain, Surface Properties of Semiconductors (Nobel Lecture in Physics, 1956), in *Nobel Lectures in Physics, 1942–1962*, World Scientific, 1998.

Quantum-Confined Materials and Nanostructures

See also the reading list on quantum-confined and microcavity devices in Chapter 18.

S. Y. Ren, *Electronic States in Crystals of Finite Size: Quantum Confinement of Bloch Waves*, Springer-Verlag, 2nd ed. 2017.

M. V. Fischetti and W. G. Vandenberghe, *Advanced Physics of Electron Transport in Semiconductors and Nanostructures*, Springer-Verlag, 2016.

P. Harrison and A. Valavanis, *Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures*, Wiley, 4th ed. 2016.

A. Zhang, G. Zheng, and C. M. Lieber, *Nanowires: Building Blocks for Nanoscience and Nanotechnology*, Springer-Verlag, 2016.

C. Jagadish and E. R. Weber, eds., *Semiconductors and Semimetals*, S. A. Dayeh, A. Fontcuberta i Morral, and C. Jagadish, eds., Volume 94, *Semiconductor Nanowires II: Properties and Applications*, Academic Press/Elsevier, 2016.

C. R. Kagan, E. Lifshitz, E. H. Sargent, and D. V. Talapin, Building Devices from Colloidal Quantum Dots, *Science*, vol. 353, aac5523, 2016.

Y. S. Zhao, ed., *Organic Nanophotonics: Fundamentals and Applications*, Springer-Verlag, 2015.

H. Ünlü and N. J. M. Horing, eds., *Low Dimensional Semiconductor Structures: Characterization, Modeling and Applications*, Springer-Verlag, 2013.

D. Vollath, *Nanomaterials: An Introduction to Synthesis, Properties and Applications*, Wiley–VCH, 2nd ed. 2013.

G. Cao and Y. Wang, *Nanostructures and Nanomaterials: Synthesis, Properties, and Applications*, World Scientific, 2nd ed. 2011.

F. W. Wise, ed., *Selected Papers on Semiconductor Quantum Dots*, SPIE Optical Engineering Press (Milestone Series Volume 180), 2005.

L. Esaki, A Bird's-Eye View on the Evolution of Semiconductor Superlattices and Quantum Wells, *IEEE Journal of Quantum Electronics*, vol. QE-22, pp. 1611–1624, 1986.

Single-Photon Emitters

X. He, N. F. Hartmann, X. Ma, Y. Kim, R. Ihly, J. L. Blackburn, W. Gao, J. Kono, Y. Yomogida, A. Hirano, T. Tanaka, H. Kataura, H. Htoon, and S. K. Doorn, Tunable Room-Temperature Single-Photon Emission at Telecom Wavelengths from sp^3 Defects in Carbon Nanotubes, *Nature Photonics*, vol. 11, pp. 577–582, 2017.

C. Palacios-Berraquero, D. M. Kara, A. R.-P. Montblanch, M. Barbone, P. Latawiec, D. Yoon, A. K. Ott, M. Loncar, A. C. Ferrari, and M. Atatüre, Large-Scale Quantum-Emitter Arrays in Atomically Thin Semiconductors, *Nature Communications* **8**, 15093 doi: 10.1038/ncomms15093, 2017.

I. Aharonovich and M. Toth, Quantum Emitters in Two Dimensions, *Science*, vol. 358, pp. 170–171, 2017.

I. Aharonovich, D. Englund, and M. Toth, Solid-State Single-Photon Emitters, *Nature Photonics*, vol. 10, pp. 631–641, 2016.

X. Ding, Y. He, Z.-C. Duan, N. Gregersen, M.-C. Chen, S. Unsleber, S. Maier, C. Schneider, M. Kamp, S. Höfling, C.-Y. Lu, and J.-W. Pan, On-Demand Single Photons with High Extraction Efficiency and Near-Unity Indistinguishability from a Resonantly Driven Quantum Dot in a Micropillar, *Physical Review Letters*, vol. 116, 020401, 2016.

N. Somaschi, V. Giesz, L. De Santis, J. C. Loredo, M. P. Almeida, G. Hornecker, S. L. Portalupi, T. Grange, C. Antón, J. Demory, C. Gómez, I. Sagnes, N. D. Lanzillotti-Kimura, A. Lemaítre, A. Auffeves, A. G. White, L. Lanco, and P. Senellart, Near-Optimal Single-Photon Sources in the Solid State, *Nature Photonics*, vol. 10, pp. 340–345, 2016.

P. Lodahl, S. Mahmoodian, and S. Stobbe, Interfacing Single Photons and Single Quantum Dots with Photonic Nanostructures, *Reviews of Modern Physics*, vol. 87, pp. 347–400, 2015.

M. G. Raymer and K. Srinivasan, Manipulating the Color and Shape of Single Photons, *Physics Today*, vol. 65, no. 11, pp. 32–37, 2012.

PROBLEMS

17.1-6 **Donor-Electron Ionization Energies and Radii.** Estimate the donor electron ionization energies E_D and Bohr radii a_0 for the semiconductor materials listed below (see Sec. 14.1A and Example 17.1-1). Comment, in each case, on the role of thermal excitations and use of the bulk relative permittivities in your calculations.

(a) A silicon crystal, with electron effective mass $m_c = 0.98\, m_0$ (Table 17.1-1) and relative permittivity $\epsilon/\epsilon_o = 12.3$ (Table 17.2-1).

(b) A gallium arsenide crystal, with electron effective mass $m_c = 0.07\, m_0$ (Table 17.1-1) and relative permittivity $\epsilon/\epsilon_o = 13$ (Table 17.2-1).

(c) A gallium nitride crystal, with electron effective mass $m_c = 0.20\, m_0$ (Table 17.1-1) and relative permittivity $\epsilon/\epsilon_o = 6.25$ (Table 17.2-1).

(d) A sample of Na^+-doped polyacetylene, an n-type conjugated polymer semiconductor with electron effective mass $m_c = m_0$ and relative permittivity $\epsilon/\epsilon_o = 3$. Organic light-emitting diodes (OLEDs) operate on the basis of recombination radiation from bound excitons.

17.1-7 **Fermi Level of an Intrinsic Semiconductor.** Given the expressions (17.1-12) and (17.1-13) for the thermal equilibrium carrier concentrations in the conduction and valence bands:

(a) Determine an expression for the Fermi level E_f of an intrinsic semiconductor and show that it falls exactly in the middle of the bandgap only when the effective mass of the electrons m_c is precisely equal to the effective mass of the holes m_c.

(b) Determine an expression for the Fermi level of a doped semiconductor as a function of the doping level and the Fermi level determined in (a).

17.1-8 **Electron–Hole Recombination Under Strong Injection.**Semiconductors!recombination Consider electron–hole recombination under conditions of strong carrier-pair injection such that the recombination lifetime can be approximated by $\tau = 1/\mathrm{r}\Delta\mathrm{n}$, where r is the recombination coefficient of the material and $\Delta\mathrm{n}$ is the injection-generated excess carrier concentration. Assuming that the source of injection R is set to zero at $t = t_0$, find an analytical expression for $\Delta\mathrm{n}(t)$, demonstrating that it exhibits power-law rather than exponential behavior.

17.1-9 **Bowing Parameters for Ternary Semiconductors.** The lattice constant of a ternary semiconductor alloy, say $\mathrm{A}_x\mathrm{B}_{1-x}\mathrm{C}$, typically varies linearly with the composition x, in accordance with Vegard's law. The bandgap energy E_g, on the other hand, usually varies nonlinearly with x so that a plot of bandgap energy versus lattice constant exhibits a bowed shape. This relation is usually modeled by the quadratic equation

$$E_g^{\mathrm{ABC}}(x) = E_g^{\mathrm{AC}}x + E_g^{\mathrm{BC}}(1-x) - \mathrm{b}x(1-x),$$

where b is called the **bowing parameter**. Use the curves provided in Figs. 17.1-7 and 17.1-8 to determine the bowing parameters for $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$, $\mathrm{GaAs}_{1-x}\mathrm{P}_x$, $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{N}$, $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}$, $\mathrm{Al}_x\mathrm{In}_{1-x}\mathrm{N}$, and $\mathrm{Hg}_x\mathrm{Cd}_{1-x}\mathrm{Te}$. What significance does the bowing parameter have with respect to lattice matching of the ternary compound to a substrate?

*17.1-10 **Energy Levels in a GaAs/AlGaAs Quantum Well.**

(a) Draw the energy-band diagram of a single-crystal multiquantum-well structure of GaAs/AlGaAs to scale on the energy axis when the AlGaAs has the composition $\mathrm{Al}_{0.3}\mathrm{Ga}_{0.7}\mathrm{As}$. The bandgap of GaAs, E_g(GaAs), is 1.42 eV; the bandgap of AlGaAs increases above that of GaAs by approximately 12.47 meV for each 1% increase in the Al composition. Because of the inherent characteristics of these two materials, the depth of the GaAs conduction-band quantum well is about 60% of the total conduction-plus-valence band quantum-well depths.

(b) Assume that a GaAs conduction-band well has depth as determined in (a) above and has precisely the same energy levels as the finite square well shown in Fig. 17.1-26(*b*), for which $(mV_0d^2/2\hbar^2)^{1/2} = 4$, where V_0 is the depth of the well. Find the total width d of the GaAs conduction-band well. The effective mass of an electron in the conduction band of GaAs is $m_c \approx 0.07\, m_0 = 0.64 \times 10^{-31}$ kg.

17.2-3 **Validity of the Approximation for Absorption/Emission Rates.** The derivation of the rate of spontaneous emission made use of the approximation $g_{\nu 0}(\nu) \approx \delta(\nu - \nu_0)$ in the course of evaluating the integral

$$r_{\mathrm{sp}}(\nu) = \int \left[\frac{1}{\tau_{\mathrm{r}}}\, g_{\nu 0}(\nu)\right] f_e(\nu_0)\, \varrho(\nu_0)\, d\nu_0.$$

(a) Demonstrate that this approximation is satisfactory for GaAs by plotting the functions $g_{\nu 0}(\nu)$, $f_e(\nu_0)$, and $\varrho(\nu_0)$ at $T = 300^\circ$ K and comparing their widths. GaAs is collisionally lifetime broadened with $T_2 \approx 1$ ps.

(b) Repeat (a) for the rate of absorption in thermal equilibrium.

17.2-4 **Peak Spontaneous Emission Rate in Thermal Equilibrium.**

(a) Determine the photon energy $h\nu_p$ at which the direct interband spontaneous emission rate from a semiconductor material in thermal equilibrium achieves its maximum value when the Fermi level lies within the bandgap and away from the band edges by at least several kT.

(b) Show that this peak rate (photons per sec per Hz per cm^3) is given by

$$r_{\rm sp}(\nu_p) = \frac{D_0}{\sqrt{2e}}\sqrt{kT} = \frac{2(m_r)^{3/2}}{\sqrt{e}\,\pi\hbar^2\tau_{\rm r}}\sqrt{kT}\,\exp\left(-\frac{E_g}{kT}\right).$$

(c) What is the effect of doping on this result?

(d) Assuming that $\tau_{\rm r} = 0.4$ ns, $m_c = 0.07\,m_0$, $m_v = 0.50\,m_0$, and $E_g = 1.42$ eV, find the peak rate in GaAs at $T = 300°$ K.

17.2-5 **Radiative Recombination Rate in Thermal Equilibrium.**

(a) Show that the direct interband spontaneous emission rate integrated over all emission frequencies (photons per sec per cm^3) is given by

$$\int_0^\infty r_{\rm sp}(\nu)\,d\nu = D_0\frac{\sqrt{\pi}}{2h}\,(kT)^{3/2} = \frac{(m_r)^{3/2}}{\sqrt{2}\,\pi^{3/2}\hbar^3\tau_{\rm r}}\,(kT)^{3/2}\exp\left(-\frac{E_g}{kT}\right),$$

provided that the Fermi level is within the semiconductor energy bandgap and away from the band edges. *Note*: $\int_0^\infty x^{1/2}e^{-\mu x}\,dx = (\sqrt{\pi}/2)\mu^{-3/2}$.

(b) Compare this with the approximate integrated rate obtained by multiplying the peak rate obtained in Prob. 17.2-4 by the approximate frequency width $2kT/h$ shown in Fig. 17.2-8.

(c) Using (17.1-15), set the phenomenological equilibrium radiative recombination rate $\mathtt{r_r np} = \mathtt{r_r n_i^2}$ (photons per second per cm^3) introduced in Sec. 17.1D equal to the direct interband result derived in (a) to obtain the expression for the radiative recombination coefficient

$$\mathtt{r_r} = \frac{\sqrt{2}\,\pi^{3/2}\hbar^3}{(m_c+m_v)^{3/2}}\,\frac{1}{(kT)^{3/2}\,\tau_{\rm r}}.$$

(d) Use the result in (c) to find the value of $\mathtt{r_r}$ for GaAs at $T = 300°$ K using $m_c = 0.07\,m_0$, $m_v = 0.5\,m_0$, and $\tau_{\rm r} = 0.4$ ns. Compare this with the value provided in Table 17.1-4 ($\mathtt{r_r} \approx 10^{-10}$ cm^3/s).

CHAPTER

18

LEDS AND LASER DIODES

18.1 LIGHT-EMITTING DIODES 789
- A. Injection Electroluminescence
- B. LED Characteristics
- C. Materials and Device Structures
- D. Silicon Photonics
- E. Organic LEDs
- F. LED Lighting

18.2 SEMICONDUCTOR OPTICAL AMPLIFIERS 817
- A. Gain and Bandwidth
- B. Pumping
- C. Heterostructures
- D. Quantum-Well Structures
- E. Superluminescent Diodes

18.3 LASER DIODES 831
- A. Amplification, Feedback, and Oscillation
- B. Power and Efficiency
- C. Spectral and Spatial Characteristics

18.4 QUANTUM-CONFINED LASERS 844
- A. Quantum-Well and Multiquantum-Well Lasers
- B. Quantum-Wire and Multiquantum-Wire Lasers
- C. Quantum-Dot and Multiquantum-Dot Lasers
- D. Quantum Cascade Lasers

18.5 MICROCAVITY LASERS 854
- A. Vertical-Cavity Surface-Emitting Lasers
- B. Microdisk and Microring Lasers
- C. Photonic-Crystal Lasers

18.6 NANOCAVITY LASERS 862

The operation of semiconductor laser diodes was reported nearly simultaneously in 1962 by independent research teams from the **General Electric** Corporation, **IBM** Corporation, and **Lincoln Laboratory** of the Massachusetts Institute of Technology.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Light can be emitted from a semiconductor material as a result of electron–hole recombination. Nevertheless, materials capable of emitting such light do not glow at room temperature because the concentrations of thermally excited electrons and holes are too small to produce discernible radiation. However, an external source of energy can be used to produce electron–hole pairs in sufficient numbers that they generate large amounts of spontaneous recombination radiation, causing the material to luminesce. A convenient way of achieving this is to forward bias a p–n junction, which fosters the injection of electrons and holes in the vicinity of the junction. The ensuing recombination radiation is called injection electroluminescence.

A light-emitting diode (LED) is a forward-biased p–n junction, usually fabricated from a direct-bandgap semiconductor material, that emits light via injection electroluminescence [Fig. 18.0-1(a)]. If the forward voltage is increased beyond a certain point, however, the number of electrons and holes in the junction region can become sufficiently large such that a population inversion is achieved, whereupon stimulated emission (i.e., emission induced by the presence of photons) becomes more prevalent than absorption. Under these conditions, the junction region may be used as a semiconductor optical amplifier (SOA) [Fig. 18.0-1(b)] or, with appropriate feedback, as a laser diode (LD) [Fig. 18.0-1(c)].

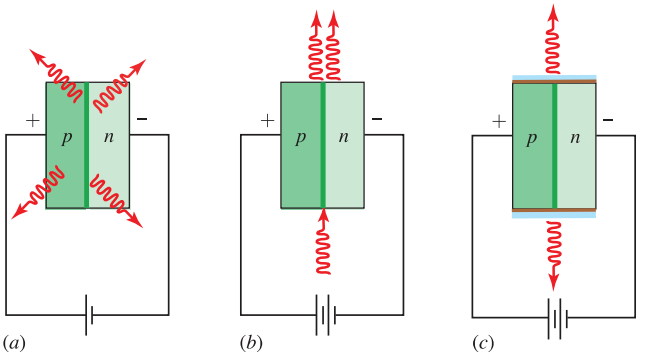


Figure 18.0-1 A forward-biased semiconductor p–n junction diode operated as: (a) a light-emitting diode (LED); (b) a semiconductor optical amplifier (SOA); and (c) a laser diode (LD).

Semiconductor photon sources in the form of both LEDs and LDs serve as highly efficient electronic-to-photonic transducers. They are indispensable in many applications by virtue of their small size, high brightness, high efficiency, high reliability, ruggedness, and durability. Visible LEDs have long been used for **indication applications** (in which the observer directly views the source); examples include indicator lights, mobile phones, signage, traffic signals, and backlighting. High-power visible LEDs are ubiquitous in **illumination applications** (in which the observer views the light scattered from objects illuminated by the source), such as architectural and street lighting, flashlights, and projection.

Infrared LEDs are often used in remote controls for consumer products such as optical mice, headphones, and keyboards. Ultraviolet LEDs are useful in applications such as water purification, surgical sterilization, resin curing, and printing. They are also used for the detection of chemical and biological agents, many of which fluoresce at particular wavelengths when exposed to ultraviolet light.

Laser diodes find extensive use in long-haul optical fiber communication systems, where it is a particular convenience that they can be readily modulated by controlling the injected current. They also find use in high-density optical data-storage systems

such as DVD players, and in scanning, reading, and high-resolution color-printing systems. Laser diodes are also employed in lidars and in directional lighting applications, such as automotive headlights. As discussed in Sec. 16.3A, banks of laser diodes are used to optically pump optical fiber amplifiers and solid-state lasers, thereby converting the relatively broadband, multimode laser-diode light into the narrowband, single-mode light emitted by diode-pumped solid-state (DPSS) lasers.

The advent of quantum-confined semiconductor lasers such as multiquantum-well, multiquantum-dot, and quantum cascade lasers, together with compact lasers such as vertical-cavity, microdisk, and nanolasers, has greatly facilitated the integration of lasers with other optical components in compact configurations, which in turn has opened the door to manifold new uses.

This Chapter

This chapter is devoted to the study of light-emitting diodes (Sec. 18.1), semiconductor optical amplifiers (Sec. 18.2), laser diodes (Sec. 18.3), quantum-confined lasers (Sec. 18.4), microcavity lasers (Sec. 18.5), and nanocavity lasers (Sec. 18.6). As background, we draw broadly on the material presented in Chapter 17. The theoretical treatments offered for semiconductor optical amplifiers and laser-diode oscillators closely parallel the analyses of laser amplifiers and laser oscillators provided in Chapters 15 and 16, respectively.

18.1 LIGHT-EMITTING DIODES

Electroluminescence is a phenomenon in which light is emitted by a material that is subjected to an electric field (Sec. 14.5). Injection electroluminescence, first observed in 1907, underlies the operation of light-emitting diodes, which are highly efficient devices capable of emitting light of just about any color. LEDs are highly important in a number of areas of photonics. We discuss the theory of injection electroluminescence in Sec. 18.1A, the characteristics of light-emitting diodes in Sec. 18.1B, representative materials and device structures in Sec. 18.1C, the use of indirect-bandgap silicon for generating light in Sec. 18.1D, organic LEDs in Sec. 18.1E, and LED lighting in Sec. 18.1F.

A. Injection Electroluminescence

Electroluminescence in Thermal Equilibrium

Electron–hole radiative recombination results in the emission of light from a semiconductor material. At room temperature the concentration of thermally excited electrons and holes is so small, however, that the generated photon flux is very small (Example 18.1-1).

EXAMPLE 18.1-1. ***Photon Emission from GaAs in Thermal Equilibrium.*** At room temperature, the intrinsic concentration of electrons and holes in GaAs is $\mathrm{n_i} \approx 1.8 \times 10^6$ cm^{-3} (Table 17.1-3). Since the radiative electron–hole recombination coefficient $\mathrm{r_r} \approx 10^{-10}$ cm^3/s under certain conditions (as specified in Table 17.1-4), the electroluminescence rate $\mathrm{r_r np} = \mathrm{r_r n_i^2} \approx 324$ photons/cm^3-s, as discussed in Sec. 17.1D. A 2-μm-thick layer of GaAs therefore produces a photon-flux density $\phi \approx 0.065$ photons/cm^2-s, which is negligible as may be understood by consulting Table 13.2-1 (light emitted from a layer of GaAs thicker than about 2 μm suffers reabsorption). Taking the photon energy $h\nu$ as the bandgap energy for GaAs, $E_g = 1.42$ eV or $1.42e = 2.27 \times 10^{-19}$ J, the emitted intensity turns out to be $I = h\nu\phi \approx 1.5 \times 10^{-20}$ W/cm^2.

If thermal equilibrium conditions are maintained, this intensity cannot be appreciably increased (or decreased) by doping the material. In accordance with the law of mass action provided in (17.1-17), the product np is fixed at $\mathrm{n_i^2}$ if the material is not too heavily doped so that the recombination rate $\mathrm{r_r np} = \mathrm{r_r n_i^2}$ depends on the doping level only through $\mathrm{r_r}$. An abundance of electrons *and* holes is required for a large recombination rate; in an n-type semiconductor n is large but p is small, whereas the converse is true in a p-type semiconductor.

Electroluminescence in the Presence of Carrier Injection

The photon emission rate can be appreciably increased by using external means to increase excess electron–hole pairs in the material. This may be accomplished, for example, by illuminating the material with light, but it is typically achieved by forward biasing a p–n junction diode, which serves to inject carrier pairs into the junction region. This process is illustrated in Fig. 17.1-21 and will be explained further in Sec. 18.1B. The photon emission rate may be calculated from the electron–hole pair injection rate R (pairs/cm^3-s), where R plays the role of the laser pumping rate (Sec. 15.2). The photon flux Φ (photons per second), generated within a volume V of the semiconductor material, is directly proportional to the carrier-pair injection rate (Fig. 18.1-1).

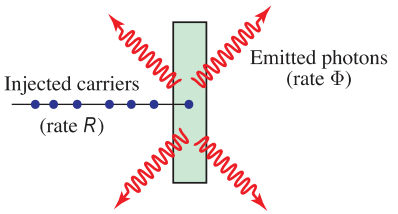


Figure 18.1-1 Spontaneous photon emission resulting from electron–hole radiative recombination, as might occur in a forward-biased p–n junction.

Denoting the equilibrium concentrations of electrons and holes in the absence of pumping as $\mathrm{n_0}$ and $\mathrm{p_0}$, respectively, we use $\mathrm{n} = \mathrm{n_0} + \Delta\mathrm{n}$ and $\mathrm{p} = \mathrm{p_0} + \Delta\mathrm{p}$ to represent the steady-state carrier concentrations in the presence of pumping (Sec. 17.1D). The excess electron concentration $\Delta\mathrm{n}$ is precisely equal to the excess hole concentration $\Delta\mathrm{p}$ because electrons and holes are produced in pairs. It is assumed that the excess electron–hole pairs recombine at a rate $1/\tau$, where τ is the overall (radiative and nonradiative) electron–hole recombination time. Under steady-state conditions, the generation (pumping) rate must precisely balance the recombination (decay) rate, so that $R = \Delta\mathrm{n}/\tau$. Thus, the steady-state excess-carrier concentration is proportional to the pumping rate, i.e.,

$$\Delta\mathrm{n} = R\tau. \tag{18.1-1}$$

For carrier injection rates that are sufficiently low, as explained in Sec. 17.1D, we have $\tau \approx 1/\mathrm{r}(\mathrm{n_0} + \mathrm{p_0})$ where r is the (radiative and nonradiative) recombination coefficient, so that $R \approx \mathrm{r}\Delta\mathrm{n}\,(\mathrm{n_0} + \mathrm{p_0})$.

Only radiative recombinations generate photons, however, and the internal quantum efficiency $\eta_i = \mathrm{r_r}/\mathrm{r} = \tau/\tau_\mathrm{r}$, defined in (17.1-28) and (17.1-30), accounts for the fact that only a fraction of the recombinations are radiative in nature. The injection of RV carrier pairs per second therefore leads to the generation of a photon flux $\Phi = \eta_i RV$ photons/s, i.e.,

$$\boxed{\Phi = \eta_i RV = \eta_i \frac{V\Delta\mathrm{n}}{\tau} = \frac{V\Delta\mathrm{n}}{\tau_\mathrm{r}}.} \tag{18.1-2}$$

The internal photon flux Φ is proportional to the carrier-pair injection rate R and therefore to the steady-state concentration of excess electron–hole pairs Δn.

The internal quantum efficiency η_i plays a crucial role in determining the performance of this electron-to-photon transducer. Direct-bandgap semiconductors are usually used to make LEDs (and laser diodes) because η_i is substantially larger than it is for indirect-bandgap semiconductors (e.g., at room temperature $\eta_i \approx 0.5$ for GaAs, whereas $\eta_i \approx 10^{-5}$ for Si, as shown in Table 17.1-4). The internal efficiency η_i depends on the doping, temperature, and defect concentration of the material.

EXAMPLE 18.1-2. ***Injection Electroluminescence Emission from GaAs.*** Under certain conditions, we have $\tau = 50$ ns and $\eta_i = 0.5$ for GaAs (Table 17.1-4), so that a steady-state excess concentration of injected electron–hole pairs $\Delta n = 10^{17}$ cm^{-3} will give rise to a photon flux concentration $\eta_i \Delta n/\tau \approx 10^{24}$ photons/cm^3-s. This corresponds to an optical power density $\approx 2.3 \times 10^5$ W/cm^3 for photons at the bandgap energy $E_g = 1.42$ eV. A 2-μm-thick slab of GaAs therefore produces an optical intensity of ≈ 46 W/cm^2, which is a factor of 10^{21} greater than the thermal-equilibrium value calculated in Example 18.1-1. Under these conditions the power emitted from a device of area 200 μm $\times$ 10 μm is ≈ 0.9 mW, which is substantial.

Spectral Intensity of Electroluminescence Photons

The spectral intensity of injection electroluminescence light may be determined by using the direct interband emission theory developed in Sec. 17.2. The rate of spontaneous emission $r_{sp}(\nu)$ (number of photons per second per Hz per unit volume), as provided in (17.2-17), is

$$r_{sp}(\nu) = \frac{1}{\tau_r} \varrho(\nu) f_e(\nu), \tag{18.1-3}$$

where τ_r is the radiative electron–hole recombination lifetime. The optical joint density of states for interaction with photons of frequency ν, as given in (17.2-9), is

$$\varrho(\nu) = \frac{(2m_r)^{3/2}}{\pi \hbar^2} \sqrt{h\nu - E_g}\,, \tag{18.1-4}$$

where m_r is related to the effective masses of the holes and electrons by $1/m_r = 1/m_v + 1/m_c$ [as given in (17.2-5)], and E_g is the bandgap energy. The emission condition [as given in (17.2-10)] provides

$$f_e(\nu) = f_c(E_2)[1 - f_v(E_1)], \tag{18.1-5}$$

which is the probability that a conduction-band state of energy

$$E_2 = E_c + \frac{m_r}{m_c}(h\nu - E_g) \tag{18.1-6}$$

is filled *and* a valence-band state of energy

$$E_1 = E_2 - h\nu \tag{18.1-7}$$

is empty, as provided in (17.2-6) and (17.2-7) and illustrated in Fig. 18.1-2.

Equations (18.1-6) and (18.1-7) guarantee that energy and momentum are conserved. The Fermi functions $f_c(E) = 1/\{\exp[(E - E_{fc})/kT] + 1\}$ and $f_v(E) = 1/\{\exp[(E - E_{fv})/kT] + 1\}$ that appear in (18.1-5), with quasi-Fermi levels E_{fc} and E_{fv}, apply to the conduction and valence bands, respectively, under conditions of quasi-equilibrium. The semiconductor parameters E_g, τ_r, m_v, and m_c, and the temperature

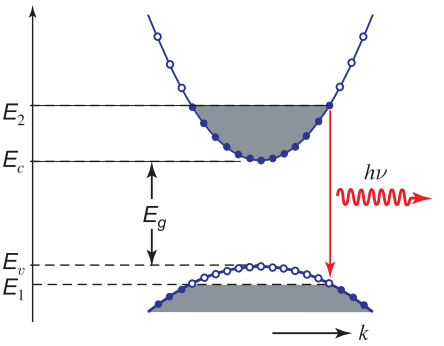


Figure 18.1-2 The spontaneous emission of a photon resulting from the recombination of an electron of energy E_2 with a hole of energy $E_1 = E_2 - h\nu$. The transition is represented by a vertical arrow because the momentum carried away by the photon, $h\nu/c$, is negligible on the scale of the figure.

T, determine the spectral distribution $r_{\rm sp}(\nu)$, given the quasi-Fermi levels E_{fc} and E_{fv}. These in turn are determined from the concentrations of electrons and holes given in (17.1-10) and (17.1-11):

$$\int_{E_c}^{\infty} \varrho_c(E) f_c(E)\, dE = \mathrm{n} = \mathrm{n}_0 + \Delta\mathrm{n}; \tag{18.1-8}$$

$$\int_{-\infty}^{E_v} \varrho_v(E)[1 - f_v(E)]\, dE = \mathrm{p} = \mathrm{p}_0 + \Delta\mathrm{n}. \tag{18.1-9}$$

The densities of states near the conduction- and valence-band edges are, respectively, as per (17.1-7) and (17.1-8),

$$\varrho_c(E) = \frac{(2m_c)^{3/2}}{2\pi^2\hbar^3}\sqrt{E - E_c}\,; \quad \varrho_v(E) = \frac{(2m_v)^{3/2}}{2\pi^2\hbar^3}\sqrt{E_v - E}\,, \tag{18.1-10}$$

where n_0 and p_0 are the concentrations of electrons and holes in thermal equilibrium (in the absence of injection), and $\Delta\mathrm{n} = R\tau$ is the steady-state injected-carrier concentration. For sufficiently weak injection, such that the Fermi levels lie within the bandgap and away from the band edges by several kT, the Fermi functions may be approximated by their exponential tails. The spontaneous photon flux (integrated over all frequencies) is then obtained from the spontaneous emission rate $r_{\rm sp}(\nu)$ by

$$\Phi = V\int_0^{\infty} r_{\rm sp}(\nu)\, d\nu = \frac{V(m_r)^{3/2}}{\sqrt{2}\,\pi^{3/2}\,\hbar^3\,\tau_{\rm r}}\,(kT)^{3/2}\,\exp\left(\frac{E_{fc} - E_{fv} - E_g}{kT}\right), \tag{18.1-11}$$

as is readily extrapolated from Prob. 17.2-5.

Increasing the pumping level R causes $\Delta\mathrm{n}$ to increase, which in turn moves E_{fc} toward (or further into) the conduction band, and E_{fv} toward (or further into) the valence band. This results in an increase in the probability $f_c(E_2)$ of finding the conduction-band state of energy E_2 filled with an electron, and the probability $1 - f_v(E_1)$ of finding the valence-band state of energy E_1 empty (filled with a hole). The net result is that the emission-condition probability $f_e(\nu) = f_c(E_2)[1 - f_v(E_1)]$ increases with R, thereby enhancing the spontaneous emission rate given in (18.1-3) and the spontaneous photon flux Φ given above.

EXERCISE 18.1-1

Quasi-Fermi Levels of a Pumped Semiconductor.

(a) Under ideal conditions at $T = 0°$ K, when there is no thermal electron–hole pair generation [Fig. 18.1-3(a)], show that the quasi-Fermi levels are related to the concentrations of injected

electron–hole pairs Δn by

$$E_{fc} = E_c + (3\pi^2)^{2/3}\frac{\hbar^2}{2m_c}(\Delta n)^{2/3} \tag{18.1-12a}$$

$$E_{fv} = E_v - (3\pi^2)^{2/3}\frac{\hbar^2}{2m_v}(\Delta n)^{2/3}, \tag{18.1-12b}$$

so that

$$E_{fc} - E_{fv} = E_g + (3\pi^2)^{2/3}\frac{\hbar^2}{2m_r}(\Delta n)^{2/3}, \tag{18.1-12c}$$

where $\Delta n \gg n_0, p_0$. Under these conditions all Δn electrons occupy the lowest allowed energy levels in the conduction band, and all Δp holes occupy the highest allowed levels in the valence band. Compare with the results of Exercise 17.1-3.

(b) Sketch the functions $f_e(\nu)$ and $r_{sp}(\nu)$ for two values of Δn. Given the effect of temperature on the Fermi functions, as illustrated in Fig. 18.1-3(*b*), determine the effect of increasing the temperature on $r_{sp}(\nu)$.

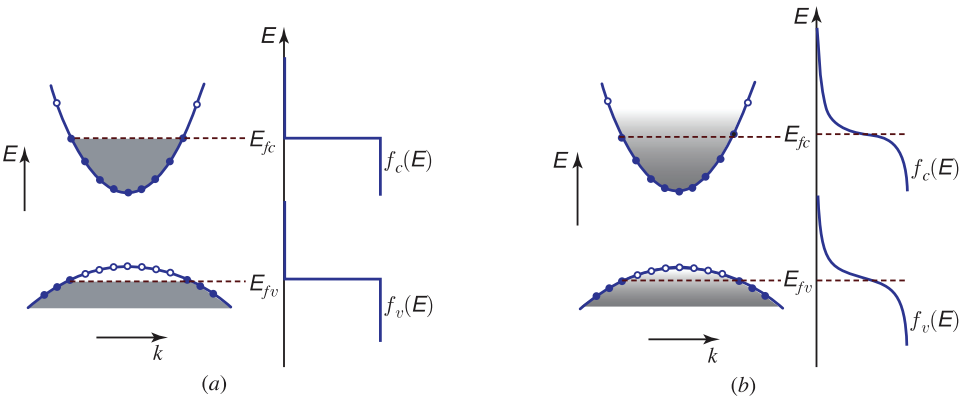


Figure 18.1-3 Energy bands and Fermi functions for a semiconductor in quasi-equilibrium (*a*) at $T = 0°$ K, and (*b*) at $T > 0°$ K.

EXERCISE 18.1-2

Spectral Intensity of Injection Electroluminescence under Weak Injection. For sufficiently weak injection, such that $E_c - E_{fc} \gg kT$ and $E_{fv} - E_v \gg kT$, the Fermi functions may be approximated by their exponential tails. Show that the luminescence rate can then be expressed as

$$r_{sp}(\nu) = D\sqrt{h\nu - E_g}\,\exp\left(-\frac{h\nu - E_g}{kT}\right), \quad h\nu \geq E_g, \tag{18.1-13a}$$

where

$$D = \frac{(2m_r)^{3/2}}{\pi\hbar^2\tau_r}\,\exp\left(\frac{E_{fc} - E_{fv} - E_g}{kT}\right) \tag{18.1-13b}$$

is an exponentially increasing function of the separation between the quasi-Fermi levels $E_{fc} - E_{fv}$. The spectral intensity of the spontaneous emission rate is shown in Fig. 18.1-4; it has precisely the same shape as the thermal-equilibrium spectral intensity shown in Fig. 17.2-8, but its magnitude is increased by the factor $D/D_0 = \exp[(E_{fc} - E_{fv})/kT]$, which can be very large in the presence of

injection. In thermal equilibrium $E_{fc} = E_{fv}$, so that (17.2-21) and (17.2-22) are recovered.

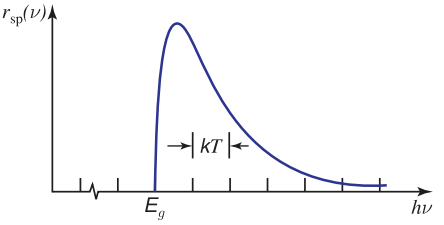


Figure 18.1-4 Spectral intensity of the direct interband injection-electroluminescence rate $r_{sp}(\nu)$ (photons per second per Hz per cm^3), versus $h\nu$, from (18.1-13), under conditions of weak injection.

EXERCISE 18.1-3

Electroluminescence Spectral Linewidth.

(a) Show that the spectral intensity of the emitted light described by (18.1-13) attains its peak value at a frequency ν_p determined by

$$h\nu_p = E_g + \tfrac{1}{2}kT. \qquad (18.1\text{-}14)$$

Peak Frequency

(b) Show that the full-width at half-maximum (FWHM) of the electroluminescence spectral intensity is

$$\Delta\nu \approx 1.8\,kT/h. \qquad (18.1\text{-}15)$$

Spectral Width (Hz)

The value of $\Delta\nu$ for active materials made of compound semiconductors can be larger than that specified in (18.1-15) by virtue of randomness in the chemical composition; this phenomenon is known as **alloy broadening**.

(c) Show that this width corresponds to a wavelength spread $\Delta\lambda \approx 1.8\lambda_p^2 kT/hc$, where $\lambda_p = c/\nu_p$. For kT expressed in eV and the wavelength expressed in μm, demonstrate that

$$\Delta\lambda \approx 1.45\,\lambda_p^2\,kT. \qquad (18.1\text{-}16)$$

(d) Calculate $\Delta\nu$ and $\Delta\lambda$ at $T = 300°$ K, for $\lambda_p = 0.8$ μm and $\lambda_p = 1.6$ μm.

B. LED Characteristics

As is clear from the foregoing discussion, the simultaneous availability of electrons and holes substantially enhances the flux of spontaneously emitted photons from a semiconductor. Electrons are abundant in n-type material, and holes are abundant in p-type material, but the generation of copious amounts of light requires that both electrons and holes be plentiful in the same region of space. This condition may be readily achieved in the junction region of a forward-biased p–n diode (Sec. 17.1E). As shown in Fig. 18.1-5, forward biasing causes holes from the p side and electrons from the n side to be forced into the common junction region by the process of minority carrier injection, where they recombine and emit photons.

The light-emitting diode (LED) is a *forward-biased* p–n *junction* with a large radiative recombination rate arising from injected minority carriers. The semiconductor material is usually *direct-bandgap* to ensure high quantum efficiency. In this section we determine the output power, as well as the spectral and spatial distributions of the light emitted from an LED, and derive expressions for the efficiency, responsivity, and response time.

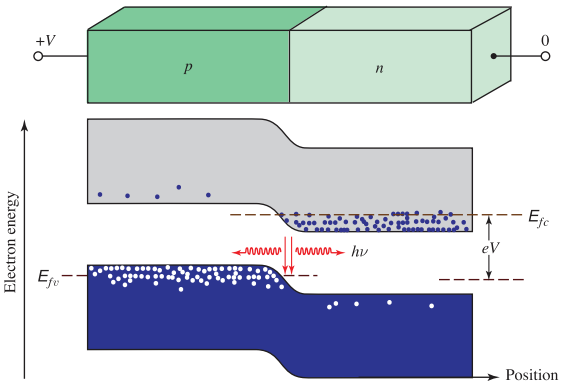


Figure 18.1-5 Energy-band diagram of a heavily doped p–n junction that is strongly forward biased by an applied voltage V (compare with the less strongly forward-biased energy-band diagram in Fig. 17.1-21). The dashed lines represent the quasi-Fermi levels, which are separated as a result of the bias. The simultaneous abundance of electrons and holes within the junction region results in strong electron–hole radiative recombination (injection electroluminescence).

Internal Photon Flux and Internal Efficiency

A schematic representation of a simple p–n homojunction diode is provided in Fig. 18.1-6. An injected DC current i leads to an increase in the steady-state carrier concentrations $\Delta \mathrm{n}$, which in turn result in radiative recombination in the active-region volume V.

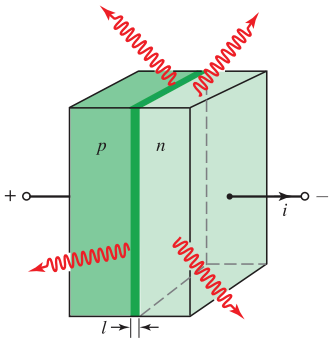


Figure 18.1-6 A simple forward-biased LED. The photons are emitted spontaneously from the junction region.

Since the total number of carriers per second passing through the junction region is i/e, where e is the magnitude of the electronic charge, the carrier injection (pumping) rate (carriers per second per cm^3) is simply

$$R = \frac{i/e}{V}. \tag{18.1-17}$$

Equation (18.1-1) provides that $\Delta \mathrm{n} = R\tau$, which results in a steady-state carrier concentration

$$\Delta \mathrm{n} = \frac{(i/e)\tau}{V}. \tag{18.1-18}$$

In accordance with (18.1-2), the **internal photon flux** Φ is then $\eta_i RV$, which, using (18.1-17), gives

$$\boxed{\Phi = \eta_i \frac{i}{e}.} \tag{18.1-19}$$

Internal Photon Flux

This simple and intuitively appealing formula governs the production of photons by electrons in an LED: a fraction η_i of the injected electron flux i/e (electrons per second)

is converted into photon flux. The **internal quantum efficiency** η_i is therefore simply the ratio of the generated photon flux to the injected electron flux.

The internal photon flux can be enhanced by making use of LEDs with double-heterostructure configurations (Sec. 17.1F), and, in particular, multiquantum-well (MQW) active regions (Sec. 17.1G). The benefit obtains because double heterostructures engender higher carrier concentrations, which enhances radiative recombination (the radiative lifetime τ_r is reduced) and thereby increases the internal quantum efficiency η_i [see (17.1-30), (17.1-31), and (18.1-19)]. To maximize η_i, the heterostructure confinement layers should be lattice matched to the active region.

Narrow quantum wells confine carriers even more tightly, further enhancing η_i. The number of quantum wells used in a device is frequently limited because of difficulties in populating all of them. To achieve good performance, it is important to make use of materials of the highest quality, which minimizes defect concentrations, and to avoid the presence of surfaces to which both carrier types have access, which minimizes nonradiative recombination.

Yet another approach for increasing η_i relies on making use of a **plasmonic LED**, in which metallic nanoparticles are embedded in a layer adjacent to a MQW active region. This engenders coupling between localized surface plasmons (LSPs) of the metallic nanoparticles (Sec. 8.2C), or of collective plasmonic resonances in periodic nanostructure arrays, and the light emitted from the proximate MQWs. The result can be a substantial enhancement of the spontaneous-emission rate $r_{sp}(\nu)$ via the Purcell effect (Sec. 14.3E), which in turn leads to an increase in the internal quantum efficiency η_i. The usefulness of this approach is evident in increased LED output power, including for devices that operate in the green. The polarization and directionality of the emitted light are also modified.

Extraction Efficiency

The photon flux generated in the junction is radiated uniformly in all directions; however, the flux that emerges from the device depends on the direction of emission. This is readily illustrated by considering the photon flux transmitted through a planar material along three possible ray directions, denoted A, B, and C in the geometry of Fig. 18.1-7:

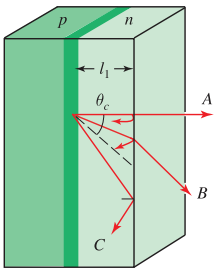


Figure 18.1-7 Not all light generated in an LED with a planar surface is able to emerge. Ray A is partly reflected. Ray B suffers more reflection. Ray C lies outside the critical angle and therefore undergoes total internal reflection, so that it is trapped in the structure.

- The photon flux traveling in the direction of ray A is attenuated by the factor

$$\eta_1 = \exp(-\alpha l_1), \tag{18.1-20}$$

where α is the absorption coefficient of the n-type material and l_1 is the distance from the junction to the surface of the device. Furthermore, for normal incidence, reflection at the semiconductor–air boundary permits only a fraction of the light,

$$\eta_2 = 1 - \frac{(n-1)^2}{(n+1)^2} = \frac{4n}{(n+1)^2}, \tag{18.1-21}$$

to be transmitted, where n is the refractive index of the semiconductor material [see Fresnel's equations (6.2-15)]. For GaAs, $n = 3.6$, so that $\eta_2 = 0.68$. The overall transmittance for the photon flux traveling in the direction of ray A is therefore $\eta_A = \eta_1\eta_2$.

- The photon flux traveling in the direction of ray B has farther to travel and therefore suffers a larger absorption; it also has greater reflection losses. Thus, $\eta_B < \eta_A$.
- The photon flux emitted along directions lying outside a cone of (critical) angle $\theta_c = \sin^{-1}(1/n)$, such as illustrated by ray C, suffers total internal reflection in an ideal material and is not transmitted [see (1.2-5)]. The area of the spherical cap atop this cone is $A = \int_0^{\theta_c} 2\pi r \sin\theta\, r\, d\theta = 2\pi r^2(1 - \cos\theta_c)$ while the area of the entire sphere is $4\pi r^2$. Thus, the fraction of the emitted light that lies within the solid angle subtended by this cone is $A/4\pi r^2$, so that

$$\eta_3 = \tfrac{1}{2}(1 - \cos\theta_c) = \tfrac{1}{2}\left(1 - \sqrt{1 - 1/n^2}\right) \approx 1/4n^2. \qquad (18.1\text{-}22)$$

For a material with refractive index $n = 3.6$, as an example, only 1.9% of the total generated photon flux can be transmitted. For a parallelepiped of refractive index $n > \sqrt{2}$, the ratio of isotropically radiated light energy that can emerge, to the total generated light energy, is $3[1 - \sqrt{1 - 1/n^2}\,]$, as shown in Exercise 1.2-6. However, some fraction of the photons emitted outside the critical angle can be absorbed and reemitted within this angle, so that in practice, η_3 may assume a value larger than that specified by (18.1-22). Loss and Fresnel reflection must also be incorporated for these rays.

The efficiency with which the internal photons can be extracted from the LED structure is known as the **extraction efficiency** η_e. Antireflection coatings (Exercise 7.1-1) can be used to reduce Fresnel reflection and thereby increase η_e.

EXERCISE 18.1-4

Extraction of Light from a Planar-Surface LED.

(a) Derive (18.1-22).

(b) Determine the critical angles for light escaping into air from: GaAs ($n = 3.6$), GaN ($n = 2.5$), and a transparent polymer ($n = 1.5$). Calculate the fraction of light that can be extracted in the three cases if absorption and Fresnel reflection are ignored.

(c) What is the enhancement in the fraction of extracted light that can be achieved if a planar GaAs LED is coated with a transparent polymer of refractive index $n = 1.5$, assuming that absorption and Fresnel reflection at the semiconductor-polymer boundary are ignored?

(d) Determine whether it might be useful to employ a material of intermediate refractive index (e.g., a polymer layer) to maximize the fraction of light emitted from the LED into air, if absorption is ignored but Fresnel reflection at both the semiconductor–polymer and polymer–air interfaces is accommodated.

The extraction efficiency can be enhanced in a multitude of ways. One approach involves selecting a geometry for the **LED die** (**LED chip**) that allows a greater fraction of the light to escape. A spherical dome surrounding a point source at its center, for example, permits all rays to escape, although they remain subject to Fresnel reflection. As illustrated in Fig. 18.1-8, several other geometries offer enhanced extraction

efficiencies in comparison with the parallelepiped: hemispherical domes, cylindrical structures (which have an escape ring along the perimeter in addition to the escape cone toward the top surface), inverted cones, and truncated inverted pyramids. However, geometries that entail complex processing steps are often avoided in practice because of increased manufacturing costs. Simple planar-surface-emitting LEDs are suitable when the intended viewing angle deviates little from the normal or when the light is coupled into an optical fiber, as it is in telecommunications applications.

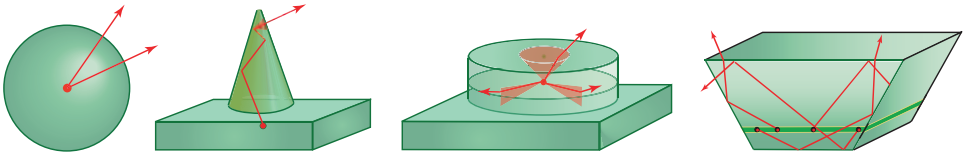


Figure 18.1-8 LED-die geometries that offer enhanced extraction efficiencies relative to the parallelepiped.

Another approach is to roughen the planar surface, which enhances the extraction efficiency by permitting rays beyond the critical angle to escape via scattering, as illustrated in Fig. 18.1-9. Indeed, an irregular surface appears automatically under certain growth conditions. Alternatively, the emission surface can be textured, such as with an array of microscopic cones or pyramids, or with nanoparticles.

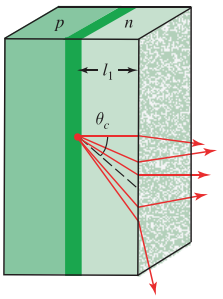


Figure 18.1-9 An LED with a roughened planar surface permits rays beyond the critical angle to escape, thereby increasing the extraction efficiency η_e.

The morphology of the light-emitting organs of some biological organisms, such as fireflies, serves to enhance light extraction by reducing refractive-index mismatch and total internal reflection. *Bioinspired surface patterning* has been successfully used to increase the extraction efficiency of LEDs.

Top-emitting LEDs often make use of current-spreading layers (also referred to as window layers), which are transparent conductive semiconductor layers that spread the region of light emission beyond that surrounding the electrical contact. Current-blocking layers, which prevent current from entering the active region below the top contact, can also be used to control the light emission. The contact geometry can be designed to maximize light transmission.

A whole host of other techniques are also used to enhance the extraction efficiency. These include the use of distributed Bragg reflectors (see Sec. 7.1C) between the active layer and an absorbing substrate to reflect the light back toward the desired direction of emission, and reflective and transparent contacts. Another favored approach is the use of a transparent substrate in conjunction with **flip-chip packaging**, which allows the light to be extracted through the substrate rather than through the top surface of the device. The LED extraction efficiency can also be enhanced by guiding light to the surface of the device via a 2D photonic crystal (Sec. 7.3A), such as a regular array of 100–250-nm diameter holes formed in the current-spreading layer.

Spatial Pattern of Emitted Light

The far-field radiation pattern for light emitted into air from a planar surface-emitting LED is similar to that of a Lambertian radiator. The intensity varies as $\cos\theta$, where θ is the angle from the emission-plane normal; the intensity decreases to half its value at $\theta = 60°$. This pattern arises as a result of Snell's law: light rays bend away from the normal as they exit the semiconductor–air interface.

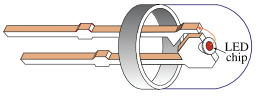


Figure 18.1-10 Polymer-encapsulated LED in a 5-mm-diameter dual in-line package (DIP). Encapsulation protects the LED chip (die), increases light extraction by reducing refractive-index mismatch, and serves as a lens to shape the beam.

LEDs are often encapsulated in transparent polymer lens domes such as epoxy or silicone for a number of reasons (Fig. 18.1-10). Lenses of different shapes alter the emission pattern in different ways, as illustrated schematically for hemispherical and parabolic lenses in Fig. 18.1-11. Polymer lenses can also enhance the extraction efficiency η_e. A lens with a refractive index close to that of the semiconductor optimizes the extraction of light from the semiconductor into the polymer. The shape of the lens can then be tailored so as to maximize the extraction of light at the polymer–air interface. Polymer materials usually have refractive indices that are intermediate between those of semiconductors and air and, in practice, yield an enhancement in light extraction by a factor of 2 to 3. Molded acrylic or polycarbonate collimators that make use of total internal reflection in conjunction with refraction are often used to provide parallel light rays for LED lighting applications, as illustrated in Fig. 1.2-14.

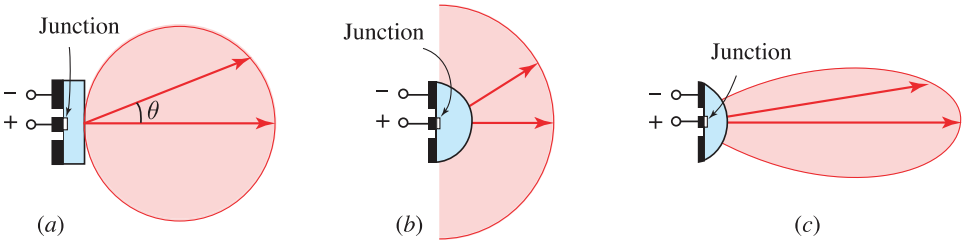


Figure 18.1-11 Radiation patterns of surface-emitting LEDs: (a) Lambertian spatial pattern in the absence of a lens; (b) spatial pattern with a hemispherical lens; (c) spatial pattern with a parabolic lens.

The radiation pattern from edge-emitting LEDs and laser diodes is usually quite narrow and can often be empirically described by the function $\cos^s\theta$, with $s > 1$. If $s = 10$, for example, the intensity decreases to half its value at $\theta \approx 21°$.

Output Photon Flux and External Quantum Efficiency

The **output photon flux** Φ_o (also called the **external photon flux**) is related to the internal photon flux by

$$\Phi_o = \eta_e \Phi = \eta_e \eta_i \frac{i}{e}, \tag{18.1-23}$$

where the internal efficiency η_i relates the internal photon flux to the injected electron flux, and the extraction efficiency η_e specifies how much of the internal photon flux is

transmitted out of the structure. A single quantum efficiency that accommodates both of these processes is the **external quantum efficiency (EQE)** η_{ex}:

$$\eta_{\text{ex}} \equiv \eta_e \eta_i. \tag{18.1-24}$$
External Efficiency

The output photon flux in (18.1-23) can therefore be written as

$$\Phi_o = \eta_{\text{ex}} \frac{i}{e}, \tag{18.1-25}$$
External Photon Flux

so that the external efficiency η_{ex} is simply the ratio of the external photon flux Φ_o to the injected electron flux i/e. Because the pumping rate generally varies locally within the junction region, so too does the generated photon flux. The LED output optical power P_o is directly related to the output photon flux since each photon has energy $h\nu$:

$$P_o = h\nu \Phi_o = \eta_{\text{ex}}\, h\nu \frac{i}{e}. \tag{18.1-26}$$
Output Power

The internal efficiency η_i for LEDs ranges between 50% and nearly 100%, while the extraction efficiency η_e for properly designed devices can extend up to 50%. The external efficiency η_{ex} of LEDs is thus typically below 50%.

As discussed in Sec. 16.2A, another measure of performance is the **wall-plug efficiency** (also called the **power-conversion efficiency** or **overall efficiency**), which is defined at the ratio of the emitted optical power P_o to the applied electrical power $P_{\text{e}} = iV$,

$$\eta_c \equiv \frac{P_o}{iV} = \eta_{\text{ex}} \frac{h\nu}{eV}, \tag{18.1-27}$$

where V is the voltage drop across the device. For $h\nu \approx eV$, as is the case for some commonly encountered LEDs, we obtain $\eta_c \approx \eta_{\text{ex}}$.

Resonant-cavity LEDs. The quantum efficiencies η_{ex} and η_c may be enhanced by making use of a **resonant-cavity light-emitting diode (RCLED)**. A pair of mirrors (e.g., distributed Bragg reflectors) is used to confine injection electroluminescence to a wavelength-sized, resonant microcavity in one dimension (Secs. 11.1B and 11.4). RCLEDs exhibit a number of attractive features: 1) the spontaneous-emission rate is enhanced by the Purcell effect (Sec. 14.3E), which results in an increase in the internal quantum efficiency η_i ; 2) the spectral width of the emitted light is reduced below kT when the cavity resonance is narrower than the spectral-intensity profile; 3) the temperature stability is then also enhanced because the cavity is less sensitive to temperature changes than is the semiconductor energy gap; and 4) the emission is more narrowly confined in angle, which results in an increase in the extraction efficiency η_e. As illustrated in Fig. 18.1-12, a substantial fraction of the light is emitted into a resonant mode whose angular extent falls principally within the extraction cone.

A photonic-crystal structure can also be incorporated in an RCLED to guide much of the residual light toward the surface of the device, thereby further increasing η_e. The increased values of η_i and η_e for RCLEDs lead directly to enhanced values of the external and wall-plug efficiencies $\eta_{\text{ex}} = \eta_e \eta_i$ and $\eta_c = \eta_{\text{ex}}(h\nu/eV)$, respectively.

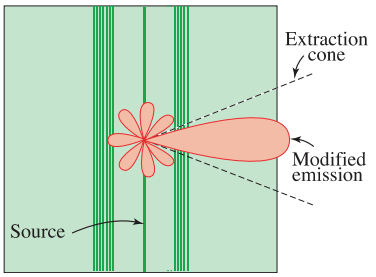


Figure 18.1-12 A plane-parallel-reflector resonant-cavity light-emitting diode (RCLED). Two closely spaced reflectors (the one at left with a reflectance near 100% and the one at right with a reflectance of, say, 50%) form a wavelength-size cavity in one dimension that confines the light and funnels a large portion of it into a spatial region that lies within the extraction cone.

However, RCLEDs are inherently low power devices by virtue of the small sizes of their active regions. The use of microresonators for enhancing the properties of photon sources is discussed further in Sec. 18.5.

Responsivity

The responsivity R of an LED is defined as the ratio of the emitted optical power P_o to the injected current i, i.e., $\mathrm{R} = P_o/i$. Using (18.1-26), we obtain

$$\mathrm{R} = \frac{P_o}{i} = \frac{h\nu\,\Phi_o}{i} = \eta_{\mathrm{ex}}\frac{h\nu}{e}\,. \tag{18.1-28}$$

The responsivity in W/A, when λ_o is expressed in μm, is then

$$\boxed{\mathrm{R} = \eta_{\mathrm{ex}}\frac{1.24}{\lambda_o}\,.} \tag{18.1-29}$$

LED Responsivity
(W/A; λ_o in μm)

For example, if $\lambda_o = 1.24\ \mu$m, then $\mathrm{R} = \eta_{\mathrm{ex}}$ W/A; if η_{ex} were unity, the maximum optical power that could be produced by an injection current of 1 mA would be 1 mW. Thus, for $\eta_{\mathrm{ex}} = {}^1/_2$ at $\lambda_o = 1.24\ \mu$m, we have $\mathrm{R} = {}^1/_2$ mW/mA.

In accordance with (18.1-26), the LED output power P_o is proportional to the injected current i. In practice, however, this relationship is valid only over a restricted range. For the particular device whose **light–current (L–i) curve** is shown in Fig. 18.1-13, the emitted optical power is proportional to the injection (drive) current only when the latter is less than about 20 mA. In this range, the responsivity has a constant value of about 0.3 mW/mA, as determined from the slope of the curve. For larger drive currents, saturation causes the proportionality to fail; the responsivity then declines with increasing drive current. Since $\lambda_o = 0.420\ \mu$m for this LED, (18.1-29) reveals that it has an external quantum efficiency (EQE) $\eta_{\mathrm{ex}} = 0.10$.

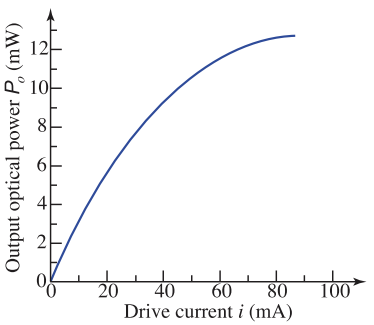


Figure 18.1-13 Optical power at the output of an LED versus injection (drive) current. This MQW InGaN/GaN LED emits in the violet region of the spectrum at $\lambda_o = 420$ nm; the device structure is exhibited in Fig. 18.1-21.

Spectral Distribution

The spectral intensity $r_{sp}(\nu)$ of light spontaneously emitted from a semiconductor in quasi-equilibrium has been determined, as a function of the concentration of injected carriers Δn, in Exercises 18.1-2 and 18.1-3. This theory is applicable to the electroluminescence light emitted from an LED in which quasi-equilibrium conditions are established by injecting current into a p–n junction.

Under conditions of weak pumping, such that the quasi-Fermi levels lie within the bandgap and are at least a few kT away from the band edges, the spectral intensity achieves its peak value at the frequency $\nu_p = (E_g + kT/2)/h$ (Exercise 18.1-3). In accordance with (18.1-15) and (18.1-16), the FWHM of the spectral intensity is $\Delta\nu \approx 1.8kT/h$, which is independent of ν, and $\Delta\nu = 10$ THz for $T = 300°$ K. When expressed in terms of wavelength, however, the width does depend on λ,

$$\Delta\lambda \approx 1.45\,\lambda_p^2\,kT, \tag{18.1-30}$$

Spectral Width (μm)

where kT is specified in eV, the wavelength is specified in μm, and $\lambda_p = c/\nu_p$.

The dependence of $\Delta\lambda$ on λ_p^2 is apparent in Fig. 18.1-14, which illustrates the observed wavelength spectral intensities for selected LEDs operating in the ultraviolet (indicated as magenta) and visible regions of the spectrum. AlN has the largest III–nitride energy bandgap, producing light at 210 nm; AlGaN is typically employed in the mid and near ultraviolet; InGaN is the material of choice in the violet, blue, and green; and AlInGaP usually serves the yellow, orange, and red. Typical spectral intensities for LEDs that operate in the near infrared are displayed in Fig. P18.1-5; these devices are generally fabricated from InGaAsP. The spectral width increases roughly as λ_p^2, in accordance with (18.1-30). If $\lambda_p = 1$ μm at $T = 300°$ K, for example, (18.1-30) provides $\Delta\lambda \approx 36$ nm. However, alloy broadening can result in a further increase in the spectral width, as is evident in the spectrum for the green LED in Fig. 18.1-14.

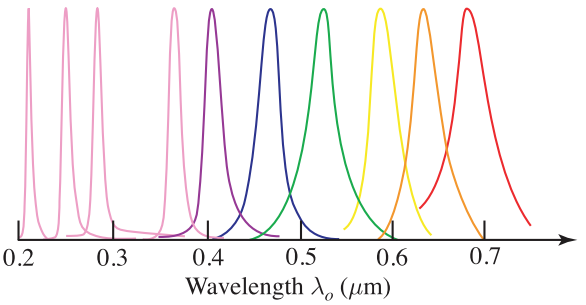


Figure 18.1-14 Spectral intensities versus wavelength for LEDs that operate in the ultraviolet and visible regions of the spectrum. The peak intensities are all normalized to the same value. Results for LEDs operating in the infrared are presented in Fig. P18.1-5.

Response Time

The response time of LEDs used for illumination is usually limited by the RC time constant of the device because the junction area, and therefore the capacitance, is large. The response time of communication-system LEDs, in contrast, is generally limited principally by the lifetime τ of the injected minority carriers that are responsible for radiative recombination. For a sufficiently small injection rate R, the injection/recombination process can be described by a first-order linear differential equation (Sec. 17.1D), and therefore by the response to sinusoidal signals. An experimental determination of the highest frequency at which an LED can be effectively modulated is easily obtained by measuring the output light power in response to sinusoidal electric

currents of different frequencies. If the injected current assumes the form $i = i_0 + i_1 \cos(\Omega t)$, where i_1 is sufficiently small so that the emitted optical power P varies linearly with the injected current, the emitted optical power behaves as $P = P_0 + P_1 \cos(\Omega t + \varphi)$.

The associated transfer function, which is defined as $H(\Omega) = (P_1/i_1)\exp(j\varphi)$, assumes the form

$$H(\Omega) = \frac{\mathrm{R}}{1 + j\Omega\tau}, \tag{18.1-31}$$

which is characteristic of a resistor–capacitor circuit. The rise time of the LED is τ (s) and its 3-dB bandwidth is $B = 1/2\pi\tau$ (Hz). A larger bandwidth B is therefore attained by decreasing the rise time τ, which comprises contributions from both the radiative lifetime τ_r and the nonradiative lifetime τ_nr through the relation $1/\tau = 1/\tau_\mathrm{r} + 1/\tau_\mathrm{nr}$. However, reducing τ_nr results in an undesirable reduction of the internal quantum efficiency $\eta_i = \tau/\tau_\mathrm{r}$. It may therefore be desirable to maximize the internal quantum efficiency–bandwidth product $\eta_i B = 1/2\pi\tau_\mathrm{r}$ rather than the bandwidth alone. This requires a reduction of only the radiative lifetime τ_r, without a reduction of τ_nr, which may be achieved by careful choice of the semiconductor material and doping level. Typical rise times of LEDs are in the range 1 to 50 ns, corresponding to bandwidths of hundreds of MHz.

Electronic Circuitry

An LED is usually driven by a current source, as illustrated schematically in Fig. 18.1-15(a), most simply implemented by means of a constant-voltage source in series with a resistor, as shown in Fig. 18.1-15(b). The emitted light is readily modulated by simply modulating the injected current. Analog and digital modulation are portrayed in Figs. 18.1-15(c) and 18.1-15(d), respectively. The performance of LED drivers may be improved by adding circuitry that regulates bias current, matches impedance, and provides nonlinear compensation to limit the maximum current. Fluctuations in the intensity of the emitted light may be stabilized by monitoring it with a photodetector, whose output is used as a feedback signal to control the injected current.

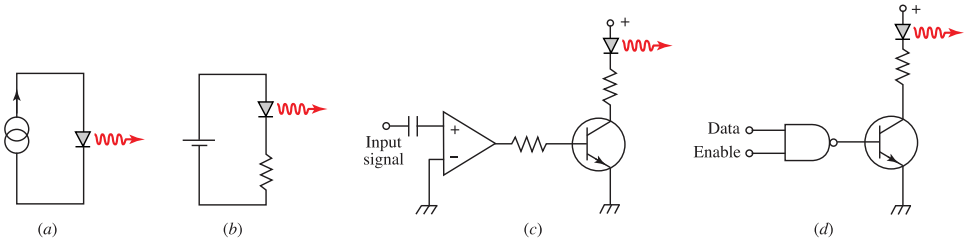


Figure 18.1-15 Various circuits can be used as LED drivers. These include (a) an ideal DC current source; (b) a DC current source provided by a constant-voltage source in series with a resistor; (c) transistor control of the current injected into the LED to provide analog modulation of the emitted light; and (d) transistor switching of the current injected into the LED to provide digital modulation of the emitted light.

C. Materials and Device Structures

Photonics was revolutionized in the 1950s by the growth of single-crystal binary III–V semiconductors, compounds that do not occur in nature. Many of these alloys have direct bandgaps and therefore exhibit large values of the internal quantum efficiency.

Photon sources fabricated from III–V materials also offer long lifetimes, unlike those that make use of II–VI alloys. GaAs was the first such material to be fabricated in the form of LEDs and LDs (see p. 787). Today's LED industry is almost exclusively built around ternary and quaternary III–V material systems, particularly InGaAsP, AlInGaP, and AlInGaN (Fig. 18.1-16)

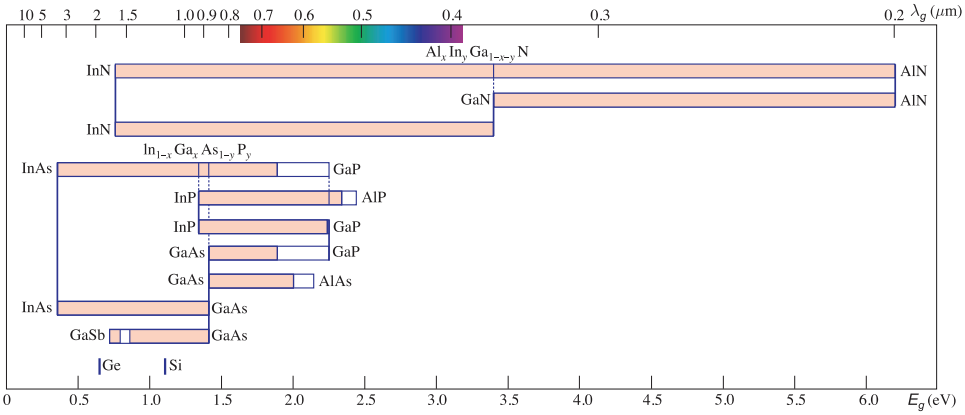


Figure 18.1-16 Bandgap wavelength λ_g, and corresponding bandgap energy E_g, for selected elemental and III–V binary, ternary, and quaternary semiconductors. Successive rows, starting at the top, represent AlInGaN, AlGaN, InGaN, InGaAsP, AlInGaP, InGaP, GaAsP, AlGaAs, InGaAs, and GaAsSb. The shaded regions indicate direct-bandgap compositions.

These III–V materials allow high-brightness light to be generated over a spectral range that stretches from the infrared to the ultraviolet, as exemplified by Figs. 18.1-14, 18.1-17, and P18.1-5.

Figure 18.1-17 LED traffic signal based on III–V materials.

LEDs may be constructed either in surface-emitting or edge-emitting geometries (Fig. 18.1-18). The surface-emitting LED emits light from a face of the device that is parallel to the plane of the active region. The edge-emitting LED, in contrast, emits light from the edge of the active region.

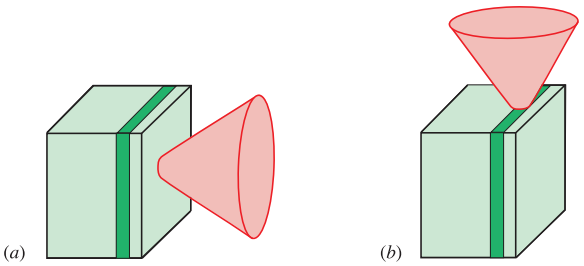


Figure 18.1-18 (*a*) Surface-emitting LED. (*b*) Edge-emitting LED.

We proceed to provide brief descriptions of the principal III–V semiconductor compounds used to fabricate LEDs, along with schematic illustrations of several representative device structures. *These compounds are also used to fabricate semiconductor optical amplifiers, laser diodes, quantum-confined lasers, microcavity lasers, and nanocavity lasers, as considered in Secs.* 18.2–18.6*, respectively.* Along the way, we highlight a number of applications for LEDs and LDs in the IR, visible, and UV.

GaAs. The first III–V material to play an important role in photonics was GaAs. This direct-bandgap, binary semiconductor was used to fabricate the first practical LED in 1961, with a peak emission wavelength at $0.890\ \mu$m (near the bandgap wavelength of GaAs at $\lambda_g = 0.873\ \mu$m). Not long thereafter, several other direct-bandgap, binary III–V semiconductors, grown by vapor-phase epitaxy (VPE) and liquid-phase epitaxy (LPE), were also shown to exhibit electroluminescence near their bandgap wavelengths (as provided in Table 17.1-2): GaSb ($\lambda_g = 1.70\ \mu$m), InP ($\lambda_g = 0.919\ \mu$m), InAs ($\lambda_g = 3.44\ \mu$m), and InSb ($\lambda_g = 7.29\ \mu$m).

GaAsP. Increasing the mole-fraction of phosphorus in the ternary semiconductor $\mathrm{GaAs_{1-x}P_x}$ causes the bandgap wavelength to move into the visible region of the spectrum, offering emission in the red (Fig. 18.1-16). Though the nature of the bandgap ultimately changes from direct to indirect as the bandgap wavelength decreases further, emission in the orange, yellow, and green can nevertheless be achieved by using nitrogen-doped versions of these materials (GaAsP:N and GaP:N). The nitrogen impurities (zinc and oxygen co-dopants can also be used) are incorporated into the material at sharply localized positions so that they can accommodate the substantial momentum changes associated with indirect transitions. However, the external quantum efficiencies of such LEDs are typically $< 1\%$, in part because of a lattice-constant mismatch with the GaAs substrate. Nevertheless, LEDs made of GaAs, GaAsP, GaAsP:N, and GaP:N are inexpensive to fabricate and continue to be used in low-brightness applications such as indicator lamps and remote controls for consumer appliances.

InGaAs. Adding indium to GaAs has the opposite effect of adding phosphorus; it serves to increase the bandgap wavelength, allowing it to extend all the way to the value for InAs. The ternary semiconductor $\mathrm{In_xGa_{1-x}As}$ is a direct-bandgap material that can be lattice matched to an InP substrate. Its bandgap is compositionally tunable over the near infrared and a portion of the mid infrared: $0.873\ \mu$m (GaAs) $\leq \lambda_g \leq 3.44\ \mu$m (InAs) (Fig. 18.1-16). LEDs fabricated from $\mathrm{In_xGa_{1-x}As}$ are used in consumer applications. Strained-layer InGaAs laser-diode arrays are used to pump $\mathrm{Yb^{3+}}$-doped DPSS and silica-fiber lasers at $\lambda_o = 940$ nm as well as $\mathrm{Er^{3+}}$: silica-fiber lasers and amplifiers at $\lambda_o = 980$ nm (Secs. 16.3A and 16.3B). InGaAs laser-diode arrays are also used for the in-band pumping of $\mathrm{Nd^{3+}}$:$\mathrm{YVO_4}$ DPSS lasers at 914 nm (Sec. 16.3A). In the domain of photodetectors, InGaAs is widely used in the fabrication of PIN detectors and avalanche photodiodes for use in optical fiber communication systems that operate in the 1.3–1.6-μm telecommunications band (Example 19.4-2 and Sec. 25.1D).

InGaAsP. The quaternary $\mathrm{In_{1-x}Ga_xAs_{1-y}P_y}$ is a versatile alloy that is widely used in the near-infrared region of the spectrum. Its bandgap is compositionally tunable over a substantial range of wavelengths [$0.549\ \mu$m (GaP) $\leq \lambda_g \leq 3.44\ \mu$m (InAs)], and lattice matching to an InP substrate can be maintained if the compositional mixing ratios x and y are judiciously chosen [stippled area in Fig. 17.1-7(a)]. Only a portion of this range enjoys the benefit of a direct bandgap, however (Fig. 18.1-16). InGaAsP laser-diode arrays are used for the in-band pumping of $\mathrm{Er^{3+}}$:silica-fiber amplifiers and lasers at 1480 nm (Secs. 15.3C and 16.3B).

Applications of IR LEDs and LDs. InGaAsP can be used to fabricate LEDs for short-haul, modest-bit-rate communication systems operating near $\lambda_o = 1330$ nm (Fig. 18.1-19). Long-haul, high-bit-rate communication systems generally operate in the vicinity of $\lambda_o = 1550$ nm and make use of laser diodes rather than LEDs since it is easier to couple the collimated light emitted by an LD into a single-mode fiber (Sec. 25.1B).

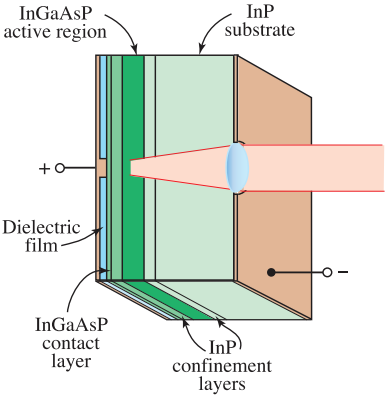


Figure 18.1-19 Saul–Lee–Burrus-type surface-emitting InGaAsP LED for use in an optical fiber communication system operating at a wavelength of 1.3 μm. The active region is lattice matched to the InP substrate. The device is mounted upside down in the package (flip-chip packaging) so the light emerges through the substrate. An integrated lens collimates the light for enhanced coupling to a fiber.

InGaAsSb. The bandgap wavelength may be further increased by replacing the P in InGaAsP with Sb, yielding the quaternary semiconductor $\mathrm{In}_{1-x}\mathrm{Ga}_x\mathrm{As}_{1-y}\mathrm{Sb}_y$. Using a GaSb substrate, this compositionally tunable alloy can be used to fabricate devices such as quantum-well lasers that operate in the mid-infrared region ($2 \leq \lambda_o \leq 4$ μm). However, III–antimonide sources have largely been replaced by mid-infrared quantum cascade lasers (Sec. 18.4D), which generally offer superior performance.

AlGaAs. Just as adding phosphorus to GaAs increases its bandgap energy, so too does the addition of aluminum. Like $\mathrm{GaAs}_{1-x}\mathrm{P}_x$, the ternary alloy $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ can be compositionally tuned in the red and near infrared (AlGaAs has a direct bandgap in the wavelength range $630 \leq \lambda_g \leq 900$ nm, as is evident from Fig. 18.1-16). Unlike GaAsP, however, AlGaAs has the merit that lattice matching to GaAs is maintained for all mole fractions of aluminum [Fig. 17.1-7(a)] so that the material can serve as a high-brightness source in the red. However, since $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}/\mathrm{GaAs}$ multiquantum-well structures can be adversely affected by nonuniform carrier distributions in the active region, LEDs are often fabricated using a double-heterostructure configuration of the form $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}/\mathrm{Al}_y\mathrm{Ga}_{1-y}\mathrm{As}$, in which the compositions of the barriers and wells differ. Collections of AlGaAs laser diodes emitting at 808 or 880 nm can be arrayed in the form of bars or stacks to provide four-level and in-band pumping, respectively, for Nd^{3+}:YVO_4 and Nd^{3+}:YAG solid-state lasers (Sec. 16.3A). Similarly, AlGaAs laser stacks emitting at 888 nm provide in-band pumping for the enormously powerful Nd^{3+}:glass laser amplifiers used in the HAPLS laser system (Example 23.2-3). AlGaAs laser-diode arrays emitting at 793 nm are also used to pump Tm^{3+}:silica-fiber lasers (Sec. 16.3B).

AlInGaP. The quaternary semiconductor $(\mathrm{Al}_x\mathrm{Ga}_{1-x})_y\mathrm{In}_{1-y}\mathrm{P}$ is a direct-bandgap material over a substantial range of the near infrared and the longer reaches of the visible spectrum (Fig. 18.1-16). Lattice matching to GaAs is attained for compositions in the range $(\mathrm{Al}_x\mathrm{Ga}_{1-x})_{0.5}\mathrm{In}_{0.5}\mathrm{P}$. The quantum efficiency of AlInGaP LEDs is enhanced by making use of multiquantum-well (MQW) active regions, wafer-bonded transparent GaP substrates in place of GaAs, and resonant-cavity (RC) configurations that offer decreased bandwidth and directed emission patterns.

Applications of visible LEDs. Under daylight conditions, human vision is maximally sensitive at 555 nm in the yellow-green region of the spectrum (Sec. 18.1F). This makes AlInGaP the material of choice for high-brightness applications such as traffic lights and signage, at least in the red, orange, yellow-orange (amber), and yellow regions. AlInGaP/InGaP LEDs also find occasional use in plastic-fiber communication systems that operate in the 650-nm wavelength region (Fig. 18.1-20). The lattice-matched ternary compound $In_{0.5}Ga_{0.5}P$, with a bandgap wavelength of 650 nm, is widely used for red laser pointers.

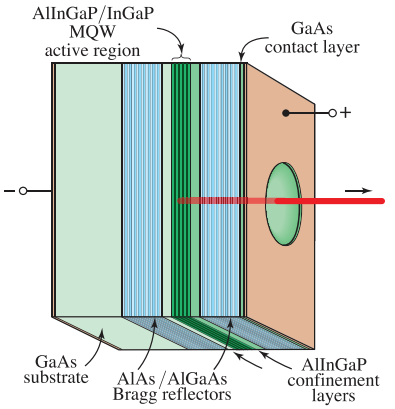


Figure 18.1-20 Surface-emitting AlInGaP/InGaP 650-nm MQW RCLED for use in short-haul, plastic-fiber communication systems. A top-emitting structure is used because of the opacity of the GaAs substrate in this device. The distributed Bragg reflectors comprise AlAs/AlGaAs layers with an aluminum content that is sufficiently high so that the 650-nm light is transmitted. A lens enhances coupling of the light to a fiber.

GaN. Gallium nitride is a direct-bandgap, binary semiconductor with a bandgap wavelength $\lambda_g = 0.366$ μm that falls in the near-ultraviolet region of the spectrum. It may be grown by MBE, MOCVD, or HVPE. GaN is the progenitor of the important ternary and quaternary compounds InGaN, AlGaN, and AlInGaN, much as GaAs was the progenitor of InGaAs, AlGaAs, and InGaAsP. These materials are often grown on sapphire or Si substrates, despite substantial lattice mismatch; unlike the arsenide and phosphide III–V compounds, the III–nitrides can tolerate large dislocation concentrations so that lattice mismatch is well-tolerated. Buffer layers can also be used to accommodate differences in thermal-expansion coefficients. A better lattice match is offered by SiC, which is sometimes used as a substrate.

InGaN. The ternary semiconductor $In_xGa_{1-x}N$ is a direct-bandgap material with a bandgap wavelength that spans the region 366 nm (GaN) $\leq \lambda_g \leq$ 1.61 μm (InN). InGaN is the material of choice for high-brightness LEDs in the wavelength range $366 \leq \lambda_g \leq 580$ nm, comprising the near-ultraviolet, violet, blue, and green regions of the spectrum (Fig. 18.1-16). This III–nitride alloy is thus complementary to AlInGaP, which best accommodates the red, orange, and yellow regions. As with AlInGaP, the quantum efficiency is enhanced by making use of MQW structures such as GaN/InGaN (Fig. 18.1-21). The substrate is often GaN on sapphire. However, the number of quantum wells is generally limited because of population limits imposed by the hole diffusion length; low and/or thin barriers are preferred. Performance can also be enhanced by the use of resonant-cavity devices. Other configurations include arrays of quantum dots that self-assemble on growth and arrays of nanorods.

AlGaN. $Al_xGa_{1-x}N$ is also a ternary III–nitride direct-bandgap semiconductor, but its bandgap wavelength falls in the range 206 nm (AlN) $\leq \lambda_g \leq$ 366 nm (GaN) (Fig. 18.1-16), which covers the mid- and near-ultraviolet regions ($200 \leq \lambda_o \leq 390$ nm). LEDs comprising AlGaN/AlGaN heterostructures have been fabricated across this wavelength region although achieving high efficiency is more challenging at

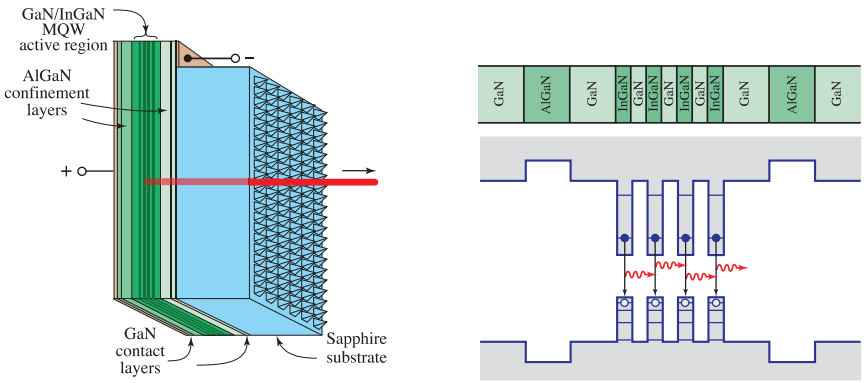


Figure 18.1-21 Flip-chip packaged, surface-emitting GaN/InGaN MQW LED operating at $\lambda_o = 420$ nm in the violet spectral region. The light is extracted through the GaN-on-sapphire transparent substrate, which is textured with an array of tiny pyramids to increase the extraction efficiency. In the structure illustrated, the active region comprises 5-nm GaN barriers and 2.5-nm $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}$ wells.

these shorter wavelengths. As with InGaN, LED quantum efficiency is enhanced by making use of double-heterostructure or MQW active regions with layers of the form $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{N}/\mathrm{Al}_y\mathrm{Ga}_{1-y}\mathrm{N}$. High-quality AlN, or templates of AlGaN/AlN/sapphire, serve as transparent lattice-matched substrates for ultraviolet AlGaN-based emitters. AlN LEDs emitting at 210 nm have also been fabricated. By virtue of their large bandgap and large thermal conductivity, III–nitride materials are also employed in specialized electronic components such as transistors for high-power and high-temperature applications.

Applications of UV LEDs. Ultraviolet LEDs operating in the UVA band ($315 \leq \lambda_o \leq 400$ nm) find use in applications such as printing, curing, and counterfeit detection. Source emitting in the UVC band ($100 \leq \lambda_o \leq 280$ nm) are useful for sterilization as well as germicidal and water-treatment applications. UV LEDs are also employed for detecting chemical and biological agents (many of which fluoresce at particular wavelengths when exposed to ultraviolet light) and for non-line-of-sight covert communications.

AlInGaN. It is clear from the foregoing that the ternary III–nitride compounds InGaN and AlGaN are suitable for fabricating sources that stretch across the visible and ultraviolet regions of the spectrum. However, the quaternary semiconductor $(\mathrm{Al}_x\mathrm{In}_y\mathrm{Ga}_{1-x-y})\mathrm{N}$ has the merit that it can be lattice matched to a GaN template for certain values of x and y [Fig. 17.1-7(b)], thereby increasing the quantum efficiency of the resulting devices. This lattice matching is analogous to that of AlInGaP to GaAs and of InGaAsP to InP. AlInGaN LEDs with lattice matching to a GaN substrate are useful over wavelengths ranging from 366 nm, the bandgap wavelength of GaN, to $\lambda_o \approx 250$ nm, the wavelength of AlInN that is lattice matched to GaN. AlInGaN/InGaN/AlInGaN quantum-well structures serve as active-region materials for devices. AlInGaN can also serve as a transparent contact layer.

D. Silicon Photonics

Silicon has long been the leading materials platform for **integrated electronics**, for a whole host of reasons: it is 1) abundant and inexpensive; 2) readily grown in pure form and in bulk; 3) easy to dope, oxidize, and manipulate; 4) stable at high temperatures; and 5) compatible with CMOS technology. Its ubiquity, availability, and properties have also made it an attractive platform for **integrated photonics**. The high refractive-index contrast of silicon and its oxides allows strong optical confinement in

a compact volume. This, together with its transparency in the 1.3–1.6-μm telecommunications band, and its CMOS compatibility, promotes its use for devices such as high-performance optical waveguides, filters, splitters, multiplexers, demultiplexers, and wavelength converters, as well as other components (Sec. 25.1E).

A notable exception to the adaptability of Si centers on its use as an active medium for LEDs and laser diodes. The development of Si-based light sources has been hampered by its indirect bandgap, which restricts its ability to generate light efficiently via interband transitions (Fig. 17.2-7). Over the years, extensive efforts have been devoted to surmounting this roadblock, either by mitigating the indirect nature of silicon's bandgap or by avoiding it altogether. Early efforts directed toward increasing the efficiency of light emission involved the use of alternatives to its crystalline form, such as porous silicon (in which nanopores pervade the diamond structure); silicon nanocrystals, superlattices, and quantum dots (Example 17.2-2); and Er^{3+}-doped silicon-based hosts and superlattices. None of these approaches has been particularly successful, however. A more fruitful approach was to co-opt light-emitting interactions in silicon other than those associated with interband transitions. In particular, the silicon Raman laser relies on stimulated Raman scattering and is thus indifferent to the nature of the bandgap (Example 16.3-4). Still, Raman devices require optical rather than electrical pumping, which reduces their appeal for many applications. Yet, silicon Raman lasers have been successfully integrated with direct-bandgap emitters such as InP that serve as an optical pump.

Fortunately, substantial progress has been made in recent years in implementing silicon-based on-chip light sources for use in **photonic integrated circuits (PICs)**. Three approaches are currently in use, each with its own limitations and merits:

1. *Flip-chip integration* (*direct-mounting integration*) of III–V laser diodes into a separately fabricated silicon platform, often with optical butt coupling. This approach, which makes use of solder bumps, requires sub-micrometer-scale alignment precision and is not scalable to large wafer volumes or complex laser designs, but it is straightforward.
2. *Heterogeneous integration* (*hybrid approach*) of III–V lasers into prepatterned silicon circuits, typically via wafer bonding and with optical evanescent coupling to evade lattice-matching limitations. This approach is incompatible with the clean CMOS-foundry environment. However, it accommodates a whole host of materials and can also relegate photon storage to the undoped silicon platform (with its low loss and high Q) via hybrid modes, thereby facilitating the fabrication of narrow-linewidth, dense-comb, and mode-locked lasers.
3. *Direct heteroepitaxial growth* of III–V lasers on Si substrates using intermediate buffer layers to minimize dislocations in the light-emitting region. This approach is encumbered with the large lattice-constant and thermal mismatches between Si and III–V materials, which result in dislocations that reduce efficiency by acting as nonradiative recombination centers. However, this can be largely counterbalanced by employing quantum-dot, rather than quantum-well emitters, since: 1) quantum dots are less affected by the threading dislocations initiated by lattice and thermal mismatches, and 2) quantum dots enjoy substantially reduced sensitivity to temperature changes.

On balance, direct heteroepitaxy appears to be the most attractive alternative for large-scale, low-cost, fabrication of silicon-based on-chip light sources.

It is worth noting that group-IV photonics also offers a route to the development of on-chip light sources via combinations and alloys of indirect-bandgap semiconductors such as Si, Ge, Sn, and C. Germanium-based structures are leading the way, although considerable challenges remain (see, e.g., Example 17.2-1 and Sec. 18.5B). Interestingly, the use of such materials is not new: the first LED, which dates to 1907, was a forward-biased SiC Schottky diode.

E. Organic LEDs

OLEDs

Organic light-emitting diodes can be fabricated from small organic molecules or conjugated polymer chains (Sec. 17.1B). Small-molecule **organic light-emitting diodes**, called **OLEDs** or **SMOLEDs**, are efficient generators of electroluminescence in the blue, green, and red. A device is formed from two thin ($\approx$ 100-nm) organic semiconductor films that are juxtaposed to form an organic heterostructure. As shown in Fig. 18.1-22(a), this structure is sandwiched between two inorganic electrodes, an anode that injects holes and one or more cathodes that inject electrons. This contrasts with the process of carrier injection in inorganic LEDs, which makes use of heavily doped p- and n-type crystalline materials and strong forward bias.

The injected carriers are transported to the heterojunction (active region), forming bound excitons that generate spontaneous emission upon recombination. Different heterostructure materials yield different recombination-radiation wavelengths, so several heterostructures can be patterned on a single substrate to provide a multicolor OLED. Such heterostructures can be fabricated in a side-by-side, or *striped configuration*, forming a color-tunable *horizontal stack* as shown in Fig. 18.1-22(a). Alternatively, they can be fabricated one atop the other, creating a *vertical stack* with a blue emitter on top, a green emitter in the middle, and a red emitter on bottom, as shown in Fig. 18.1-22(b). White organic light-emitting diodes (**WOLEDs**), which are the elements of white OLED light panels, are usually fabricated using vertical stacks.

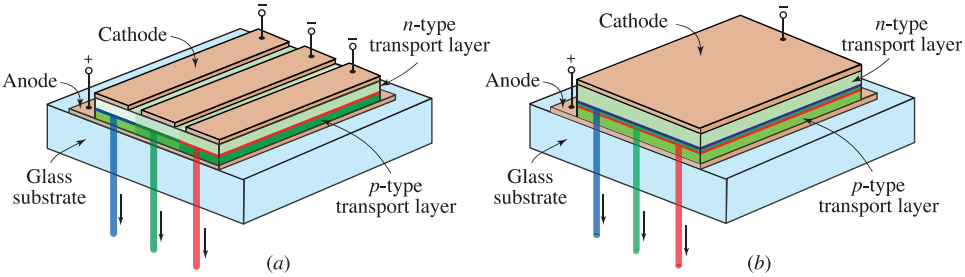


Figure 18.1-22 OLED structures fabricated in the form of (a) a horizontal stack of blue, green, and red emitters, which requires patterning but is color tunable, and (b) a vertical stack. Calcium and indium tin oxide are commonly used as the cathode and transparent anode materials, respectively. Exciton recombination radiation emitted at the organic heterojunctions exits through the transparent anode and glass substrate. Organic semiconductors used to fabricate OLEDs include hole-transporting TPD (triphenyl diamine derivative) and electron-transporting Alq_3 [aluminum tris(8-hydroxyquinoline)]. Luminescent dopants can be infused into the active regions to enhance the internal quantum efficiency and to create white light. A white element is sometimes added to the horizontal stack to facilitate color refinement.

The energy levels of bound excitons in organic materials resemble those of electrons in dye molecules, comprising both singlet (S) and triplet (T) states, as schematized in Fig. 14.1-8. The electron spin in a singlet state is antiparallel to that of the remainder of the molecule, resulting in a total spin angular-momentum quantum number $\mathcal{S} = 0$ and spin multiplicity $2\mathcal{S} + 1 = 1$, as explained in Secs. 14.1A and 14.1C. The electron spin in a triplet state, in contrast, is parallel to that of the remainder of the molecule, which results in $\mathcal{S} = 1$ and $2\mathcal{S} + 1 = 3$. The spin multiplicity of the triplet state is thus three times that of the singlet state, whence its appellation.

As explained in Sec. 14.5A, radiative transitions that take place between two states of the same multiplicity (S $\rightarrow$ S or T $\rightarrow$ T) are spin-allowed, in which case the

luminescence process is known as **fluorescence**. Luminescence from spin-forbidden transitions (e.g., T $\rightarrow$ S), in contrast, is called **phosphorescence**. The lifetimes of phosphorescent transitions are usually far longer than those of fluorescent transitions (e.g., msec vs. nsec) because of the forbidden nature of the former. The ground states of most organic compounds are singlet states so that the radiative decay of singlet excitons is strongly favored.

However, triplet-exciton radiative recombination can be fostered by infusing the active region of the device with fluorophores that bind to the organic molecules or conjugated polymer chains comprising the heterostructure. Triplet excitons can then efficiently transfer their energy to the fluorophore while concomitantly transferring their spin angular momentum to the organic molecule or polymer to which the exciton is bound. By virtue of the triplet-state multiplicity, this serves to increase the internal quantum efficiency of the device by a factor of four. This approach also allows the colors of the emitted light to be determined by the choice of fluorophores rather than by the exciting excitons.

PLEDs

Polymer light-emitting diodes, called **PLEDs** or **P-OLEDs**, are similar in construction to OLEDs except that they typically have an n-type active region into which holes are injected by a p-type organic layer. Light-emitting polymer materials (**LEPs**) that comprise such devices are often derivatives of poly(p-phenylene vinylene) (PPV) and polyfluorene. The color of the emitted light can be modified by substituting appropriate side chains on the polymer backbone. PLEDs are less expensive to fabricate than OLEDs and are readily printed, but they generally have lower efficiencies and shorter life spans.

The desirable features of small-molecule and large-molecule polymeric organic materials can be brought together in molecules known as *phosphorescent dendrimers*. These are large molecular balls containing a heavy-metal ion core, such as Ir(2-phenylpyridine)$_3$, which facilitates triplet-exciton radiative recombination via spin–orbit coupling; layers of branching-ring structures are bonded around it. Alternatively, high efficiencies can be attained by making use of efficient **thermally activated delayed fluorescence (TADF)** emitters, which have an energy gap between their singlet and triplet excited states (S_1 and T_1, respectively) that is sufficiently small so that temperature fluctuations can drive transitions to the singlet state.

F. LED Lighting

Only about 5% of the optical power radiated by a typical incandescent lamp is emitted in the form of visible light; the remaining 95% is emitted in the infrared as heat. Light-emitting diodes, in contrast to incandescent and fluorescent sources, are efficient and versatile, and have long operational lives. LED and OLED sources, discussed in Secs. 18.1C and 18.1E, respectively, are widely used in residential, architectural, automotive, and street lighting. The light emitted by LEDs can be dynamically controlled and can assume an enormous palette of colors, including white, with excellent color rendering quality. **LED lighting** is also called **solid-state lighting**.

The human visual system is constructed in such a way that the combination of light from a small number of judiciously chosen LEDs, in spite of their narrow individual spectral profiles, can nonetheless appear white to the observer. Such light, termed **metameric white light**, can even be generated by a single LED when it is endowed with a photoluminescent phosphor. As a prelude to considering the features of LED lighting, it is useful to set forth some of the units and metrics used in the field. This is followed by brief discussions of single-die (discrete) LEDs, white LEDs, array LEDs, chip-on-board (COB) LEDs, retrofit LED lamps, and white OLED light panels.

Units and Metrics

In earlier chapters our attention was focused exclusively on **radiometric units**, which characterize the strength of a light source in terms of its physical properties. **Photometric units**, on the other hand, characterize the effectiveness of a light source in terms of its ability to excite the human visual system. Both radiometric and photometric units are important in LED lighting. Examples of radiometric units are the **radiant flux** P, also called the optical power (specified in W) and the **irradiance** I, commonly called the intensity (specified in W/m^2). The corresponding photometric units are, respectively, the **luminous flux** $P_{\rm v}$ (specified in lumens, which is abbreviated lm) and the **illuminance** $M_{\rm v}$ (specified in lm/m^2, which is the same as lux). Photometric quantities often carry the subscript "v" to indicate their connection with vision.

The **luminous flux** $P_{\rm v}$ is a measure of the brightness of a light source as perceived by the eye. Dictated by historical considerations, the lumen is defined such that 683 lm corresponds to 1 W of optical power at 555 nm, the wavelength of maximum human visual sensitivity under daylight conditions where photopic vision prevails. The **photopic luminosity function** $V(\lambda_o)$, which has a value of unity at $\lambda_o = 555$ nm, specifies the relative sensitivity of the eye over the range of visible wavelengths, which stretches from 380 to 780 nm.

The **wall-plug luminous efficacy** $\eta_{\rm v}$, also called the **overall luminous efficacy** or the **luminous efficacy of the source**, is the most widely used efficiency metric for LED lighting devices. It is defined as the ratio of the luminous flux generated (lm) to the electrical power provided to the device (W), and hence has units of lm/W:

$$\eta_{\rm v} \equiv P_{\rm v}/iV. \tag{18.1-32}$$

The electrical power iV is the product of the applied current and voltage. An LED that optimally converts electrical power to visible light would exhibit a luminous efficacy of 683 lm/W, a value that can be attained only if the device has a power-conversion efficiency $\eta_c = 1$ [see (18.1-27)] and if it emits monochromatic yellow-green light at a wavelength of 555 nm. Since metameric white light comprises a wavelength spectrum broader than just a single component at 555 nm, and since the relative sensitivity of the eye diminishes for wavelengths both below and above this value in accordance with the photopic luminosity function, it is clear that a white LED is constrained to provide $\eta_{\rm v} < 683$ lm/W. Observed values of luminous efficacy for three distinct lighting devices are provided for context: 1) a 100-W tungsten incandescent lamp offers $\eta_{\rm v} \approx 15$ lm/W; 2) an equivalent compact fluorescent lamp provides $\eta_{\rm v} \approx 70$ lm/W; and 3) an equivalent commercially available white LED lamp yields $\eta_{\rm v} \approx 120$ lm/W, although white LED lamps with $\eta_{\rm v} > 300$ lm/W have been reported.

The **correlated color temperature (CCT)** of a source is the temperature of a blackbody radiator, as described in Sec. 14.4B, whose color most closely resembles that of the source. As the temperature of a blackbody radiator transitions from low to high, its color goes from deep red, to orange, to yellow, to yellowish-white, to white, and ultimately to bluish-white at sufficiently high temperatures. Sources in the range $2700^\circ \lesssim \text{CCT} \lesssim 3500^\circ$ K (yellowish) are called "warm white," while those in the range $5000^\circ \lesssim \text{CCT} \lesssim 7500^\circ$ K (bluish-white) are referred to as "cool white." The value of the CCT ascribed to a source of light bears no connection to its thermodynamic temperature.

Another important measure is the **color rendering index (CRI)**, which indicates how realistically a source can render colors. This metric, which is defined only for sources that are approximately white, is calculated by measuring the light reflected from a standardized sample set for various colors. The CRI assumes values between 0 and 100, with 100 considered ideal.

Salutary Features

LED lighting offers many salutary features in comparison with its incandescent and fluorescent counterparts:

- *Long operational life, slow failure, and low cost.* LEDs have life spans that can exceed 100 000 hours, far longer than the 1 500 hours for typical incandescent sources and 10 000 hours for compact fluorescent lamps; their failure is also gradual rather than sudden. These features result in reduced long-term replacement and maintenance costs.
- *Low energy consumption.* High values of the wall-plug luminous efficacy η_v indicate that LEDs use far less electrical power than their incandescent and fluorescent cousins to generate a given luminous flux. Because of this, LEDs can be powered by solar panels.
- *Broad choice of colors and high-quality color rendering.* LEDs can generate light with colors that span the gamut of human vision, including a continuum of whites. They offer high values of the color rendering index (CRI), indicating that colored objects appear natural under illumination.
- *Dynamic and smart-networking capabilities.* The colors, temporal irradiance patterns, and spatial distributions of light produced by LEDs can be dynamically programmed. The electronic drivers can also communicate wirelessly with each other and with collections of sensors to provide smart networks.

Discrete LEDs

Individually, LEDs emit narrowband, colored light — the color is determined by the bandgap wavelength of the material from which the LED is fabricated, as exemplified in Fig. 18.1-14. The optical and electrical characteristics of some typical individual, single-die LEDs of different colors are set forth in Table 18.1-1. Such devices are commonly called **discrete LEDs**.

Table 18.1-1 Representative parameter values for 3-mm-diameter, discrete LEDs (blue, green, and red): peak wavelength λ_p (nm), forward voltage V (V), forward current i (A), electrical power consumed iV (W), external quantum efficiency η_{ex}, power-conversion efficiency η_c, output radiant flux (optical power) P_o (W), luminous flux P_v (lm), and wall-plug luminous efficacy η_v (lm/W). The quantities η_{ex}, η_c, and P_o are interrelated via (18.1-24), (18.1-26), and (18.1-27).

Color	λ_p	V	i^a	iV	η_{ex}	η_c	P_o	P_v	η_v
Blue	465	3.1	0.35	1.1	0.4	0.35	0.38	50	45
Green	528	3.2	0.35	1.1	0.3	0.22	0.25	125	115
Red	625	2.2	0.35	0.8	0.4	0.36	0.28	75	95

[a] Tripling the current to ≈ 1 A results in an approximate doubling of the radiant flux P_o and luminous flux P_v, but this comes at the expense of a reduction in the luminous efficacy η_v.

White LEDs

White is by far the most important color for illumination. There are several methods by means of which colored LED light can be converted into metameric white light. The *first method*, widely used because of its simplicity and low cost, makes use of a *phosphor-conversion LED*, which is an LED die coupled with one or more phosphors that generate photoluminescence (Sec. 14.5B). In its simplest implementation, a violet LED and a yellow phosphor give rise to metameric white light (any two colors whose combination results in white light are known as **complementary colors**). A red phosphor can be added to the combination to yield warmer metameric white light and

quantum dots can be used in place of phosphors to increase efficiency. The *second method*, known as *additive color mixing*, relies on superposing the light generated by several LEDs of different colors — this approach has the merit of offering color-tunable LED lighting. The *third method*, often referred to as the *hybrid approach*, makes use of two or more LEDs of different colors (e.g., blue and red) in conjunction with one or more phosphors. This method can offer favorable color-quality attributes at the expense of increased complexity and cost. We proceed to examine the first and second methods in turn.

Method 1: Phosphor-conversion LEDs (PC-LEDs). The evolution of the PC-LED since the year 2000 is illustrated in Fig. 18.1-23. Early devices, such as that portrayed in Fig. 18.1-23(a), made use of InGaN LED chips with a central emission wavelength $\approx$ 465 nm and a FWHM spectral width $\approx$ 35 nm. Some of the blue LED light that impinged on the Ce^{3+}:YAG phosphor generated yellow photoluminescence light with a spectral bandwidth $\approx$ 500–700 nm that was relatively broad. The result was metameric white light. A contemporary PC-LED, such as that shown in Fig. 18.1-23(b), operates on the same principle, but takes the form of a **surface-mounted device (SMD)**, with its electrical contacts lateral to the housing; this offers greatly improved heat-sinking and efficiency as well as reduced size. The LED die is supported by a ceramic base and is overlaid with a thin yellow phosphor sheet. The entire device is encapsulated in a hemispherical silicone lens. The device illustrated also makes use of an InGaN LED chip, but with a shorter center wavelength ($\approx$ 445 nm in the violet) and a narrower FWHM spectral width ($\approx$ 10 nm). The yellow photoluminescence has a central wavelength of $\approx$ 570 nm and a spectral band of $\approx$ 510–630 nm, corresponding to a FWHM of $\approx$ 120 nm, somewhat narrower than that of the first-generation device. Violet and yellow are complementary colors so metameric cool-white light results.

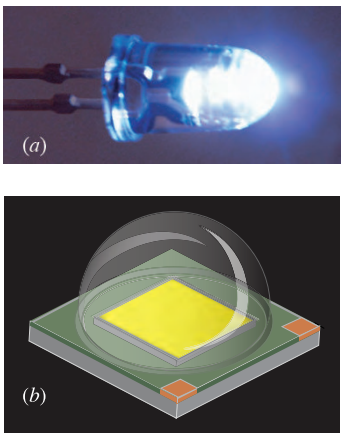


Figure 18.1-23 Evolution of the phosphor-conversion white LED. (a) White-light emission from an early device (ca. 2000) containing an InGaN LED die and a yellow phosphor in a 5-mm-diameter dual in-line package (DIP). This device generated metameric cool-white light with a wall-plug luminous efficacy $\eta_v \approx 20$ lm/W. (b) A contemporary device comprising an InGaN LED die overlaid with a thin yellow phosphor sheet and a 3-mm-diameter hemispherical lens in a surface-mounted-technology package. Devices such as these provide metameric cool-white light with $P_v > 500$ lm and $\eta_v > 300$ lm/W, a factor of 15 larger than that of early devices such as that portrayed in (a).

Method 2: Additive color mixing. The second method of obtaining metameric white light relies on multiple dies that generate different colors. Appropriate combinations of red, green, and blue light are perceived as white, a phenomenon illustrated in Fig. 18.1-24. Additive color mixing is used in luminaires to provide light of tunable color. In the terminology of LED lighting, a luminaire is a light fixture containing one or more LED lamps along with optics that shape and guide the emitted light to the exterior.

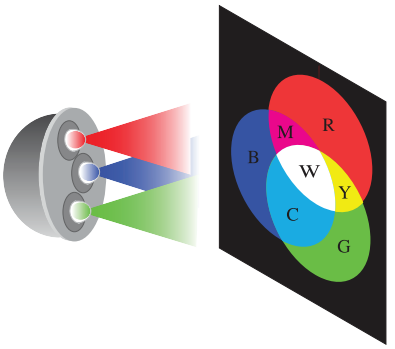


Figure 18.1-24 Additive color mixing. A device that generates light of tunable color often contains LEDs that emit Red (R), Blue (B), and Green (G), as shown. When observed, the overlapping light beams yield the following colors, as portrayed in the figure:

$$\begin{aligned} \text{B} + \text{G} &\rightarrow \text{C} \ \text{(Cyan)} \\ \text{R} + \text{B} &\rightarrow \text{M} \ \text{(Magenta)} \\ \text{R} + \text{G} &\rightarrow \text{Y} \ \text{(Yellow)} \\ \text{R} + \text{B} + \text{G} &\rightarrow \text{W} \ \text{(White)} \end{aligned}$$

Array LEDs

Modern color-mixing LEDs, which contain red, green, and blue dies within a single LED package, have the merit that they can be electrically tuned to emit essentially any color within the gamut of human vision. An example is provided by the array LED illustrated in Fig. 18.1-25, which contains red, green, and blue dies, together with a white phosphor-conversion emitter, all in close proximity and individually addressable. A 5-mm-diameter hemispherical silicone lens caps this multicolor array LED.

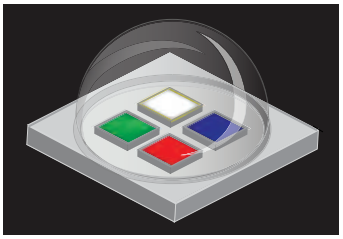


Figure 18.1-25 A color-mixing LED comprising individually addressable red, green, and blue dies, along with a white phosphor-conversion emitter, in an SMD package capped by a 5-mm-diameter hemispherical lens. Array LEDs such as these can be electrically tuned to emit essentially any color in the gamut of human vision, including metameric white light.

COB LEDs

A **chip-on-board (COB) LED**, sometimes called an **LED integrated array**, offers a modular alternative to a collection of discrete LEDs. As displayed in Fig. 18.1-26, high luminous flux is attained by making use of a large number of dies on a chip (often tens but sometimes hundreds), with high packing density, configured in the form of a single circuit and mounted on a printed-circuit board or other substrate. The device serves as a uniform diffuse source of light and is suitable for many single-color lighting applications, directional and non-directional alike. Chip-on-board LEDs are available with a broad range of parameters, including size, die density, operating voltage, color, light output, and efficiency.

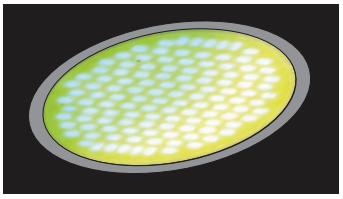


Figure 18.1-26 An illuminated chip-on-board (COB) device containing 120 InGaN dies embedded in a 3-cm-diameter, 1.5-mm-thick layer of yellow phosphor. Drawing 125 W of electrical power, a device such as this provides metameric cool-white light with a luminous flux $\approx 18\,000$ lm and a wall-plug luminous efficacy ≈ 145 lm/W.

LED Retrofit Lamps

An LED lamp designed as a drop-in replacement for an incandescent lamp is known as an **LED retrofit lamp**. These devices, generally called *bulbs*, typically comprise single- and multiple-die LEDs incorporated into plastic housings. The operating characteristics of the lamp, usually indicated on the packaging, include the luminous flux (lm), wall-plug luminous efficacy (lm/W), correlated color temperature (° K), color rendering index (CRI), electrical power consumption (W), as well as the electrical power consumption of an incandescent lamp with the same luminous flux.

A white LED retrofit lamp, the top portion of which is illustrated in the cutaway view depicted in Fig. 18.1-27, closely resembles its incandescent counterpart, at least from a morphological perspective, and it has roughly the same weight. It incorporates a collection of surface-mounted devices, like the one depicted in Fig. 18.1-23(*b*), or chip-on-board devices such as that illustrated in Fig. 18.1-26.

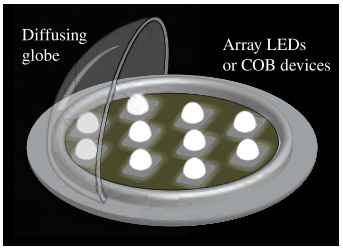


Figure 18.1-27 A white LED retrofit lamp contains an array of LEDs or chip-on-board (COB) devices enclosed in a diffusing globe. The cooling means, drive and dimmer electronics, and screw base are not shown. A device such as that depicted here might consume 10 W of electrical power and generate light indistinguishable from that produced by an incandescent bulb consuming 100 W of electrical power. An omnidirectional and dimmable lamp comprising 10 LEDs, each generating a luminous flux of 150 lm, produces metameric warm-white light with an overall luminous flux $P_v = 1500$ lm and a wall-plug luminous efficacy $\eta_v = 150$ lm/W. The emitted light often has a correlated color temperature (CCT) in the vicinity of 2700° K and a color rendering index (CRI) of about 90. Such lamps have life spans in excess of 25 000 hours.

The bulb can contain a heat sink or a plastic shroud that is vented at the top and bottom, allowing the lamp to be cooled by convection. Contemporary retrofit lamps exhibit wall-plug luminous efficacies in excess of 200 lm/W and are available at various luminous-flux levels and correlated color temperatures. The optimization of efficiency and of the distribution of emitted light has led to the development of lamps with a variety of shapes.

Retrofit lamps usually use screw bases and operate at line voltage. The circuitry incorporated within the bulb serves as a built-in driver. A collection of LEDs connected in series can be driven by a DC current obtained by rectifying AC line-voltage with diodes and capacitors. The LEDs can also be directly driven by the AC current, emitting light every other half cycle. Alternatively, wiring them as two antiparallel strands of series-connected LEDs results in half of them emitting light every half cycle. LED lamps can be dimmed either by reducing the applied voltage or by using a pulse-width-modulated current driver.

OLED Light Panels

OLED light panels are large-area light sources fashioned from organic light-emitting diodes (OLEDs), which are efficient generators of electroluminescence in the blue, green, and red. **White organic light-emitting diodes** (**WOLEDs**), fabricated in the manner prescribed in Fig. 18.1-22, generate white light via additive color mixing, as

depicted in Fig. 18.1-24. They have near-unity internal quantum efficiency and provide excellent color rendition. A white **OLED light panel** comprises a single, broad-area, vertically stacked OLED, such as that displayed in Fig. 18.1-22(b). Though typically designed to emit white light, OLED panels can alternatively be configured to emit light of any color, and to accommodate dynamic color tuning. Such panels have a spatial light-emission pattern that is nearly Lambertian so they offer large-area homogeneous lighting without glare. Though their luminous flux is limited, OLED light panels are available in a broad range of sizes, shapes, and color temperatures. When fabricated on transparent plastic substrates, they are lightweight, thin, and flexible, so they can be configured in unique shapes. A white OLED light panel is displayed in Fig. 18.1-28.

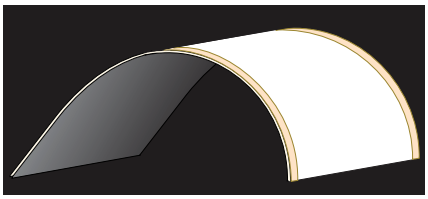


Figure 18.1-28 A $^1/_4$-mm-thick white OLED light panel that generates metameric white light with a luminous flux $P_v = 75$ lm, a luminous efficacy $\eta_v = 60$ lm/W, a CCT = 3000° K, and a CRI = 90. It has a life span of some 40 000 hours. Such panels offer large-area homogeneous illumination.

18.2 SEMICONDUCTOR OPTICAL AMPLIFIERS

Semiconductor optical amplifiers (SOAs), also called **semiconductor laser amplifiers (SLAs)**, are often used as photonic switches. As discussed in Sec. 24.3B, an SOA may be rapidly switched on and off by applying and removing an injected electric current. In the presence of gain (when the device is on), it acts as an amplifier, while in the absence of gain (when the device is off), it acts as an absorber. The SOA thus behaves as a fast photonic switch with a large extinction ratio. Moreover, arrays of SOAs may be interconnected via optical fibers to form complex photonic switches. Semiconductor optical amplifiers are also useful for wavelength conversion, as discussed in Sec. 24.3D, as well as for optical demultiplexing and optical clock recovery. They also serve as photonic logic gates in optical processing (Sec. 24.4). Though laser diodes enjoy wide use as sources in optical fiber communications, amplification in such systems is better served by optical fiber amplifiers (OFAs) such as EDFAs, REFAs, and RFAs, as explained in Sec. 25.1C.

The principle underlying the operation of an SOA is the same as that for other laser amplifiers: the creation of a population inversion that renders stimulated emission more prevalent than absorption. The population inversion is usually achieved by electric-current injection in some form of a p–n junction diode; a forward bias causes carrier pairs to be injected into the junction region where they recombine to emit stimulated emission. However, the theory of the SOA is somewhat more complex than that presented in Chapter 15 for the conventional laser amplifier inasmuch as the transitions take place between bands of closely spaced energy levels rather than between well-separated discrete energy levels or manifolds. For purposes of comparison, nevertheless, the SOA may be viewed as a device that operates via a form of in-band pumping.

The extension of the laser amplifier theory set forth in Chapter 15 to semiconductor devices was provided in Chapter 17. We proceed to use the results derived in Sec. 17.2 to obtain expressions for the gain and bandwidth of semiconductor optical amplifiers. We then consider pumping schemes suitable for attaining a population inversion and highlight the benefits of using heterostructure, quantum-well, and quantum-dot amplifier configurations. We then compare the performance of semiconductor optical amplifiers and optical fiber amplifiers, and finally consider superluminescent diodes.

The theoretical underpinnings of SOA operation carry over to laser-diode operation, considered in Sec. 18.3.

A. Gain and Bandwidth

Light of frequency ν can interact with the carriers of a semiconductor material of bandgap energy E_g via band-to-band transitions, provided that $\nu > E_g/h$. The incident photons may be absorbed, resulting in the generation of electron–hole pairs, or they may produce additional photons through stimulated electron–hole recombination radiation (Fig. 18.2-1). When emission is more likely than absorption, net optical gain ensues and material can serve as a coherent optical amplifier.

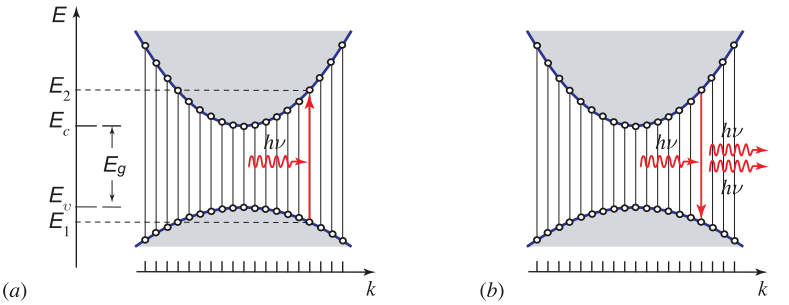


Figure 18.2-1 (*a*) The absorption of a photon results in the generation of an electron–hole pair. (*b*) Electron–hole recombination can be induced by a photon; the result is the stimulated emission of an identical photon.

Expressions for the rate of stimulated emission $r_{\rm st}(\nu)$ and the rate of photon absorption $r_{\rm ab}(\nu)$ are provided in (17.2-18) and (17.2-19), respectively. These quantities depend on the photon-flux spectral intensity ϕ_ν; the quantum-mechanical strength of the transition for the particular material under consideration (which is implicit in the value of the electron–hole radiative recombination lifetime $\tau_{\rm r}$); the optical joint density of states $\varrho(\nu)$; and the occupancy probabilities for emission $f_e(\nu)$ and absorption $f_a(\nu)$.

The optical joint density of states $\varrho(\nu)$ is determined by the E–k relations for electrons and holes and by the conservation of energy and momentum. With the help of the parabolic approximation for the E–k relations near the conduction- and valence-band edges, it was shown in (17.2-6) and (17.2-7) that the energies of the electron and hole that interact with a photon of energy $h\nu$ are

$$E_2 = E_c + \frac{m_r}{m_c}(h\nu - E_g), \quad E_1 = E_2 - h\nu, \tag{18.2-1}$$

respectively, where m_c and m_v are their effective masses and $1/m_r = 1/m_c + 1/m_v$. The resulting optical joint density of states that interacts with a photon of energy $h\nu$ was determined to be [see (17.2-9)]

$$\varrho(\nu) = \frac{(2m_r)^{3/2}}{\pi\hbar^2}\sqrt{h\nu - E_g}, \quad h\nu \geq E_g. \tag{18.2-2}$$

It is apparent that $\varrho(\nu)$ increases as the square root of photon energy above the bandgap.

The occupancy probabilities $f_e(\nu)$ and $f_a(\nu)$ are determined by the pumping rate through the quasi-Fermi levels E_{fc} and E_{fv}. The quantity $f_e(\nu)$ is the probability that a conduction-band state of energy E_2 is filled with an electron and a valence-band state of energy E_1 is filled with a hole. The quantity $f_a(\nu)$, on the other hand, is the

probability that a conduction-band state of energy E_2 is empty and a valence-band state of energy E_1 is filled with an electron. The Fermi inversion factor [see (17.2-25)]

$$f_g(\nu) = f_e(\nu) - f_a(\nu) = f_c(E_2) - f_v(E_1) \tag{18.2-3}$$

represents the degree of population inversion. The quantity $f_g(\nu)$ depends on both the Fermi function for the conduction band, $f_c(E) = 1/\{\exp[(E - E_{fc})/kT] + 1\}$, and the Fermi function for the valence band, $f_v(E) = 1/\{\exp[(E - E_{fv})/kT] + 1\}$. It is a function of temperature and of the quasi-Fermi levels E_{fc} and E_{fv}, which in turn are determined by the pumping rate. A complete population inversion can in principle be achieved in a semiconductor optical amplifier $[f_g(\nu) = 1]$, so it behaves like a four-level laser system in that respect (Sec. 15.2B).

The results provided above were combined in (17.2-24) to provide an expression for the net gain coefficient, $\gamma_0(\nu) = [r_{\rm st}(\nu) - r_{\rm ab}(\nu)]/\phi_\nu$,

$$\boxed{\gamma_0(\nu) = \frac{\lambda^2}{8\pi\tau_{\rm r}}\, \varrho(\nu)\, f_g(\nu).} \tag{18.2-4}$$

Gain Coefficient

Comparing (18.2-4) with (15.1-4), it is apparent that the quantity $\varrho(\nu) f_g(\nu)$ in the semiconductor optical amplifier plays the role of $Ng(\nu)$ in other laser amplifiers, and that $\sigma(\nu) \approx \gamma_0(\nu)/\Delta{\rm n}$.

Amplifier Bandwidth

In accordance with (18.2-3) and (18.2-4), a semiconductor medium provides net optical gain at the frequency ν when $f_c(E_2) > f_v(E_1)$. Conversely, net attenuation ensues when $f_c(E_2) < f_v(E_1)$. Thus, a semiconductor material in thermal equilibrium (undoped or doped) cannot provide net gain whatever its temperature; this is because the conduction- and valence-band Fermi levels coincide ($E_{fc} = E_{fv} = E_f$). External pumping is required to separate the Fermi levels of the two bands in order to achieve amplification.

The condition $f_c(E_2) > f_v(E_1)$ is equivalent to the requirement that the photon energy be smaller than the separation between the quasi-Fermi levels, i.e., $h\nu < E_{fc} - E_{fv}$, as demonstrated in Exercise 17.2-1. Of course, the photon energy must be larger than the bandgap energy ($h\nu > E_g$) in order that laser amplification occur by means of interband transitions. Thus, if the pumping rate is sufficiently large that the separation between the two quasi-Fermi levels exceeds the bandgap energy E_g, the medium can act as an amplifier for optical frequencies in the band

$$\boxed{\frac{E_g}{h} < \nu < \frac{E_{fc} - E_{fv}}{h}\,.} \tag{18.2-5}$$

Amplifier Bandwidth

For $h\nu < E_g$ the medium is transparent, whereas for $h\nu > E_{fc} - E_{fv}$ it is an attenuator instead of an amplifier. Equation (18.2-5) demonstrates that the amplifier bandwidth increases with $E_{fc} - E_{fv}$, and therefore with pumping level. In this respect it is unlike the atomic laser amplifier, which has an unsaturated bandwidth $\Delta\nu$ that is independent of pumping level (Fig. 15.1-2).

Computation of the gain properties is simplified considerably if thermal excitations can be ignored (i.e., if $T = 0^\circ$ K). The Fermi functions are then simply $f_c(E_2) = 1$ for $E_2 < E_{fc}$ and 0 otherwise; $f_v(E_1) = 1$ for $E_1 < E_{fv}$ and 0 otherwise. In that case the

Fermi inversion factor is

$$f_g(\nu) = \begin{cases} +1, & h\nu < E_{fc} - E_{fv} \\ -1, & \text{otherwise.} \end{cases} \tag{18.2-6}$$

Schematic plots of the functions $\varrho(\nu)$, $f_g(\nu)$, and the gain coefficient $\gamma_0(\nu)$ are presented in Fig. 18.2-2, illustrating how $\gamma_0(\nu)$ changes sign and turns into a loss coefficient when $h\nu > E_{fc} - E_{fv}$. The ν^{-2} dependence of $\gamma_0(\nu)$, arising from the λ^2 factor in the numerator of (18.2-4), varies sufficiently slowly that it may be ignored. Finite temperature smoothes the functions $f_g(\nu)$ and $\gamma_0(\nu)$, as shown by the dashed curves in Fig. 18.2-2.

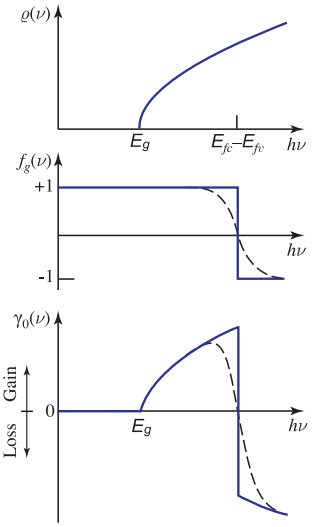


Figure 18.2-2 Dependence on energy of the optical joint density of states $\varrho(\nu)$, the Fermi inversion factor $f_g(\nu)$, and the gain coefficient $\gamma_0(\nu)$ at $T = 0°$ K (solid curves) and at room temperature (dashed curves). Photons with energy between E_g and $E_{fc} - E_{fv}$ undergo laser amplification.

Dependence of the Gain Coefficient on Pumping Level

The gain coefficient $\gamma_0(\nu)$ increases both in its width and in its magnitude as the pumping rate R is elevated. As provided in (18.1-1), a constant pumping rate R (number of injected excess electron–hole pairs per cm^3 per second) establishes a steady-state concentration of injected electron–hole pairs in accordance with $\Delta\mathrm{n} = \Delta\mathrm{p} = R\tau$, where τ is the electron–hole recombination lifetime (which includes both radiative and nonradiative contributions). Knowledge of the steady-state total concentrations of electrons and holes, $\mathrm{n} = \mathrm{n}_0 + \Delta\mathrm{n}$ and $\mathrm{p} = \mathrm{p}_0 + \Delta\mathrm{n}$, respectively, permits the Fermi levels E_{fc} and E_{fv} to be determined via (18.1-8) and (18.1-9). Once the Fermi levels are known, the computation of the gain coefficient can proceed using (18.2-4). The dependence of $\gamma_0(\nu)$ on $\Delta\mathrm{n}$, and thereby on R, is illustrated in Example 18.2-1.

EXAMPLE 18.2-1. ***Gain Coefficient for an InGaAsP SOA.*** A sample of the quaternary material $\mathrm{In}_{0.72}\mathrm{Ga}_{0.28}\mathrm{As}_{0.6}\mathrm{P}_{0.4}$, with bandgap energy $E_g = 0.95$ eV, is operated as a semiconductor optical amplifier at a wavelength of $\lambda_o = 1300$ nm at $T = 300°$ K. The sample is undoped but has residual concentrations of $\approx 2 \times 10^{17}$ cm^{-3} donors and acceptors, and a radiative electron–hole recombination lifetime $\tau_\mathrm{r} \approx 2.5$ ns. The effective masses of the electrons and holes are $m_c \approx 0.06\, m_0$ and $m_v \approx 0.4\, m_0$, respectively, and the refractive index $n \approx 3.5$. Given the steady-state injected-carrier concentration $\Delta\mathrm{n}$ (which is controlled by the injection rate R and the overall recombination time τ), the gain coefficient $\gamma_0(\nu)$ may be computed from (18.2-4) in conjunction with (18.1-8)

and (18.1-9). As illustrated in Fig. 18.2-3, both the amplifier bandwidth and the peak value of the gain coefficient γ_p increase with Δn. The energy at which the peak occurs also increases with Δn, as expected from the behavior shown in Fig. 18.2-2. Furthermore, the minimum energy at which amplification occurs decreases slightly with increasing Δn as a result of band-tail states, which reduce the bandgap energy. At the largest value of Δn shown ($\Delta n = 1.8 \times 10^{18}$ cm^{-3}), photons with energies falling between 0.91 and 0.97 eV undergo amplification. This corresponds to a full amplifier bandwidth of 14.5 THz, and a wavelength range of 80 nm. A more suitable measure is the bandwidth at the full-width at half-maximum (FWHM) of the gain profile, also called the 3-dB gain bandwidth, which is 10 THz, corresponding to about 50 nm at $\lambda_o = 1300$ nm (see Table 15.3-1 for a comparison with other laser transitions). The calculated peak gain coefficient $\gamma_p = 270$ cm^{-1} at this value of Δn is large in comparison with most atomic laser amplifiers.

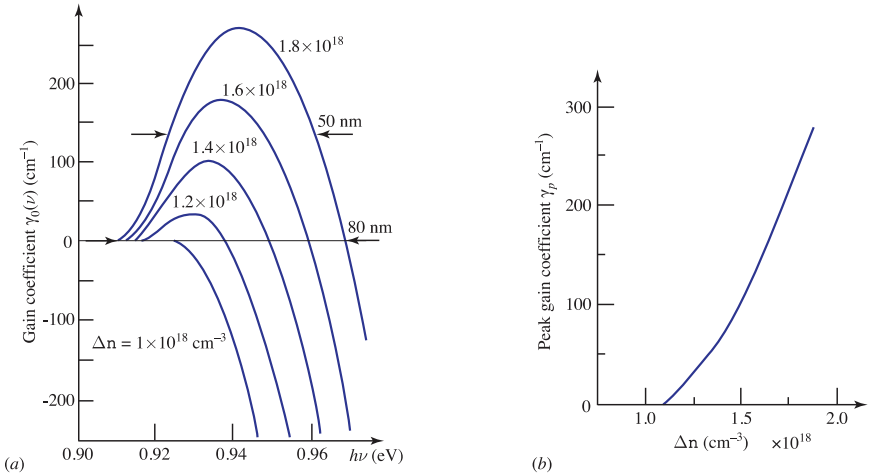


Figure 18.2-3 (a) Calculated gain coefficient $\gamma_0(\nu)$ for an InGaAsP SOA versus photon energy $h\nu$, with the injected-carrier concentration Δn as a parameter ($T = 300°$ K). The band of frequencies over which amplification occurs (centered near 1300 nm) increases with increasing Δn. At the largest value of Δn shown, the FWHM amplifier bandwidth is 10 THz, corresponding to 0.04 eV in energy and 50 nm in wavelength. (Adapted from N. K. Dutta, Calculated Absorption, Emission, and Gain in $In_{0.72}Ga_{0.28}As_{0.6}P_{0.4}$, *Journal of Applied Physics*, vol. 51, pp. 6095–6100, 1980, Fig. 8.) (b) Calculated peak gain coefficient γ_p as a function of Δn. At the largest value of Δn, the peak gain coefficient $\approx$ 270 cm^{-1}. (Adapted from N. K. Dutta and R. J. Nelson, The Case for Auger Recombination in $In_{1-x}Ga_xAs_yP_{1-y}$, *Journal of Applied Physics*, vol. 53, pp. 74–92, 1982, Fig. 17.)

The onset of gain saturation in semiconductor optical amplifiers is not unlike that of other homogeneously broadened laser amplifiers, as considered in Sec. 15.4. The relatively large semiconductor transition cross section (Table 15.3-1) implies a small saturation photon-flux density [$\phi_s \approx 1/\tau_r \sigma(\nu)$] and therefore a reduced gain coefficient [see (15.4-2) and (15.4-3)]. This in turn limits the overall gain that an SOA can provide.

In common with other optical amplifiers, SOAs suffer from amplified spontaneous emission noise (Sec. 15.5); however, they are also affected by noise associated with temperature and carrier fluctuations.

Approximate Peak Gain Coefficient

The complex dependence of the gain coefficient on the injected-carrier concentration makes the analysis of the semiconductor amplifier (and laser) somewhat difficult. Because of this, it is customary to adopt an empirical approach in which the peak gain coefficient γ_p is assumed to be linearly related to Δn for values of Δn near the operating point. As the example in Fig. 18.2-3(b) illustrates, the approximation is reasonable when γ_p is sufficiently large. The dependence of the peak gain coefficient

γ_p on Δn may then be modeled by the linear relation

$$\gamma_p \approx \alpha \left(\frac{\Delta n}{\Delta n_T} - 1 \right), \tag{18.2-7}$$

Peak Gain Coefficient (Linear Approximation)

which is illustrated in Fig. 18.2-4. The parameters α and Δn_T are chosen to satisfy the following limits:

- When $\Delta n = 0$, $\gamma_p = -\alpha$, where α represents the absorption coefficient of the semiconductor in the absence of current injection.
- When $\Delta n = \Delta n_T$, $\gamma_p = 0$. Thus, Δn_T is the injected-carrier concentration at which emission and absorption just balance so that the medium is transparent.

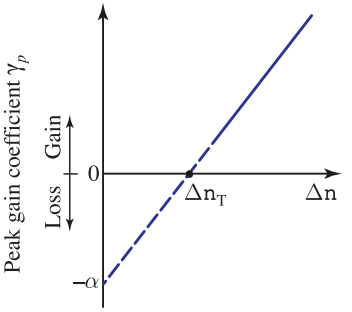


Figure 18.2-4 Peak value of the gain coefficient γ_p as a function of injected-carrier concentration Δn for the approximate linear model. The quantity α represents the attenuation coefficient in the absence of injection, whereas Δn_T represents the injected-carrier concentration at which emission and absorption just balance each other. The solid portion of the straight line matches the more realistic calculation considered in the preceding subsection.

EXAMPLE 18.2-2. ***Approximate Peak Gain Coefficient for an InGaAsP SOA.*** The peak gain coefficient γ_p versus Δn for InGaAsP presented in Fig. 18.2-3(*b*) may be approximately fit by a linear relation that takes the form of (18.2-7), with the parameters $\Delta n_T \approx 1.25 \times 10^{18}$ cm^{-3} and $\alpha = 600$ cm^{-1}. For $\Delta n = 1.4\,\Delta n_T = 1.75 \times 10^{18}$ cm^{-3}, the linear model yields a peak gain coefficient $\gamma_p = 240$ cm^{-1}. For an InGaAsP crystal of length $d = 350$ μm, this corresponds to a total gain $\exp(\gamma_p d) \approx 4447$ or 36.5 dB. In practice, this value is reduced by gain saturation, as discussed above, as well as by coupling losses, which are typically 3 to 5 dB per facet.

Increasing the injected-carrier concentration from below to above the transparency value Δn_T results in the semiconductor changing from a strong absorber of light $[f_g(\nu) < 0]$ into a high-gain amplifier of light $[f_g(\nu) > 0]$. The very same large transition probability that makes the semiconductor a good absorber also makes it a good amplifier, as may be understood by comparing (17.2-18) and (17.2-19).

B. Pumping

Optical Pumping

Pumping may be achieved by means of external light, as depicted in Fig. 18.2-5, provided that its photon energy is sufficiently large ($> E_g$). Pump photons are absorbed by the semiconductor, resulting in the generation of carrier pairs. The generated electrons and holes decay to the bottom of the conduction band and the top of the valence band, respectively. If the intraband relaxation time is much shorter than the interband

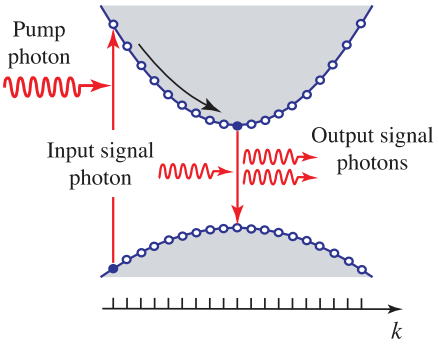


Figure 18.2-5 Optical pumping of a semiconductor optical amplifier.

relaxation time, as is usually the case, a steady-state population inversion between the bands may be established, as discussed in Sec. 15.2 for conventional laser amplifiers.

Current Pumping

A more practical scheme for pumping a semiconductor optical amplifier is by means of electron–hole injection in a heavily doped p–n junction — a diode. As with the LED (Sec. 18.1) the junction is forward biased so that minority carriers are injected into the junction region (electrons into the p-type region and holes into the n-type region). Figure 18.1-5 shows the energy-band diagram of a forward-biased heavily doped p–n junction. The conduction-band and valence-band quasi-Fermi levels, E_{fc} and E_{fv}, lie within the conduction and valence bands, respectively, and a state of quasi-equilibrium exists within the junction region. The quasi-Fermi levels are sufficiently well separated so that a population inversion is achieved and net gain may be obtained over the bandwidth $E_g \leq h\nu \leq E_{fc} - E_{fv}$ within the active region. The thickness l of the active region is an important parameter of the diode that is determined principally by the diffusion lengths of the minority carriers at both sides of the junction. Typical values of l for InGaAsP are 1–3 μm.

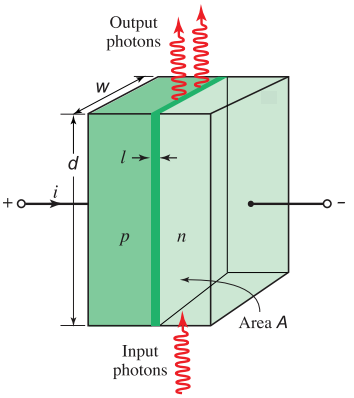


Figure 18.2-6 Geometry of a simple semiconductor optical amplifier. Charge carriers travel perpendicularly to the p–n junction, whereas photons travel in the plane of the junction.

If an electric current i is injected through an area $A = wd$, where w and d are the width and height of the device, respectively, into a volume lA (as portrayed in Fig. 18.2-6), then the steady-state carrier injection rate is $R = i/elA = J/el$ per second per unit volume, where $J = i/A$ is the injected current density. The resulting injected-

carrier concentration is then

$$\Delta \mathrm{n} = \tau R = \frac{\tau}{elA}\, i = \frac{\tau}{el}\, J. \tag{18.2-8}$$

The injected-carrier concentration is therefore directly proportional to the injected current density so that the results shown in Figs. 18.2-3(b) and 18.2-4 with $\Delta \mathrm{n}$ as a parameter may just as well have J as a parameter. In particular, it follows from (18.2-7) and (18.2-8) that within the linear approximation implicit in (18.2-7), the peak gain coefficient is linearly related to the injected current density J, i.e.,

$$\boxed{\gamma_p \approx \alpha\left(\frac{J}{J_{\mathrm{T}}} - 1\right).} \tag{18.2-9}$$

Peak Gain Coefficient

The transparency current density J_{T} is given by

$$\boxed{J_{\mathrm{T}} = \frac{el}{\eta_i \tau_{\mathrm{r}}}\, \Delta \mathrm{n}_{\mathrm{T}},} \tag{18.2-10}$$

Transparency Current Density

where $\eta_i = \tau/\tau_{\mathrm{r}}$ again represents the internal quantum efficiency.

When $J = 0$, the peak gain coefficient $\gamma_p = -\alpha$ becomes the attenuation coefficient, as is apparent in Fig. 18.2-7. When $J = J_{\mathrm{T}}$, $\gamma_p = 0$ so that the material is transparent and neither amplifies nor attenuates. Net gain can be achieved only when the injected current density J exceeds its transparency value J_{T}. Note that J_{T} is directly proportional to the junction thickness l so that a lower transparency current density J_{T} is achieved by using a narrower active-region thickness. This is an important consideration in the design of semiconductor optical amplifiers (and lasers).

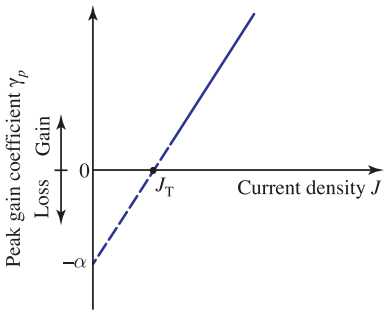


Figure 18.2-7 Peak optical gain coefficient γ_p as a function of current density J for the approximate linear model. When $J = J_{\mathrm{T}}$ the material is transparent and exhibits neither gain nor loss.

EXAMPLE 18.2-3. ***Gain of an InGaAsP SOA.*** An InGaAsP semiconductor optical amplifier operates at 300° K and has the following parameters: $\tau_{\mathrm{r}} = 2.5$ ns, $\eta_i = 0.5$, $\Delta \mathrm{n}_{\mathrm{T}} = 1.25 \times 10^{18}$ cm^{-3}, and $\alpha = 600$ cm^{-1}. The junction has thickness $l = 2$ μm, length $d = 200$ μm, and width $w = 10$ μm. Using (18.2-10), the current density that just makes the semiconductor transparent is $J_{\mathrm{T}} = 3.2 \times 10^4$ A/cm^2. A slightly larger current density $J = 3.5 \times 10^4$ A/cm^2 provides a peak gain coefficient $\gamma_p \approx 56$ cm^{-1} as is clear from (18.2-9). This gives rise to an amplifier gain $G = \exp(\gamma_p d) = \exp(1.12) \approx 3$. However, since the junction area $A = wd = 2 \times 10^{-5}$ cm^2, a rather large injection current $i = JA = 700$ mA is required to produce this current density.

Motivation for Heterostructures

If the thickness l of the active region in Example 18.2-3 were reduced from 2 μm to, say, 0.1 μm, the current density J_T would be reduced by a factor of 20, to the more reasonable value 1600 A/cm^2. Because proportionately less volume would have to be pumped, the amplifier could then provide the same gain with a lower injected current density. Such a reduction in the thickness of the active region poses a potential problem, however, because the diffusion lengths of the electrons and holes in InGaAsP are several μm and the carriers would tend to diffuse out of this smaller region. However, it is possible to confine carriers to an active region whose thickness is smaller than their diffusion lengths by making use of a heterostructure device, as discussed in Sec. 18.2C. Indeed, light can simultaneously be confined in such a structure, providing an additional advantage.

C. Heterostructures

As is apparent from (18.2-9) and (18.2-10), the diode-laser peak amplifier gain coefficient γ_p varies inversely with the thickness l of the active region. It is therefore advantageous to use the smallest thickness possible. The active region is defined by the diffusion distances of minority carriers on both sides of the junction. The concept of the double heterostructure is to form heterojunction potential barriers on both sides of the p–n junction to provide a potential well that limits the distance over which minority carriers may diffuse. The junction barriers define a region of space within which minority carriers are confined, allowing active regions of thickness l as small as 0.1 μm to be achieved. Yet thinner confinement regions, ≈ 0.01 μm, can be attained by making use of quantum-well structures, as will be discussed in Sec. 18.2D.

Electromagnetic confinement of the amplified optical beam can be achieved simultaneously if the material of the active layer is selected such that its refractive index is slightly greater than that of the two surrounding layers, in which case the structure acts as an optical waveguide (Sec. 9.2).

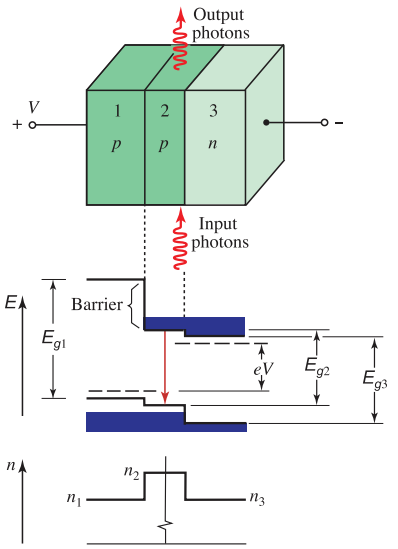


Figure 18.2-8 Energy-band diagram and refractive index as functions of position for a double-heterostructure semiconductor optical amplifier.

The double-heterostructure design therefore calls for three layers of different lattice-matched materials, as illustrated in Fig. 18.2-8:

Layer 1: p-type, energy bandgap E_{g1}, refractive index n_1
Layer 2: p-type, energy bandgap E_{g2}, refractive index n_2
Layer 3: n-type, energy bandgap E_{g3}, refractive index n_3

The semiconductor materials are selected such that E_{g1} and E_{g3} are greater than E_{g2}, which achieves carrier confinement, while n_2 is greater than n_1 and n_3, which achieves light confinement. The active layer (layer 2) is made quite thin (0.1 to 0.2 μm) to minimize the transparency current density J_{T} and thereby to maximize the peak gain coefficient γ_p. Stimulated emission takes place in the p–n junction between layers 2 and 3.

In summary, the double-heterostructure design offers the following advantages:

- Increased amplifier gain, for a given injected current density, as a result of decreased active-layer thickness, in accordance with (18.2-9) and (18.2-10). Injected minority carriers are confined within the thin active layer between the two heterojunction barriers and are prevented from diffusing to the surrounding layers.
- Increased amplifier gain resulting from the confinement of photons within the active layer as a result of its larger refractive index. The active medium acts as an optical waveguide.
- Reduced loss, resulting from the inability of layers 1 and 3 to absorb the guided photons because the bandgaps of these layers, E_{g1} and E_{g3}, are larger than the photon energy ($h\nu = E_{g2} < E_{g1},\ E_{g2}$).

Two examples of double-heterostructure semiconductor optical amplifiers follow:

- *InGaAsP/InP Double-Heterostructure Laser Diode Amplifier.* The active layer, $\mathrm{In}_{1-x}\mathrm{Ga}_x\mathrm{As}_{1-y}\mathrm{P}_y$, is surrounded by layers of InP. The composition parameters x and y are selected so that the materials are lattice matched. Operation is thereby restricted to a range of values of x and y for which E_{g2} corresponds to the wavelength band 1.1–1.7 μm.
- *GaAs/AlGaAs Double-Heterostructure Laser Diode Amplifier.* The active layer (layer 2) is fabricated from GaAs ($E_{g2} = 1.42$ eV, $n_2 = 3.6$). The surrounding layers (1 and 3) are fabricated from $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ with $E_g > 1.43$ eV and $n < 3.6$ (by 5–10%). This amplifier typically operates within the 0.82–0.88-μm wavelength band when the AlGaAs composition parameter is in the range $x = 0.35$–0.5.

D. Quantum-Well Structures

As discussed in Sec. 18.2C, heterostructures offer a reduced thickness of the active layer within which carriers and photons are confined. This in turn provides increased amplifier gain and reduced amplifier loss. When the thickness of the active layer is reduced yet further, say to 5–10 nm (which is smaller than the de Broglie wavelength of a thermalized electron), quantum effects play a key role. Since the active layer in a double heterostructure has a bandgap energy smaller than that of the surrounding layers, the structure then acts as a quantum well (Sec. 17.1G), and is referred to as a **quantum-well device**.

The band structure and energy–momentum (E–k) relations of a quantum well are different from those of a bulk material. The conduction band is split into a number of subbands, labeled by the quantum number $q = 1, 2, \ldots$, each with its own energy–momentum relation and density of states. The bottoms of these subbands have energies $E_c + E_q$, where $E_q = \hbar^2(q\pi/l)^2/2m_c,\ q = 1, 2, \ldots,$ are the energies of an electron of effective mass m_c in a one-dimensional quantum well of thickness l (see Figs. 17.1-25 and 17.1-27; q_1 and d_1 in Chapter 17 correspond to q and l here). Each subband has a parabolic E–k relation and a constant density of states that is independent of

energy. The overall density of states in the conduction band, $\varrho_c(E)$, therefore assumes a staircase distribution [see (17.1-37)] with steps at energies E_c+E_q, $q = 1, 2, \ldots$. The valence band has similar subbands at energies $E_v - E'_q$, where $E'_q = \hbar^2(q\pi/l)^2/2m_v$ are the energies of a hole of effective mass m_v in a quantum well of thickness l.

The interactions of photons with electrons and holes in a quantum well take the form of energy- and momentum-conserving transitions between the conduction and valence bands. The transitions must also conserve the quantum number q, as illustrated in Fig. 18.2-9(a); they obey rules similar to those that govern transitions between the conduction and valence bands in bulk semiconductors. The expressions for the transition probabilities and gain coefficient in the bulk material (Sec. 17.2) apply to the quantum-well structure if we simply replace the bandgap energy E_g with the energy gap between the subbands, $E_{gq} = E_g + E_q + E'_q$, and use a constant density of states rather than one that varies as the square root of energy. The total gain coefficient is the sum of the gain coefficients provided by all of the subbands ($q = 1, 2, \ldots$).

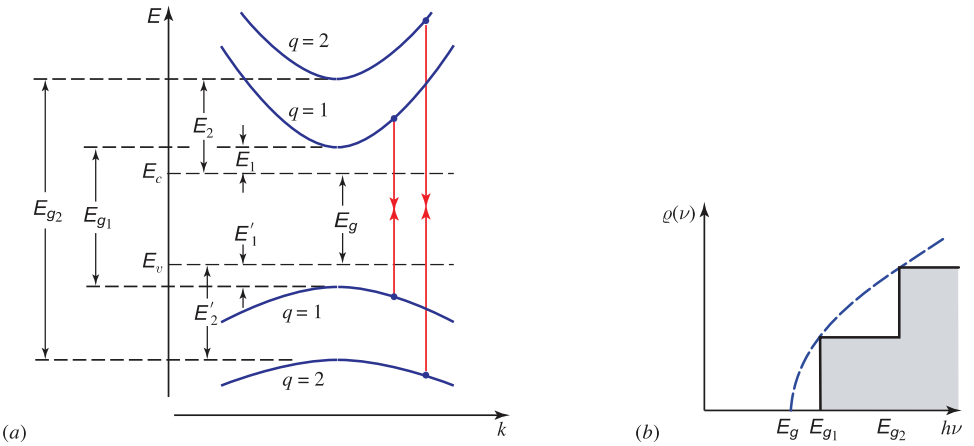


Figure 18.2-9 (a) E–k relations of different subbands. (b) Optical joint density of states for a quantum-well structure (staircase curve) and for a bulk semiconductor (dashed curve). The first jump occurs at energy $E_{g1} = E_g + E_1 + E'_1$ (where E_1 and E'_1 are, respectively, the lowest energies of an electron and a hole in the quantum well).

Density of States

Consider transitions between the two subbands of quantum number q. To satisfy the conservation of energy and momentum, a photon of energy $h\nu$ interacts with states of energies $E = E_c + E_q + (m_r/m_c)(h\nu - E_{gq})$ in the upper subband and $E - h\nu$ in the lower. The optical joint density of states $\varrho(\nu)$ is related to $\varrho_c(E)$ by $\varrho(\nu) = (dE/d\nu)\,\varrho_c(E) = (hm_r/m_c)\,\varrho_c(E)$. It follows from (17.1-37) that

$$\varrho(\nu) = \begin{cases} \dfrac{hm_r}{m_c}\dfrac{m_c}{\pi\hbar^2 l} = \dfrac{2m_r}{\hbar l}, & h\nu > E_g + E_q + E'_q \\ 0, & \text{otherwise.} \end{cases} \tag{18.2-11}$$

Including transitions between all subbands $q = 1, 2, \ldots$, we arrive at a $\varrho(\nu)$ that has a staircase distribution with steps at the energy gaps between subbands of the same quantum number [Fig. 18.2-9(b)].

Gain Coefficient

The gain coefficient of the device is given by the usual expression [see (17.2-24)]

$$\gamma_0(\nu) = \frac{\lambda^2}{8\pi\tau_r}\,\varrho(\nu)\,f_g(\nu)\,, \tag{18.2-12}$$

where the Fermi inversion factor $f_g(\nu)$ depends on the quasi-Fermi levels and temperature, and is the same for bulk and quantum-well lasers. The density of states $\varrho(\nu)$, however, differs in the two cases, as we have shown. The frequency dependences of $\varrho(\nu)$, $f_g(\nu)$, and their product are illustrated in Fig. 18.2-10 for quantum-well and bulk double-heterostructure configurations. The quantum-well structure has a smaller peak gain coefficient and a narrower gain profile. It is assumed in the construction of

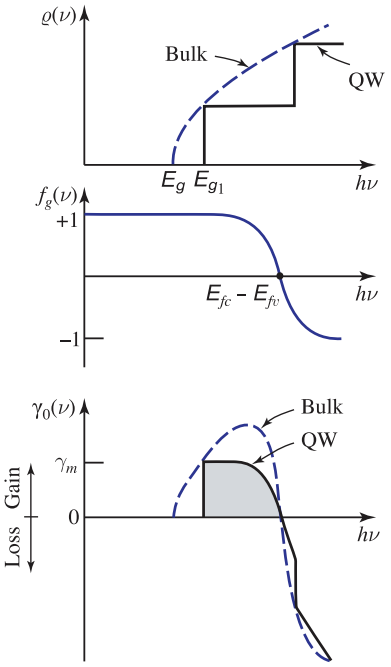


Figure 18.2-10 Density of states $\varrho(\nu)$, Fermi inversion factor $f_g(\nu)$, and gain coefficient $\gamma_0(\nu)$ in quantum-well (solid) and bulk (dashed) structures.

Fig. 18.2-10 that only a single step of the staircase function $\varrho(\nu)$ occurs at an energy smaller than $E_{fc} - E_{fv}$, as is the case under usual injection conditions.

The maximum gain coefficient γ_m may then be determined by substituting $f_g(\nu) = 1$ and $\varrho(\nu) = 2m_r/\hbar l$ in (18.2-12), which yields

$$\gamma_m = \frac{\lambda^2 m_r}{2\tau_r h l}\,. \tag{18.2-13}$$

Relation Between Gain Coefficient and Current Density

By increasing the injected current density J, the concentration of excess electrons and holes Δn is increased and, therefore, so is the separation between the quasi-Fermi levels $E_{fc} - E_{fv}$. The effect of this increase on the gain coefficient $\gamma_0(\nu)$ may be assessed by examining the diagrams in Fig. 18.2-10. For sufficiently small J there is no gain. When J is such that $E_{fc} - E_{fv}$ just exceeds the gap E_{g1} between the $q = 1$ subbands, the medium provides gain. The peak gain coefficient increases sharply and

saturates at the value γ_m. An increase of J increases the gain spectral width but not its peak value. If J is increased yet further, to the point where $E_{fc} - E_{fv}$ exceeds the gap E_{g2} between the $q = 2$ subbands, the peak gain coefficient undergoes another jump, and so on. The gain profile can therefore be quite broad, thereby providing a wide tuning range for such devices.

Materials and Device Structures

The structure of a semiconductor optical amplifier resembles that of a laser diode operated above transparency but below the threshold of oscillation (Sec. 18.3). Semiconductor optical amplifiers can be made to operate in any region of the optical spectrum by judiciously choosing the semiconductor material and using compositional tuning. The center wavelength, bandwidth, and gain depend both on the material and on the structure of the device.

SOAs designed for applications in the near infrared are usually fabricated from InGaAsP, InGaAs, or InP. In the 1300–1600-nm telecommunications band, achievable bandwidths are $\Delta\lambda \approx 50$ nm, corresponding to $\Delta\nu \approx 10$ THz at $\lambda_o = 1300$ nm (Example 18.2-1). This is broader than the bandwidths offered by EDFAs but is similar to those provided by RFAs (see Secs. 15.3C and 15.3D, respectively). Quantum-well SOAs offer a substantial reduction in the drive current required to achieve transparency but otherwise behave similarly to bulk devices.

The gain of an SOA is usually limited to ≈ 15 dB because of gain saturation and insertion losses of 3–5 dB per facet (Example 18.2-2). Saturation leads to interchannel and intersymbol interference, rendering SOAs unsuitable for use in DWDM communication systems. Furthermore, the short semiconductor recombination time (Table 15.3-1) leaves the SOA susceptible to high-frequency noise that might reside in the pumping current and optical signal, leading to noise figures $\approx$ 8–10 dB as opposed to 3 dB for EDFAs. It is important to note that if an SOA is to be operated as a broadband single-pass device (i.e., as a traveling-wave amplifier), the facet reflectances must be reduced to a minimum. Failure to do so can lead to multiple reflections and a gain profile that is modulated by the resonator modes; this can also result in oscillation, which, of course, obviates the possibility of controllable amplification. Techniques for reducing reflectances include the use of antireflection coatings and tilted waveguides.

As a result of the issues discussed above, optical-transmission applications of SOAs have, for the most part, been limited to metropolitan optical networks where low gain suffices for overcoming losses associated with multiple optical add–drop nodes. The appeal of SOAs, rather, resides in their use as photonic switches (see Sec. 24.3B), wavelength converters (see Sec. 24.3D), and logic gates in optical processing. They also are useful for optical demultiplexing and optical clock recovery.

EXAMPLE 18.2-4. ***Waveguide Amplifiers.*** Multiquantum-well semiconductor optical amplifiers can be constructed in the form of optical waveguides, providing operation in fundamental optical modes at increased output saturation powers, and employing direct butt coupling to single-mode fibers. Such devices have relatively low losses and a small optical confinement factor. As an example, a 1550-nm InGaAsP/InP quantum-well amplifier with a length of 1 cm provides a fiber-to-fiber gain of 13 dB.

Comparison of Quantum-Dot and Quantum-Well SOAs

The quantum-dot semiconductor optical amplifier (QD-SOA) offers many of the advantages and disadvantages offered by its quantum-well counterpart, and indeed their

performance under CW operation is comparable. As a result of inhomogeneous broadening, QD-SOAs enjoy bandwidths that can extend up to 200 nm, corresponding to $\Delta\nu \approx 25$ THz at $\lambda_o = 1550$ nm; however, a concomitant reduction in the gain per unit bandwidth and saturated output power ensues. Nevertheless, the QD-SOA is distinguished by its greater unsaturated gain and faster gain dynamics, which provide efficient amplification for short optical pulses and pulse trains. Gain recovery times can extend down to 100 fs, corresponding to operation at > 200 Gb/s.

Comparison of SOAs and OFAs

The semiconductor optical amplifier enjoys advantages and disadvantages with respect to optical fiber amplifiers such as the erbium-doped fiber amplifier and the Raman fiber amplifier:

Advantages of SOAs:

- Central wavelength selectable by choice of material
- Compatible with photonic integrated circuits
- Electrical pumping
- Readily modulated via injection current
- Compact
- Low cost

Disadvantages of SOAs:

- Low gain
- Low saturated output power
- High noise
- Substantial interchannel and intersymbol interference
- Sensitivity to thermal effects from heat dissipation
- Sensitivity to facet reflections
- Sensitivity to signal polarization
- Control of transverse-mode characteristics
- High insertion loss
- Incompatibility with fiber geometry

On balance, the performance of the SOA is generally inferior to that of the EDFA and the RFA, and its use is generally restricted to special applications. The relative merits of EDFAs and RFAs were considered in Sec. 15.3D.

E. Superluminescent Diodes

Superluminescent diodes (**SLEDs**) are semiconductor optical amplifiers (SOAs) operated without an optical signal presented to the input. The light emitted from a SLED is **amplified spontaneous emission (ASE)** produced by the device itself. The ASE takes the same form as the optical noise emitted by a conventional laser amplifier, as discussed in Sec. 15.5. The SLED is distinguished from an LED in that the level of current injection is sufficiently high so that stimulated emission outweighs spontaneous emission.

An example of a SLED is the multiquantum-well InGaAsP/InP structure displayed in Fig. 18.2-11. The optical output power generated by a SLED is generally greater than that of an LED but less than that of an LD (Fig. 18.3-5), while the normalized spectral intensity is typically narrower than that of an LED but broader than that of an LD (Fig. 18.3-7). A SLED can offer diffraction-limited and spatially coherent emission comparable to that of an edge-emitting LD, which facilitates coupling the output light into a single-mode fiber. As with the semiconductor optical amplifier, it is important to minimize optical feedback to avoid lasing from occurring. This may be achieved in any

number of ways, such as by making use of a stripe contact that injects current only over a portion of the device, by using a tapered-stripe geometry, or by antireflection-coating or tilting the facets of the device.

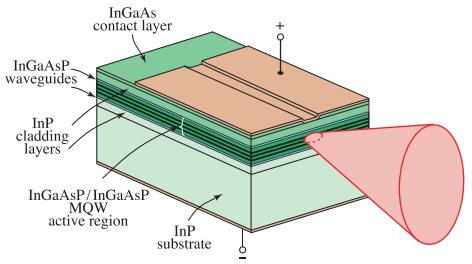


Figure 18.2-11 A MQW InGaAsP/InP superluminescent diode. SLEDs can generate light with substantial optical power and with a bandwidth intermediate between that of an LED and an LD. It is important to minimize feedback so that laser oscillation does not occur. One way of achieving this is to use a stripe contact that injects current only over a portion of the device (illustrated), which increases the loss via absorption.

Superluminescent diodes are used in applications where the long coherence time of laser light is troublesome, either because its spectrum is too narrow or because of randomly occurring interferences (speckle). Examples of such applications include interferometric instrumentation such as optical coherence tomography (Sec. 12.2B), fiber-optic gyroscopy, and certain fiber-optic sensors. Optical fiber amplifiers are also sometimes used as sources of superluminescence light.

18.3 LASER DIODES

In this section we consider the general characteristics of conventional laser diodes. The earliest devices, fabricated in 1962, comprised single p–n junctions of GaAs and GaAsP, which emitted in the near infrared and red, respectively (see p. 787). Semiconductor lasers have been fabricated in a bewildering variety of forms. They operate at wavelengths that stretch from the mid-ultraviolet to the far-infrared — and at output powers that range from nW (for nanolasers) to W (for individual laser diodes) to kW (for banks of laser diodes). Today's laser diodes take many forms. Quantum-confined lasers are discussed in Sec. 18.4, and compact lasers in the form of microcavity and nanocavity devices are considered in Secs. 18.5 and 18.6, respectively.

A. Amplification, Feedback, and Oscillation

A laser diode is a semiconductor optical amplifier that is endowed with a path for optical feedback. As discussed in the preceding section, a semiconductor optical amplifier is a forward-biased heavily doped p–n junction fabricated from a direct-bandgap semiconductor material. The injected current is sufficiently large so as to provide optical gain. In its simplest configuration, the optical feedback is provided by plane mirrors, which are usually implemented by cleaving the semiconductor material along its crystal planes. The sharp refractive index difference between the crystal and the surrounding air causes the cleaved surfaces to act as reflectors. Thus, the semiconductor crystal acts both as a gain medium and as a Fabry–Perot optical resonator, as illustrated in Fig. 18.3-1. Provided that the gain coefficient is sufficiently large, the feedback converts the optical amplifier into an optical oscillator, i.e., a laser. The device is called a **laser diode** or a **diode laser** (it is also sometimes referred to as a **semiconductor injection laser**).

The laser diode (LD) bears considerable similarity to the light-emitting diode (LED) discussed in Sec. 18.1. In both devices, the source of energy is an electric current

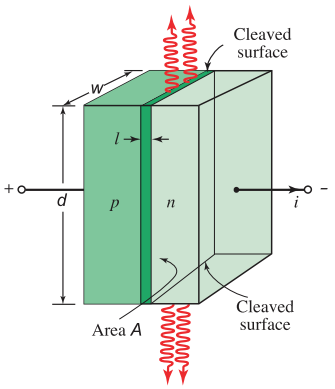


Figure 18.3-1 In its simplest configuration, a laser diode is a forward-biased p–n junction in which two surfaces that are perpendicular to the plane of the junction act as reflectors. These surfaces are often cleaved along crystal planes to ensure that they are parallel. The other two surfaces perpendicular to the plane of the junction are often roughened to eliminate feedback.

injected into a p–n junction. However, the light emitted from an LED is generated by spontaneous emission, whereas the light from an LD arises from stimulated emission.

Laser diodes enjoy a number of advantages with respect to other types of lasers: small size, compatibility with electronic components, high power, high efficiency, as well as ease of pumping and modulation by electric-current injection. Laser diodes have manifold uses, as will be considered subsequently.

We begin our discussion of the conditions required for laser oscillation, and the characteristics of the emitted light, with a brief summary of the basic results that describe the semiconductor optical amplifier and the Fabry–Perot optical resonator.

Laser Amplification

The gain coefficient $\gamma_0(\nu)$ of a semiconductor optical amplifier has a peak value γ_p that is approximately proportional to the injected-carrier concentration, which in turn is proportional to the injected current density J. Thus, as provided in (18.2-9) and (18.2-10), and as illustrated in Fig. 18.2-7,

$$\gamma_p \approx \alpha \left(\frac{J}{J_{\mathrm{T}}} - 1 \right), \quad J_{\mathrm{T}} = \frac{e\,l}{\eta_i \tau_{\mathrm{r}}} \Delta \mathrm{n}_{\mathrm{T}}, \tag{18.3-1}$$

where τ_{r} is the radiative electron–hole recombination lifetime, $\eta_i = \tau/\tau_{\mathrm{r}}$ is the internal quantum efficiency, l is the thickness of the active region, α is the thermal-equilibrium absorption coefficient, and $\Delta \mathrm{n}_{\mathrm{T}}$ and J_{T} are the injected-carrier concentration and current density required to just make the semiconductor transparent.

Feedback

The feedback is often obtained by cleaving the crystal planes normal to the plane of the junction, or by polishing two parallel surfaces of the crystal. The active region of the p–n junction illustrated in Fig. 18.3-1 then also serves as a planar-mirror optical resonator of length d and cross-sectional area lw. Semiconductor materials typically have large refractive indices, so that the power reflectance for normal incidence at the semiconductor–air interface,

$$\mathcal{R} = \left(\frac{n-1}{n+1} \right)^2, \tag{18.3-2}$$

is substantial [see (6.2-15) and Table 17.2-1]. Thus, if the gain of the medium is sufficiently large, the refractive-index discontinuity can itself serve as an adequate reflective surface and no external mirrors are necessary. For GaAs, for example, $n = 3.6$, so that (18.3-2) yields $\mathcal{R} = 0.32$.

Resonator Losses

The principal source of loss in the Fabry–Perot resonator arises from the partial reflection at the surface of the crystal. This loss constitutes the transmitted useful laser light. For a resonator of length d, the reflection loss coefficient is [see (11.1-22)]

$$\alpha_m = \alpha_{m1} + \alpha_{m2} = \frac{1}{2d} \ln\left(\frac{1}{\mathcal{R}_1 \mathcal{R}_2}\right); \tag{18.3-3}$$

if the two surfaces have the same reflectances $\mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R}$, then $\alpha_m = (1/d)\ln(1/\mathcal{R})$. The total loss coefficient is

$$\alpha_r = \alpha_s + \alpha_m, \tag{18.3-4}$$

where α_s represents other sources of loss, including free-carrier absorption in the semiconductor material (Fig. 17.2-2) and scattering from optical inhomogeneities. The quantity α_s increases as the concentration of impurities and interfacial imperfections in heterostructures increase. It can attain values in the range 10 to 100 cm^{-1}.

Of course, the term $-\alpha$ in the expression for the gain coefficient (18.3-1), corresponding to absorption in the material, also contributes substantially to the losses. This contribution is accounted for, however, in the net peak gain coefficient γ_p given by (18.3-1). This is apparent from the expression for $\gamma_0(\nu)$ given in (17.2-24), which is proportional to $f_g(\nu) = f_e(\nu) - f_a(\nu)$ (i.e., to stimulated emission less absorption).

Another important contribution to the loss results from the spread of optical energy outside the active layer of the amplifier (in the direction perpendicular to the junction plane). This can be especially detrimental if the thickness of the active layer l is small. The light then propagates through a thin amplifying layer (the active region) surrounded by a lossy medium so that large losses are likely. This problem may be alleviated by the use of a double heterostructure (Sec. 18.2C and Fig. 18.2-8), in which the middle layer is fabricated from a material of elevated refractive index that acts as a waveguide confining the optical energy.

Losses caused by optical spread may be phenomenologically accounted for by defining a **confinement factor** Γ to represent the fraction of the optical energy lying within the active region, as illustrated in Fig. 18.3-2.

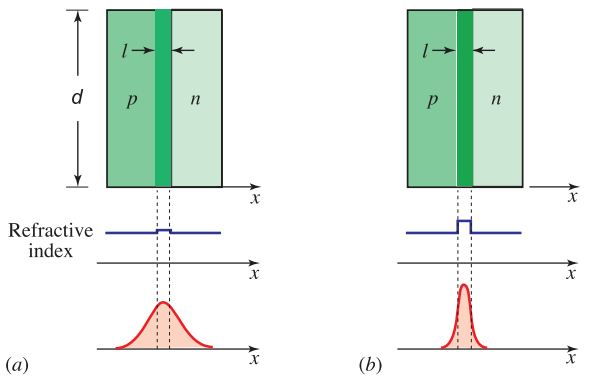


Figure 18.3-2 Spatial spread of the laser light in the direction perpendicular to the plane of the junction for: (*a*) homostructure, and (*b*) heterostructure lasers.

Assuming that the energy outside the active region is totally wasted, Γ is therefore the factor by which the gain coefficient is reduced, or equivalently, the factor by which the

loss coefficient is increased. Equation (18.3-4) must therefore be modified to reflect this increase, so that

$$\alpha_r = \frac{1}{\Gamma}(\alpha_s + \alpha_m). \tag{18.3-5}$$

There are three types of simple laser-diode structures based on the mechanism used to confine the carriers or light in the transverse (or lateral) direction (i.e., in the junction plane): **broad-area** (in which there is no mechanism for transverse confinement), **gain-guided** (in which transverse variations of the gain are used for confinement), and **index-guided** (in which transverse refractive-index variations are used for confinement).

Gain Condition: Laser Threshold

The laser oscillation condition is that the gain exceed the loss, $\gamma_p > \alpha_r$, as indicated in (16.1-12). The threshold gain coefficient is therefore α_r. Setting $\gamma_p = \alpha_r$ and $J = J_{\rm t}$ in (18.3-1) corresponds to a threshold injected current density $J_{\rm t}$ given by

$$J_{\rm t} = \frac{\alpha_r + \alpha}{\alpha} J_{\rm T}, \tag{18.3-6}$$

Threshold Current Density

where the transparency current density,

$$J_{\rm T} = \frac{e\,l}{\eta_i \tau_{\rm r}} \Delta n_{\rm T}, \tag{18.3-7}$$

Transparency Current Density

is the current density that just makes the medium transparent. The threshold current density is larger than the transparency current density by the factor $(\alpha_r + \alpha)/\alpha$, which is ≈ 1 when $\alpha \gg \alpha_r$. Since the current $i = JA$, where $A = wd$ is the cross-sectional area of the active region, we can define $i_{\rm T} = J_{\rm T}A$ and $i_{\rm t} = J_{\rm t}A$, corresponding to the currents required to achieve transparency of the medium and laser-oscillation threshold, respectively.

The threshold current density $J_{\rm t}$ is a key parameter in characterizing the laser-diode performance; smaller values of $J_{\rm t}$ indicate superior performance. In accordance with (18.3-6) and (18.3-7), $J_{\rm t}$ is minimized by maximizing the internal quantum efficiency η_i; and by minimizing the resonator loss coefficient α_r, the transparency injected-carrier concentration $\Delta n_{\rm T}$, and the active-region thickness l. As l is reduced beyond a certain point, however, the loss coefficient α_r becomes larger because the confinement factor Γ decreases [see (18.3-5)]. Consequently, $J_{\rm t}$ decreases with decreasing l until it reaches a minimum value, beyond which any further reduction causes $J_{\rm t}$ to increase (see Fig. 18.3-3). In double-heterostructure lasers, however, the confinement factor remains near unity for lower values of l because the active layer behaves as an optical waveguide (see Fig. 18.3-2). The result is a lower minimum value of $J_{\rm t}$, as shown in Fig. 18.3-3, and therefore superior performance. The reduction in $J_{\rm t}$ is illustrated in the following examples.

Because the parameters $\Delta n_{\rm T}$ and α in (18.3-1) are temperature dependent, so too are the threshold current density $J_{\rm t}$ and the frequency of peak gain. Temperature control can be used to stabilize the laser output and to modify the output frequency.

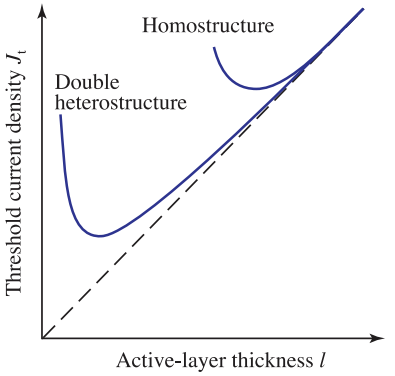


Figure 18.3-3 Dependence of the threshold current density J_t on the thickness of the active layer l. The double-heterostructure laser exhibits a lower value of J_t than the homostructure laser, and therefore superior performance. The increase of J_t at small values of l is a result of the reduction in confinement for thin active layers.

EXAMPLE 18.3-1. ***Threshold Current for an InGaAsP Homostructure Laser Diode.*** Consider an InGaAsP homostructure Fabry–Perot laser diode with the same material parameters as in Examples 18.2-1 and 18.2-2: $\Delta n_T = 1.25 \times 10^{18}$ cm^{-3}, $\alpha = 600$ cm^{-1}, $\tau_r = 2.5$ ns, $n = 3.5$, and $\eta_i = 0.5$ at $T = 300°$ K. Assume that the dimensions of the junction are $d = 200$ μm, $w = 10$ μm, and $l = 2$ μm. The current density necessary for transparency is then calculated to be $J_T = 3.2 \times 10^4$ A/cm^2. We now determine the threshold current density for laser oscillation. Using (18.3-2), the surface reflectance is $\mathcal{R} = 0.31$. The corresponding mirror loss coefficient is $\alpha_m = (1/d)\ln(1/\mathcal{R}) = 59$ cm^{-1}. Assuming that the loss coefficient due to other effects is also $\alpha_s = 59$ cm^{-1} and that the confinement factor $\Gamma \approx 1$, the total loss coefficient is then $\alpha_r = 118$ cm^{-1}. The threshold current density is therefore $J_t = [(\alpha_r + \alpha)/\alpha]\, J_T = [(118 + 600)/600][3.2 \times 10^4] = 3.8 \times 10^4$ A/cm^2. The corresponding threshold current $i_t = J_t wd \approx 760$ mA, which is rather high. Homostructure lasers are rarely used because of the difficulties of achieving CW operation without cooling to dissipate heat.

EXAMPLE 18.3-2. ***Threshold Current for an InGaAsP Heterostructure Laser Diode.*** We turn now to an InGaAsP/InP double-heterostructure Fabry–Perot laser diode (see Fig. 18.2-8) with the same parameters and dimensions as in Example 18.3-1 except for the active-layer thickness, which is now taken to be $l = 0.1$ μm instead of 2 μm. If the confinement of light is assumed to be perfect ($\Gamma = 1$), we may use the same values for the resonator loss coefficient α_r. The transparency current density is then reduced by a factor of 20 to become $J_T = 1600$ A/cm^2, and the threshold current density assumes a more reasonable value of $J_t = 1915$ A/cm^2. The corresponding threshold current is $i_t = 38$ mA. It is this significant reduction in threshold current that made CW operation of the double-heterostructure laser diode feasible at room temperature.

B. Power and Efficiency

Internal Photon Flux

When the laser current density is increased above its threshold value (i.e., $J > J_t$), the amplifier peak gain coefficient γ_p exceeds the loss coefficient α_r. Stimulated emission then outweighs absorption and other resonator losses so that oscillation can begin and the photon flux Φ in the resonator can increase. As with other homogeneously broadened lasers, saturation sets in as the photon flux becomes larger and the population difference becomes depleted [see (16.1-2)]. As shown in Fig. 16.2-1, the gain coefficient then decreases until it becomes equal to the loss coefficient, whereupon steady state is reached.

As with the internal photon-flux density and the internal photon-number density considered for other types of lasers [see (16.2-2) and (16.2-13)], the steady-state internal photon flux Φ is proportional to the difference between the pumping rate R and the

threshold pumping rate R_t. Since $R \propto i$ and $R_t \propto i_t$, in accordance with (18.2-8), Φ may be written as

$$\Phi = \begin{cases} \dfrac{\eta_i i_t}{e}\left(\dfrac{i}{i_t} - 1\right) = \eta_i \dfrac{i - i_t}{e}, & i > i_t \\ 0, & i \le i_t. \end{cases} \tag{18.3-8}$$

Steady-State Internal Photon Flux

Thus, the steady-state laser internal photon flux (photons/s generated within the active region) is equal to the electron flux (injected electrons/s) in excess of that required for threshold, multiplied by the internal quantum efficiency η_i.

The internal laser power above threshold is simply related to the internal photon flux Φ by the relation $P = h\nu\Phi$, so that we obtain

$$P \approx \eta_i (i - i_t)\frac{1.24}{\lambda_o}, \tag{18.3-9}$$

Internal Laser Power
λ_o (μm), P (W), i (A)

where λ_o is expressed in μm, i in amperes, and P in watts.

Output Photon Flux and Efficiency

The laser output photon flux Φ_o is the product of the internal photon flux Φ and the **extraction efficiency** η_e [see (16.2-16)], which is the ratio of the loss associated with the useful light transmitted through the mirrors to the total resonator loss α_r. If only the light transmitted through mirror 1 is used, then $\eta_e = \alpha_{m1}/\alpha_r$; on the other hand, if the light transmitted through both mirrors is used, then $\eta_e = \alpha_m/\alpha_r$. In the latter case, if both mirrors have the same reflectance $\mathcal{R}$, we obtain $\eta_e = [(1/d)\ln(1/\mathcal{R})]/\alpha_r$. The laser output photon flux is therefore given by

$$\Phi_o = \eta_e \eta_i \frac{i - i_t}{e}. \tag{18.3-10}$$

Laser Output Photon Flux

The proportionality between the laser output photon flux and the injected electron flux above threshold set forth in (18.3-10) is governed by a quantity known as the **external differential quantum efficiency**,

$$\eta_d = \eta_e \eta_i. \tag{18.3-11}$$

External Differential Quantum Efficiency

The quantity η_d thus represents the rate of change of the output photon flux with respect to the injected electron flux above threshold,

$$\eta_d = \frac{d\Phi_o}{d(i/e)}. \tag{18.3-12}$$

It is related to the differential power-conversion efficiency (slope efficiency) set forth in (16.2-20) via

$$\eta_d = \frac{eV}{h\nu}\,\eta_s\,. \tag{18.3-13}$$

When $h\nu \approx eV$, as is usually the case, we have $\eta_d \approx \eta_s$.

The laser output power above threshold is $P_o = h\nu\Phi_o = \eta_d(i - i_{\rm t})(h\nu/e)$, which is written more simply as

$$\boxed{P_o \approx \eta_d(i - i_{\rm t})\frac{1.24}{\lambda_o}\,,} \tag{18.3-14}$$

Laser Output Power
λ_o (μm), P_o (W), i (A)

when λ_o is expressed in μm. This relationship is called the **light–current (L–i) curve**. The slope of this curve above threshold is known as the **differential responsivity** of the laser, which is usually specified in units of W/A:

$$\mathsf{R}_d = \frac{dP_o}{di} \approx \eta_d\frac{1.24}{\lambda_o}\,. \qquad [\lambda_o\ (\mu\text{m}),\ P_o\ (\text{W}),\ i\ (\text{A})] \tag{18.3-15}$$

Light–current curves for two laser diodes are displayed as the solid curves in Fig. 18.3-4: (a) a gain-guided MQW InGaAsP/InGaAsP device operating at 1550 nm; and (b) a MQW GaN/InGaN device operating at 405 nm. The theoretical fits provided by (18.3-14) are shown as dashed curves.

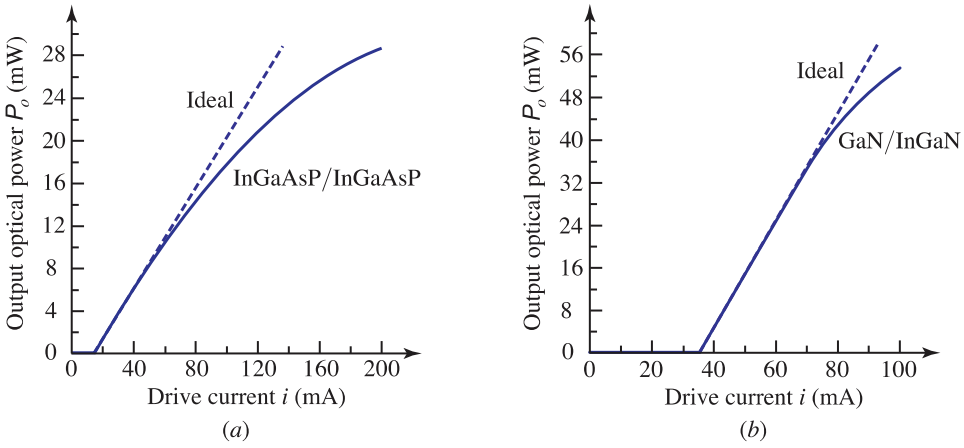


Figure 18.3-4 Measured (solid) and ideal (dashed) light–current curves for: (a) a gain-guided MQW InGaAsP/InGaAsP laser diode operated at a wavelength of 1550 nm in the near infrared (the device structure is exhibited in Fig. 18.4-4); (b) a MQW GaN/InGaN laser diode operated at a wavelength of 405 nm in the violet. Nonlinearities, which are not accounted for by the simple theory, cause the optical output power to saturate.

The parameters associated with these laser diodes are readily extracted by making use of (18.3-14) and (18.3-15); their values are presented in Table 18.3-1. Though the external differential quantum efficiency η_d is nearly identical for both devices, the differential responsivity R_d is about a factor of four greater for the GaN/InGaN device by virtue of its shorter operating wavelength, as is readily understood from (18.3-15).

Table 18.3-1 MQW laser-diode operating parameters extracted from the infrared and violet light–current curves displayed in Figs. 18.3-4(a) and (b), respectively.

Material	λ_o (nm)	$i_{\rm t}$ (mA)	η_d	R_d (W/A)
InGaAsP/InGaAsP	1550	15	0.33	0.26
GaN/InGaN	405	35	0.33	1.0

The **power-conversion efficiency** (or wall-plug efficiency) η_c is defined as the ratio of the emitted laser light power P_o to the electrical input power $P_{\rm e} = iV$, where V is the forward-bias voltage applied to the diode. Since $P_o = \eta_d(i - i_{\rm t})(h\nu/e)$, we have

$$\eta_c = \eta_d \left(1 - \frac{i_{\rm t}}{i}\right) \frac{h\nu}{eV}. \tag{18.3-16}$$

Power-Conversion Efficiency

For operation well above threshold, so that $i \gg i_{\rm t}$, and for $eV \approx h\nu$, we obtain $\eta_c \approx \eta_d$. Laser diodes can exhibit power-conversion efficiencies in excess of 70%, which is well above that for LEDs (Table 18.1-1) and for other types of lasers (Table 16.3-1). The electrical power that is not transformed into light is transformed into heat. Because laser diodes do, in fact, generate substantial amounts of heat they are usually mounted on heat sinks, which help to dissipate the heat and stabilize the temperature.

EXAMPLE 18.3-3. ***Comparison of Efficiencies for Multiquantum-Well and Double-Heterostructure InGaAsP Laser Diodes.*** Consider once again Example 18.3-2 for the InGaAsP/InP double-heterostructure Fabry–Perot laser diode with $\eta_i = 0.5$, $\alpha_m = 59\ \mathrm{cm}^{-1}$, $\alpha_r = 118\ \mathrm{cm}^{-1}$, and $i_{\rm t} = 38$ mA. If the light from both output faces is used, the extraction efficiency is $\eta_e = \alpha_m/\alpha_r = 0.5$, while the external differential quantum efficiency is $\eta_d = \eta_e\eta_i = 0.25$. At $\lambda_o = 1300$ nm, the differential responsivity of this laser is $\mathrm{R}_d = dP_o/di = 0.24$ W/A. If, for example, $i = 50$ mA, we have $i - i_{\rm t} = 12$ mA and $P_o = 12 \times 0.24 = 2.9$ mW. Comparison of these numbers with those reported in Fig. 18.3-4(a) and Table 18.3-1 for a MQW InGaAsP/InGaAsP laser diode operated at 1550 nm reveals that the MQW laser has a lower threshold current and a higher external differential quantum efficiency than the double-heterostructure laser, as expected.

Summary

There are four efficiencies associated with laser diodes:

- The internal quantum efficiency $\eta_i = \mathrm{r}_{\rm r}/\mathrm{r} = \tau/\tau_{\rm r}$, which accounts for the fact that only a fraction of the electron–hole recombinations are radiative.
- The extraction efficiency η_e, which accounts for the fact that only a portion of the light lost from the cavity is useful.
- The external differential quantum efficiency $\eta_d = \eta_e\eta_i$, which accounts for both of the above effects.
- The power-conversion (wall-plug) efficiency η_c, which is the ratio of the emitted optical power to the electrical power supplied to the device.

The differential responsivity R_d (W/A) is also a useful measure of performance.

Comparison of LED, SLED, and LD Efficiencies and Powers

It is of interest to compare the efficiencies and optical powers associated with LEDs, SLEDs, and LDs. When operated below threshold, laser diodes produce spontaneous emission and behave as light-emitting diodes (Sec. 18.1). Indeed, the presence of spontaneous emission can be discerned at low currents in LD light–current curves.

The four efficiencies attendant to LD operation have been highlighted in the summary above. There are also four efficiencies associated with LEDs, as discussed in Sec. 18.1. These are the internal quantum efficiency η_i, which accounts for the fact that only a fraction of the electron–hole recombinations are radiative in nature; the transmittance efficiency η_e, which accounts for the fact that only a fraction of the light generated in the junction region can escape from the high-index semiconductor medium; the external efficiency $\eta_{\text{ex}} = \eta_i \eta_e$, which accounts for both of the foregoing effects; and the power-conversion efficiency η_c. The responsivity R is also used as a measure of LED performance.

There are one-to-one correspondences between the quantities η_i, η_e, and η_c for the LED and the LD, respectively. Furthermore, there are correspondences between η_{ex} and η_d, R and R_d, and i and $(i - i_{\text{t}})$ for the LED and the LD, respectively. The superior performance of the laser results principally from the fact that η_e for the LD is greater than that for the LED. This is because the laser operates on the basis of stimulated emission, which causes the laser light to be concentrated in particular modes so that it can be more readily extracted. The net result is that a laser diode operated above threshold has a value of η_d that is typically larger than the value of η_{ex} for an LED.

Superluminescent laser diodes (SLEDs), which are operated with current injection that is sufficiently strong so that stimulated emission dominates spontaneous emission, exhibit behavior intermediate between that of LEDs and LDs. As discussed in Sec. 18.2E, feedback must be frustrated in these devices to avert lasing.

To forge a comparison for the performance of these three classes of devices, light–current curves for a light-emitting diode, superluminescent diode, and laser diode are provided in Fig. 18.3-5, at modest values of the drive current. All are MQW InGaAsP/InP structures operating at a wavelength of 1600 nm. The responsivity and efficiency of the LD are substantially greater than those of the other two devices. Moreover, the light–current curve of the SLED characteristically bends upward in exponential fashion (as is apparent in the inset) whereas the LED and LD curves bend downward at higher current levels as a result of saturation.

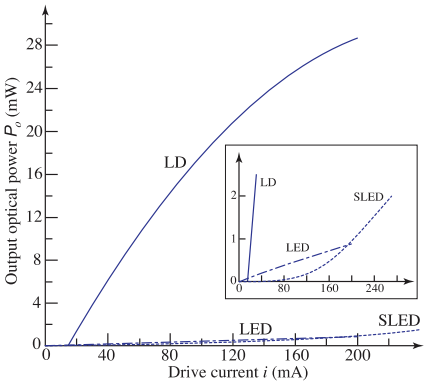


Figure 18.3-5 Light–current curves for a light-emitting diode (LED), a superluminescent diode (SLED), and a laser diode (LD). All three devices are InGaAsP/InP MQW structures operated at a wavelength of 1600 nm, and at modest values of the drive current. The inset provides an expanded view of the LED and SLED curves.

C. Spectral and Spatial Characteristics

Spectral Characteristics

The spectral intensity of laser light is governed by three factors, as described in Sec. 16.2B:

1. The bandwidth B over which the active medium small-signal gain coefficient $\gamma_0(\nu)$ is greater than the loss coefficient α_r.
2. The homogeneous or inhomogeneous nature of the line-broadening mechanism (Sec. 14.3D).
3. The resonator modes, in particular the approximate frequency spacing between the longitudinal modes $\nu_F = c/2d$, where d is the resonator length.

The spectral intensity of light emitted by a semiconductor laser diode, in particular, is characterized by the following three features:

1. The spectral width of the gain coefficient is relatively large because transitions occur between two energy bands rather than between two discrete energy levels.
2. Intraband processes are fast so that semiconductors tend to be homogeneously broadened. Nevertheless, spatial hole burning permits the simultaneous oscillation of many longitudinal modes (Sec. 16.2B). Spatial hole burning is particularly prevalent in short cavities in which there are few standing-wave cycles. This permits the fields of different longitudinal modes, which are distributed along the resonator axis, to overlap less, thereby allowing partial spatial hole burning to occur.
3. The semiconductor resonator length d is significantly smaller than that of most other types of lasers. The frequency spacing of adjacent resonator modes $\nu_F = c/2d$ is therefore relatively large. Nevertheless, many such modes can generally be supported within the broad bandwidth B over which the small-signal gain exceeds the loss [the number of possible laser modes is $M = B/\nu_F$, in accordance with (16.2-23)].

EXAMPLE 18.3-4. ***Number of Longitudinal Modes in an InGaAsP Laser Diode.*** An InGaAsP crystal ($n = 3.5$) of length $d = 400$ μm has Fabry–Perot resonator modes spaced by $\nu_F = c/2d = c_o/2nd \approx 107$ GHz. Near the central wavelength $\lambda_o = 1300$ nm, this frequency spacing corresponds to a free-space wavelength spacing λ_F, where $\lambda_F/\lambda_o = \nu_F/\nu$, so that $\lambda_F = \lambda_o\nu_F/\nu = \lambda_o^2/2nd \approx 0.6$ nm. If the spectral width $B = 1.2$ THz (corresponding to a wavelength width $\Delta\lambda = 7$ nm), then approximately 11 longitudinal modes may oscillate. A typical spectral-intensity pattern consisting of a single transverse mode and about 11 longitudinal modes is illustrated in Fig. 18.3-6. The linewidth of individual longitudinal modes is typically of the order of tens of MHz for index-guided lasers and a few GHz for gain-guided lasers. The overall spectral width of light emitted by laser diodes is greater than that of most other lasers (see Table 15.3-1). To reduce the number of modes to one within the confines of a Fabry–Perot structure, the resonator length d would have to be reduced so that $B = c/2d$, requiring a cavity of length $d \approx 36$ μm.

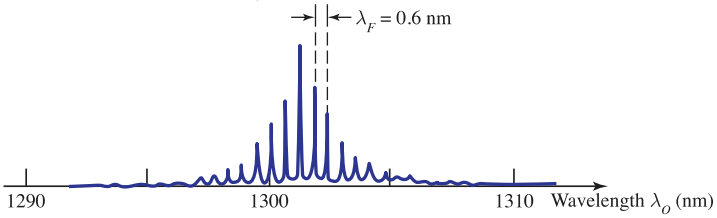


Figure 18.3-6 Spectral intensity of a 1300-nm InGaAsP index-guided buried-heterostructure laser. This distribution is considerably narrower, and differs in shape, from that of a $\lambda_o \approx 1300$-nm InGaAsP LED (see Fig. P18.1-5). The number of modes decreases as the injection current increases; the mode closest to the gain maximum increases in power while the side peaks saturate. (Adapted from R. J. Nelson, R. B. Wilson, P. D. Wright, P. A. Barnes, and N. K. Dutta, CW Electrooptical Properties of InGaAsP ($\lambda = 1.3$ μm) Buried-Heterostructure Lasers, *IEEE Journal of Quantum Electronics*, vol. QE-17, pp. 202–207, Fig. 6 ©1981 IEEE.)

Comparison of LED, SLED, and LD Spectral Intensities

The spectral intensities for an InGaAsP/InP light-emitting diode, a superluminescent diode, and a laser diode are compared in Fig. 18.3-7. The spectral narrowing associated with stimulation emission is evident in the SLED curve, and even more so in the LD curve.

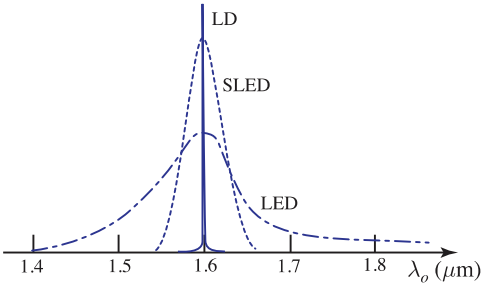


Figure 18.3-7 Normalized spectral intensities for a light-emitting diode (LED), a superluminescent diode (SLED), and a laser diode (LD). All three devices are InGaAsP/InP structures operating at a wavelength of 1600 nm. The LED has a broad spectrum, the LD has a narrow spectrum, and the SLED has a spectrum of intermediate width.

Spatial Characteristics

As with other Fabry–Perot lasers, oscillation in laser diodes takes the form of transverse and longitudinal modes. In Sec. 16.2C, the indices (l, m) were used to characterize the spatial distributions in the transverse direction, while the index q was used to represent variation along the direction of wave propagation or temporal behavior. In most other types of lasers, the laser beam resides totally within the active medium so that the spatial distributions of the different modes are determined by the shapes of the mirrors and their separations. For circularly symmetric systems, the transverse modes can be represented in terms of Hermite–Gaussian or Laguerre–Gaussian beams (Sec. 11.2D). However, the situation is different in laser diodes since the laser beam extends outside the active layer. The transverse modes (also called spatial modes) are therefore modes of the dielectric waveguide created by the different layers of the laser diode.

The transverse modes can be determined by using the theory presented in Sec. 9.3 for an optical waveguide with rectangular cross section of dimensions l and w. If l/λ_o is sufficiently small, the waveguide will admit only a single mode in the transverse direction perpendicular to the junction plane. However, w is usually larger than λ_o, so that the waveguide will support several modes in the direction parallel to the plane of the junction, as illustrated in Fig. 18.3-8. Modes in the direction parallel to the junction plane are called **transverse modes** or **lateral modes**. The larger the ratio w/λ_o, the greater the number of transverse modes possible.

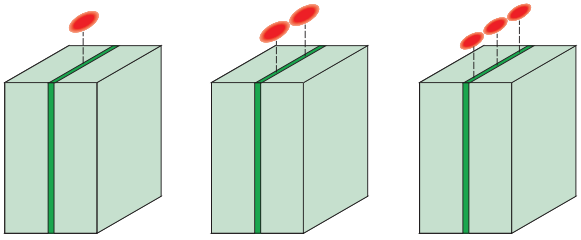


Figure 18.3-8 Schematic illustration of optical-intensity spatial distributions for the laser waveguide modes $(l, m) = (1, 1)$, $(1, 2)$, and $(1, 3)$.

Far-Field Radiation Pattern

An edge-emitting laser diode with an active layer of dimensions l and w emits light with far-field angular divergence $\approx \lambda_o/l$ (radians) in the plane perpendicular to the junction and $\approx \lambda_o/w$ in the plane parallel to the junction, as illustrated in Fig. 18.3-9. This is similar to the results for a Gaussian beam of diameter $2W_0$, provided in (3.1-21), for which the divergence angle is $\theta \approx (2/\pi)(\lambda_o/2W_0) = \lambda_o/\pi W_0$ when $\theta \ll 1$. The angular divergence determines the far-field radiation pattern, as discussed in Sec. 4.3. Because of the small size of its active layer, the laser diode is characterized by an angular divergence larger than that of most other lasers. As an example, for $l = 2$ μm, $w = 10$ μm, and $\lambda_o = 800$ nm, the divergence angles are calculated to be $\approx 23°$ and $5°$. Light from a single-transverse-mode laser diode, for which w is smaller, has an even larger angular divergence. The spatial distribution of the far-field light within the radiation cone depends on the number of transverse modes and on their optical powers. The highly asymmetric elliptical distribution of laser-diode light emitted from such a device can make collimation tricky.

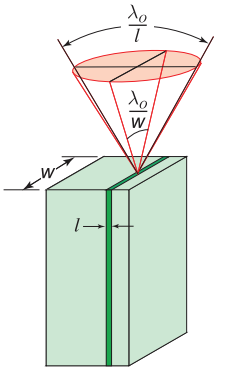


Figure 18.3-9 Angular distribution of the optical beam emitted from an edge-emitting laser diode. The directions perpendicular and parallel to the plane of the junction are called the fast axis and slow axis, respectively (not to be confused with designations of the same name that relate to polarization).

Single-Mode Operation: DBR and DFB Laser Diodes

Because higher-order *transverse* modes have a wider spatial spread, they are less confined; their loss coefficient α_r is therefore greater than that for lower-order modes. Consequently, some of the highest-order modes will fail to satisfy the oscillation conditions; others will oscillate at a lower power than the fundamental (lowest-order) mode. To achieve high-power single-spatial-mode operation, the number of waveguide modes must be reduced. This can be done by decreasing the dimensions of the active-layer cross section (l and w) so that it acts as a single-mode waveguide. The attendant reduction of the junction area also has the effect of reducing the threshold current. Higher-order transverse modes may also be eliminated by making use of gain-guided or index-guided laser-diode configurations.

Operation on a single *longitudinal* mode, which produces a single-frequency output, may be achieved by reducing the length d of the resonator so that the frequency spacing between adjacent longitudinal modes, i.e., the Fabry–Perot free spectral range $\nu_F = c/2d$, exceeds the spectral width of the amplifying medium. Single-mode operation may also be attained by making use of multiple-mirror resonators, as discussed in Sec. 16.2D and illustrated in Fig. 16.2-15.

Another approach for achieving single-frequency operation involves the use of distributed reflectors in place of the cleaved crystal surfaces that serve as lumped mirrors in the Fabry–Perot configuration. When feedback of this type is provided, the surfaces of the crystal are antireflection coated to suppress the Fabry–Perot modes. As an example, wavelength-selective reflectors such as Bragg gratings can be placed in the plane

of the junction [Fig. 18.3-10(a)]. As discussed in Secs. 2.4B and 7.1C, a Bragg grating reflects light when the grating period Λ satisfies $\Lambda = q\lambda/2$, where q is an integer. The device portrayed in Fig. 18.3-10(a) is called a **distributed Bragg reflector laser** or, more simply, a **DBR laser**. Alternatively, a DBR grating placed below or above the active region can also serve as a distributed reflector, as illustrated in Fig. 18.3-10(b).

Yet another method for providing wavelength-dependent feedback makes use of a corrugated region between the active and guiding layers, as shown in Fig. 18.3-10(c); this results in a periodic refractive index and therefore a grating. Structures of this kind are known as **distributed-feedback lasers** or, for short, **DFB lasers**. This class of lasers offers single-frequency operation that is robust in the presence of drive-current variations, operating-temperature variations, and the presence of modulation. Moreover, DFB lasers offer narrow spectral widths, large modulation bandwidths, and low noise (the DFB configuration avoids partition noise arising from competition among longitudinal modes in Fabry–Perot lasers). Consequently, DFB lasers are widely used, including as sources for optical fiber communication systems in the 1300–1600-nm wavelength range (Sec. 18.4A).

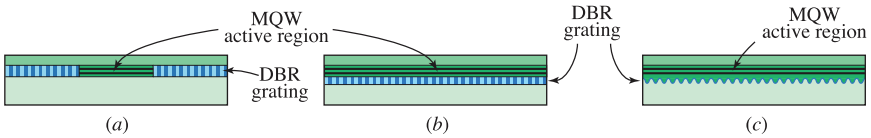


Figure 18.3-10 (a) Schematic diagram of a distributed Bragg reflector (DBR) multiquantum-well laser diode with DBR mirrors outside the active region. (b) Diagram of a distributed feedback (DFB) multiquantum-well laser diode with a DBR structure that resides below the active region and serves as a distributed reflector. (c) Structure for a distributed feedback (DFB) multiquantum-well laser diode, which incorporates a corrugation between the active and guiding layers that acts as a distributed reflector.

Linewidth-enhancement factor. The linewidth $\Delta\nu_{\mathrm{L}}$ of the light emitted by a single-mode semiconductor laser exceeds the Schawlow–Townes linewidth $\Delta\nu_{\mathrm{ST}}$, which accommodates the random-phase spontaneous-emission contributions that combine with the laser-oscillation mode (Sec. 16.2D). The increased linewidth is a result of the coupling between phase and intensity changes that arises from the effects of the semiconductor carrier density on its refractive index. This coupling is represented by a linewidth-enhancement factor α, which is the constant of proportionality between changes in the phase and the field gain. A typical MQW DFB laser has a linewidth-enhancement factor $\alpha = 4.5$ at an output power of 10 mW; quantum-dot devices have smaller values of α. The laser-diode linewidth is a factor of $(1 + \alpha)^2$ larger than the conventional laser linewidth, so that the Schawlow–Townes linewidth for a single-mode laser diode is given by $\Delta\nu_{\mathrm{ST}} = \pi(1 + \alpha)^2 h\nu\,(\delta\nu)^2/P_o$.

External-Cavity Wavelength-Tunable Laser Diodes

There are many circumstances in which it is advantageous to be able to tune the output wavelength of a single-mode laser diode. One prominent example is in a coherent optical communication system, whose operation requires a tunable local oscillator (Sec. 25.4). Other examples include wavelength-division multiplexed (WDM) systems (Sec. 25.3C), systems involving wavelength conversion, and spectroscopic applications. The wavelength at which a LD operates can be changed, for example, by modifying the refractive index of the active medium. This can be implemented via various physical mechanisms, such as carrier injection, the application of an electric field, or temperature modification.

However, it is far more convenient to achieve tuning by placing the die in an external cavity that incorporates a wavelength-selective element (see Sec. 16.2D). Aside from allowing the output wavelength to be tuned, this approach concomitantly yields a salutary reduction in its spectral width. In the Littman–Metcalf configuration illustrated in Fig. 18.3-11, the die has a highly reflective coating on one of its ends and an antireflection coating on the other. A collimating lens and an external mirror complete the cavity, into which is inserted a wavelength-selective element, usually a stationary diffraction grating. The output wavelength may then be tuned over the spectral width B where net gain is available by rotating the mirror that reflects the first-order diffracted beam back to the laser diode. A particular merit of the Littman–Metcalf configuration is that the direction of the output beam remains fixed as the wavelength is tuned. An analogous fiber-optic configuration incorporates a fiber Bragg grating (FBG).

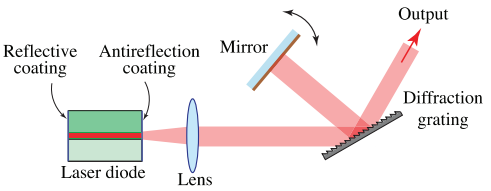


Figure 18.3-11 Littman–Metcalf configuration for a tunable external-cavity laser diode. The output wavelength is tuned by rotating the mirror.

External-cavity laser diodes are readily mode-locked since a saturable absorber (Sec. 15.4A) can be easily inserted in the cavity to achieve passive mode locking. Mode-locked external-cavity laser diodes offer advantages over mode-locked fiber lasers in certain applications (e.g., optical fiber communications).

18.4 QUANTUM-CONFINED LASERS

Quantum-confined lasers, in which carriers are confined to dimensions smaller than the de Broglie wavelength of a thermalized electron ($\lambda_{dB} \approx 50$ nm in GaAs), are the workhorses of the family of laser diodes. Confinement of the electron momentum in 0, 1, 2, and 3 dimensions corresponds to bulk, quantum-well, quantum-wire, and quantum-dot configurations, respectively; the *geometrical dimensionality* of these structures are 3, 2, 1, and 0, respectively, as depicted in Fig. 18.4-1. Bulk structures have confinement in 0 dimensions and thus have a geometrical dimensionality of 3. Quantum-dot structures have confinement in 3 dimensions and thus have a geometrical dimensionality of 0. Convention dictates that quantum-confined structures be designated by their geometrical dimensionality. Some of the elementary properties of quantum-confined structures were set forth in Secs. 14.1D and 17.1G.

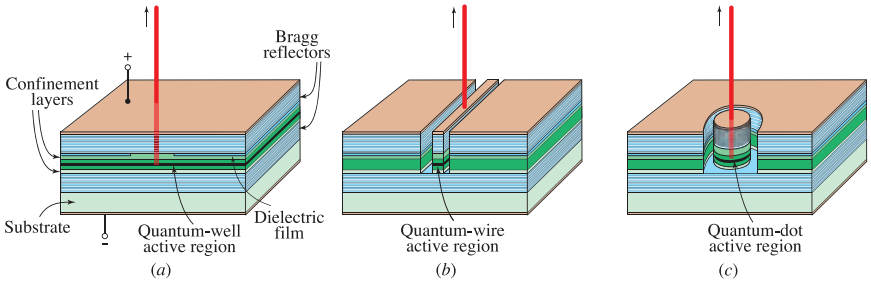


Figure 18.4-1 Schematic representation of several quantum-confined laser configurations: (*a*) quantum-well laser (2D); (*b*) quantum-wire laser (1D); and (*c*) quantum-dot laser (0D). Charge carriers are restricted to the active region by confinement layers while Bragg reflectors serve as mirrors.

The dimensionality of a quantum-confined device governs the behavior of the laser gain coefficient, threshold current, external differential quantum efficiency, and output linewidth. In general, a decrease in the geometrical dimensionality leads to a reduction of the active volume, and thus to a reduction of the output power, especially for quantum-wire and quantum-dot lasers. In this section, we discuss quantum-well, quantum-wire, and quantum-dot semiconductor lasers in turn. We then turn to quantum cascade lasers, which are multiquantum-well devices that generate substantial optical power in the infrared and THz spectral regions.

A. Quantum-Well and Multiquantum-Well Lasers

We have already encountered several examples of quantum-well and multiquantum-well structures earlier, in connection with LEDs, SOAs, SLEDs, and LDs. As discussed in Secs. 18.2 and 18.3, the performance of the single-quantum-well (SQW) device portrayed in Fig. 18.4-1(a) is superior to that of the double-heterostructure (DH) device. The benefit accrues from the small thickness of a single quantum well, which is typically < 10 nm; this is to be compared with an active-region thickness of ≈ 100 nm for a DH laser diode and ≈ 2 μm for an old-fashioned homojunction device.

The dependences of the peak gain coefficient γ_p on the current density J for SQW and bulk DH laser diodes are compared in Fig. 18.4-2. The SQW laser requires a far smaller value of the current density J_{T} to achieve transparency. Its peak gain coefficient increases sharply at first but then saturates at multiples of the maximum gain coefficient γ_m [see (18.2-13)].

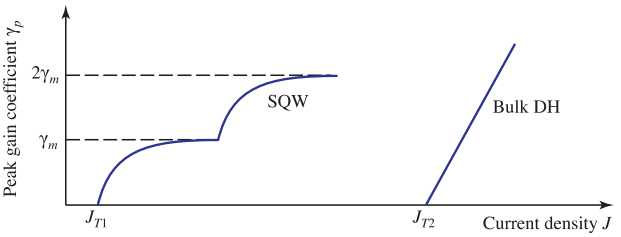


Figure 18.4-2 Peak gain coefficient γ_p versus current density J for SQW and bulk DH laser diodes. The quantum-well laser has a far smaller transparency current density J_{T}; however, its gain coefficient saturates at a lower level.

The single-quantum-well laser offers the following salutary features in comparison with its double-heterostructure counterpart:

- Smaller threshold current density
- Larger external differential quantum efficiency
- Larger power-conversion efficiency
- Narrower gain-coefficient width
- Smaller laser-mode linewidth
- Reduced temperature dependence
- Faster response and thus greater modulation frequencies

The multiquantum-well (MQW) laser, schematized in Fig. 18.4-3, offers a greater gain coefficient than the single-quantum-well laser. Indeed, the gain coefficient of a MQW laser with N wells is N times that of each of its wells. Multiquantum-well lasers offer excellent performance and are in fact the most commonly used of all laser diodes. They find extensive use in all manner of applications, and comprise the lion's share of the device structures discussed in this section and in Sec. 18.5.

However, to effect a fair comparison of the performance of SQW and MQW devices, the pumping level should be taken to be the same in both. Consider a single quantum well injected with an excess carrier density $\Delta \mathrm{n}$ and a peak gain coefficient γ_p. In the

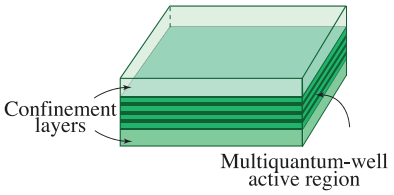


Figure 18.4-3 Schematic of the active region of a multiquantum-well laser. The confinement layers restrict charge carriers to the quantum-well region.

comparison MQW structure, each of the N wells would then be injected with only $\Delta n/N$ carriers. Because of the nonlinear dependence of the gain on Δn, the gain coefficient of each well would then be $\xi\gamma_p/N$, where ξ could be smaller or greater than unity, depending on the operating conditions. The total gain provided by the MQW laser would thus be $N(\xi\gamma_p/N) = \xi\gamma_p$. It turns out that the performance of the MQW device is typically inferior at low current densities but superior at high current densities, but by a factor smaller than N.

Strained-Layer Quantum-Well Lasers

The introduction of mechanical strain can provide a salutary effect on the performance of laser diodes, in spite of the fact that the notion is counterintuitive. **Strained-layer lasers** are widely used because of their superior properties. Quantum-confined strained-layer lasers have been fabricated in many material systems (including III–V, III–nitride, and SiGe) using various configurations, and operated at many wavelengths. Rather than being lattice-matched to the confining layers, the active region is deliberately chosen to have a different lattice constant. If sufficiently thin, it can accommodate its atomic spacings to those of the surrounding layers, and in the process become mechanically strained. If the active region is too thick, however, it will not properly accommodate and the material will develop defects and imperfections that render it unusable. The InGaAs active layer in an AlGaAs/InGaAs strained-layer quantum-well laser, for example, has a free-standing lattice constant that is significantly greater than that of its AlGaAs confining layers. The thin InGaAs layer is therefore subjected to biaxial compression in the plane of the layer, while its atomic spacings are increased above their nominal values in the direction perpendicular to the layer. Conversely, an active layer with a free-standing lattice constant smaller than that of the confining layers would be subjected to biaxial tension in the plane of the layer and would experience decreased atomic spacings in the perpendicular direction.

Compressive strain can alter the band structure in three significant ways: (1) it increases the bandgap E_g; (2) it removes the degeneracy at $k = 0$ between the heavy and light hole bands; and (3) it makes the valence bands anisotropic so that the highest band has a light effective mass in the direction parallel to the plane of the layer and a heavy effective mass in the perpendicular direction. These features can serve to significantly improve the performance of quantum-well lasers. First, the laser wavelength is altered by virtue of the dependence of E_g on the strain. Second, the laser threshold current density can be reduced by the presence of the strain, which may be understood in terms of the following argument: achieving a population inversion requires that the separation of the quasi-Fermi levels be greater than the bandgap energy, i.e., $E_{fc} - E_{fv} > E_g$, as set forth in (17.2-12); the reduced hole mass allows E_{fv} to more readily descend into the valence band, thereby permitting this condition to be satisfied at lower values of injection current. Indeed, we have already seen that strain can have an outsize effect on the performance of a photonic device; the Ge laser can operate only in its presence (Example 17.2-1).

Materials and Device Structures

Most semiconductor lasers in use today make use of active regions that comprise quantum-confined structures. We begin by considering several types of conventional multiquantum-well laser diodes; these lasers find use in a whole host of applications ranging from consumer products such as laser printers and data-storage devices to long-haul optical fiber communication systems. They also serve as highly efficient optical pumps for optical fiber amplifiers, fiber lasers, and solid-state lasers. The materials and device structures for most conventional laser diodes closely resemble those used for light-emitting diodes (Sec. 18.1C). Direct-bandgap ternary and quaternary materials are used in the near-infrared to mid-ultraviolet region because their bandgap wavelengths can be compositionally tuned. As with LEDs, AlInGaN, AlInGaP, InGaAs, and InGaAsP are particularly important quaternary materials.

Laser diodes are commonly available at the following wavelengths: 635, 650, 680, and 780 nm for use in laser pointers, optical storage and display systems, and short-haul plastic-fiber communication systems; 850 nm for short-haul silica-fiber communication systems; and 1300–1600 nm for long-haul silica-fiber communication systems. Typical wavelengths used for diode-pumped solid-state (DPSS) and diode-pumped fiber lasers are 793 nm (AlGaAs) for thulium-doped silica fiber; 808 and 880 nm (AlGaAs) for neodymium-doped yttrium vanadate and YAG; 940 nm (InGaAs) for ytterbium-doped YAG and silica fiber; and 980 nm (InGaAs) for erbium-doped silica-fiber lasers and amplifiers. Other wavelengths at which laser diodes are commonly available include 375, 405, 440, 670, 785, 830, and 920 nm. Lead-salt laser diodes can operate at wavelengths as long as about 30 μm, but they have been largely supplanted by mid-infrared quantum cascade lasers.

Individual edge-emitting laser diodes can deliver optical powers that range from milliwatts to tens of watts, with power-conversion efficiencies that can exceed 70%, speeds in the vicinity of tens of GHz, and life spans of years.

Single-mode MQW lasers. Lasers that include a narrow waveguide, in the form of a ridge or buried index step with a width between 2 and 5 μm, can accommodate only a single spatial mode (Sec. 9.3). Such lasers are used when it is important to be able to focus the laser beam to a diffraction-limited spot; applications include optical data storage, printing, metrology, and optical fiber communications. The small waveguide size in such devices limits the optical power to a maximum of about 1 W. When single spatial mode MQW lasers have a Fabry–Perot resonator configuration, as for the ridge-waveguide laser illustrated below, lasing generally occurs on multiple longitudinal modes.

Operation on a single longitudinal-mode can be instituted by making use of a distributed-feedback (DFB) configuration in place of the Fabry–Perot feedback. The former provides feedback only at a single frequency, and DFB lasers have many other salutary characteristics as well (Sec. 18.3C). The signal and pump sources for optical fiber communication systems require the narrow frequency spectrum and low noise offered by lasers that operate on a single spatial mode as well as a single longitudinal mode. Lasers for this purpose generally provide milliwatts to watts of optical power and are fiber-coupled. The buried-heterostructure DFB laser diode illustrated below is a highly reliable device that fills the bill.

Ridge-waveguide Fabry–Perot laser. The ridge-waveguide (RW) laser diode operates on a single spatial mode and can lase over a broad range of wavelengths. The ridge waveguide provides weak optical waveguiding as well as gain guiding by restricting current injection to the active region beneath the ridge. RW laser diodes often take the form of a Fabry–Perot structure with cleaved facets. The 500-μm-long device displayed in Fig. 18.4-4 has an active region comprising six 7-nm-thick, compression-strained InGaAsP quantum wells sandwiched between 10-nm-thick tension-strained

InGaAsP barriers. The laser diode depicted here has a threshold current $i_t = 15$ mA, an external differential quantum efficiency $\eta_d = 0.33$, a differential responsivity $R_d = 0.26$ W/A, and emits 20 mW of optical power.

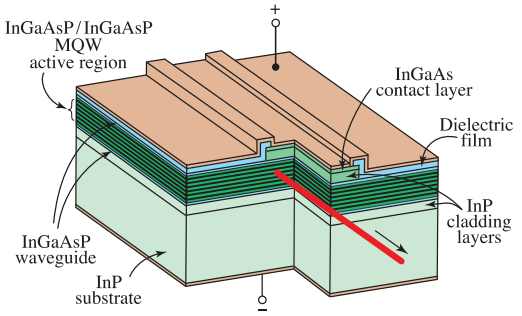


Figure 18.4-4 Schematic diagram of a strained-MQW InGaAsP/InGaAsP ridge-waveguide laser diode that operates at 1550 nm. The active region under the ridge is surrounded on all sides by material of lower refractive index so that the ridge constitutes an optical waveguide. This laser operates on a single spatial mode but on multiple longitudinal modes. The light–current curve for this device is displayed in Fig. 18.3-4(a).

Buried-heterostructure distributed-feedback laser. As illustrated in Fig. 18.4-5, alternating p- and n-type layers allow current flow only in the vicinity of the active region in this buried-heterostructure device, thereby enforcing lateral confinement. The dielectric film provides gain guiding. The distributed feedback (DFB) component of the device makes use of a corrugated-layer grating adjacent to the active region that serves as a distributed reflector (Sec. 18.3C). Lasers such as these offer ample gain at modest current levels, and can provide output powers as high as 1 W in a single spatial and longitudinal mode. These devices offer narrow spectral widths, which is crucial for the efficient operation of 1300–1600-nm wavelength-division-multiplexed (WDM) communication systems, as discussed in Sec. 25.1B. Typical values of the threshold current and differential responsivity are $i_t < 10$ mA and $R_d \approx 0.4$ W/A, respectively, and $\Delta\nu_L$ is a few MHz.

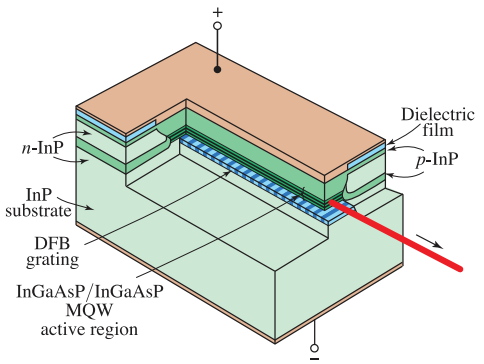


Figure 18.4-5 Buried-heterostructure multiquantum-well DFB laser used for optical fiber communications in the 1300–1600-nm wavelength range. This laser operates on a single spatial mode as well as on a single longitudinal mode.

Multimode MQW lasers. If the width of the active region of a laser diode is broadened from a few μm to, say, 200 μm, it can operate on multiple spatial modes and deliver optical powers as high as 10 W. Devices of this kind are known as **broad-area laser diodes** or **broad-stripe laser diodes**. The light from such lasers cannot be focused to a diffraction-limited spot, nor can it be efficiently coupled into a single-mode fiber, but it is suitable for applications such as pumping optical fiber amplifiers, multi-clad fiber lasers, and diode-pumped solid-state (DPSS) lasers.

Laser-diode bars and stacks. Yet higher laser powers may be obtained by configuring multimode laser diodes into bars and stacks, which can serve as pumps for diode-pumped solid-state (DPSS) lasers (Sec. 16.3A) as well as for purposes such as direct diode-laser materials processing. The usual bar comprises between 10 and 50 broad-area laser diodes, contiguously arranged in a 1D array and integrated into a single chip. A bar suitable for pumping a DPSS laser typically has a length of 1 cm, emits 100 W of partially coherent optical power, and has a power-conversion efficiency in the vicinity of 50% (Example 16.3-1). Bars are often mounted in stacks, which commensurately ramp up the optical power to kW levels. Powerful stacks developed at the Lawrence Livermore National Laboratory (LLNL) serve as pumps for the Nd^{3+}:glass laser amplifiers in the HAPLS petawatt laser system. As detailed in Example 23.2-3, an individual laser-diode stack containing more than 125 000 AlGaAs laser diodes delivers 250-J pulses of 0.3-ms duration at a wavelength of 888 nm, with a peak power of 800 kW and an average power of 2.5 kW, at a repetition rate of 10 pulses/s. The stack consists of an array of 1600 bars, each of which emits 500 W, with a bar-to-bar spacing of 350 μm. The power-conversion efficiency of the stack is $\eta_c \approx 60\%$.

B. Quantum-Wire and Multiquantum-Wire Lasers

Quantum wires (see Sec. 17.1G) can also serve as the active region of a semiconductor laser, as illustrated in Fig. 18.4-1(b). **Multiquantum-wire lasers** comprise arrays of quantum wires, as portrayed in Fig. 18.4-6. In principle, multiquantum-wire lasers offer narrower linewidths than quantum-well lasers by virtue of their tighter carrier confinement. However, the fabrication of III–V quantum-wire structures currently lags behind that of quantum-well structures, in part because of the difficulty of creating a sufficiently dense collection of wires, and hence so too does their performance.

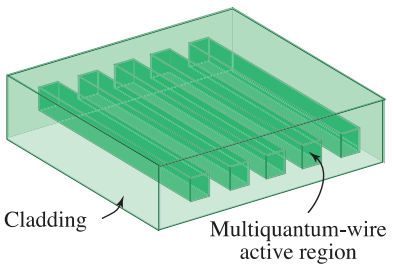


Figure 18.4-6 Schematic of the active region of a multiquantum-wire laser. Light is ordinarily emitted in all directions; laser emission can be restricted to the end faces by making use of a suitable resonator.

EXAMPLE 18.4-1. ***Performance Comparison of Multiquantum-Wire and Quantum-Well Lasers.*** A collection of five 1-mm-long, 23-nm-wide, InGaAsP active-layer quantum wires, clad with InP and spaced 80 nm apart, operates as a room-temperature CW multiquantum-wire laser at a wavelength $\lambda_o \approx 1550$ nm. The threshold current, threshold current density, external differential quantum efficiency, and power-conversion efficiency turn out to be $i_t = 140$ mA, $J_t = 800$ A/cm^2, $\eta_d = 40\%$, and $\eta_c = 2\%$, respectively.[†] As a result of the small volume of the active region and the substantial optical losses, however, the performance of this multiquantum-wire laser is inferior to that of a quantum-well laser fabricated from the same chip, which has operating parameters $i_t = 100$ mA, $J_t = 500$ A/cm^2, $\eta_d = 50\%$, and $\eta_c = 6\%$.

[†] See H. Yagi, T. Sano, K. Ohira, D. Plumwongrot, T. Maruyama, A. Haque, S. Tamura, and S. Arai, GaInAsP/InP Partially Strain-Compensated Multiple-Quantum-Wire Lasers Fabricated by Dry Etching and Regrowth Processes, *Japanese Journal of Applied Physics*, vol. 43, pp. 3401–3409, 2004.

C. Quantum-Dot and Multiquantum-Dot Lasers

Quantum dots, occasionally called **quantum boxes** or **nanocrystals**, are semiconductor particles that can take the form of cubes, spheres, disks, pyramids, or other shapes. They typically have dimensions in the range 1–50 nm (a 10-nm cube of GaAs contains some 40 000 atoms). The carriers may be confined by cladding the dots with a semiconductor of larger bandgap or by embedding them in glass or polymer. Figure 18.4-1(c) depicts a quantum-dot. The growth and characteristics of quantum dots were discussed in Sec. 14.1D and their energy levels were examined in Sec. 17.1G. The energy levels of a quantum dot are those of its excitons. Though the levels are sharp as a result of tight carrier confinement, they depend strongly on the size of the dot. As illustrated in Fig. 14.1-13, the photoluminescence photon energy increases as the dot size decreases because of the greater energy required to confine the semiconductor excitation to a smaller volume (see Sec. 14.1D).

Since quantum dots can self-assemble into ordered arrangements, it is easier than it might appear to construct a multiquantum-dot laser with an active region containing many quantum dots, as depicted in Fig. 18.4-7. The first such device, fabricated with InGaAs quantum dots, was operated in 1994.

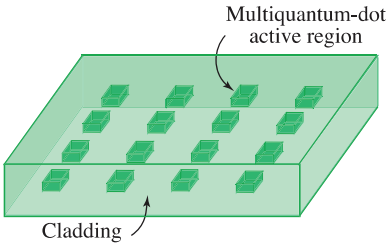


Figure 18.4-7 Schematic of the active region of a multiquantum-dot laser, which often consists of multiple layers, each containing self-assembled multiple quantum dots as shown. The typical dimensions of self-assembled quantum dots fall in the 10–50-nm range.

For collections of quantum dots, the delta-function density of states associated with an isolated dot is usually smeared into a smooth profile whose character is determined by the inhomogeneities in quantum-dot sizes and shapes. The resulting inhomogeneous broadening of the quantum-dot laser has the distinct merit in that it offers wavelength tunability. This contrasts with the nearly homogeneously broadened gain medium of the quantum-well laser.

GaAs- and InP-based multiquantum-dot lasers, which serve the 1.3 and 1.55-μm wavelength regions, respectively, offer a number of performance advantages over their quantum-well counterparts:

- Ultralow CW threshold current density at room temperature, $J_t \approx 10$ A/cm^2 per quantum-dot layer; this is a factor of 5 lower than its quantum-well counterpart.
- Superior gain and differential quantum efficiency.
- Reduced dependence of threshold current density on temperature; uncooled lasing is available from a ground-state transition for temperatures as high as 200° C, which eliminates the need for external cooling.
- Reduced sensitivity to defects facilitates the integration of III–V lasers with group-IV materials such as Si and Ge.
- Reduced linewidth-enhancement factor and reduced linewidth; the linewidth–power product of a quantum-dot DFB laser is ≈ 1 MHz-mW at an output power of 2 mW — this is an order of magnitude lower than that of a commercial quantum-well DFB laser operated at the same power.
- Increased operating bandwidth to 200 nm; wavelength-tunable operation is available in an external-cavity configuration (Sec. 18.3C).
- Increased modulation bandwidth for direct modulation (currently $\approx$ 20 GHz at room temperature).

- Insensitivity to optical feedback for light that inadvertently reenters the cavity, thereby avoiding the need for an isolator in an optical fiber communication system.
- Availability of mode-locked operation with short pulse widths and high repetition rates by virtue of broad bandwidth, fast gain dynamics, easily saturated gain and absorption, and low linewidth-enhancement factor.

The small size and low power consumption of quantum-dot lasers, along with their ability to operate uncooled at high temperatures, as well as the salutary features outlined above, make them especially suitable for use in many specialized applications. These include sensing in hot environments as well as serving in optical clock distribution and high bit-rate optical time-division multiplexing systems (Chapter 24).

D. Quantum Cascade Lasers

Most semiconductor lasers operate via radiative electron–hole recombination. The production of light in such **interband lasers** is a two-carrier, single-photon affair: a transition comprising the combination of an electron from the conduction band with a hole in the valence band generates a single photon via stimulated emission. The **quantum cascade laser** (**QCL**), in contrast, makes use of only a single kind of carrier, the electron, which makes multiple transitions and generates multiple photons via stimulated emission. The QCL is therefore a *unipolar* rather than a *bipolar* device. QCLs are constructed from a concatenated series of quantum wells, designed and biased in such a way that a single electron injected into the conduction band of the device undergoes a cascade of intersubband stimulated-emission transitions as it transits the device (Fig. 18.4-8). The operating wavelength of the device is thus unrelated to the bandgaps of the constituent semiconductor materials. Rather, the wavelength is determined by the widths of the quantum wells and barriers, which in turn determine the subband and miniband energy separations (Exercise 17.1-5). With hundreds to thousands of individual layers, the QCL is, perhaps, the epitome of **band-structure engineering**.

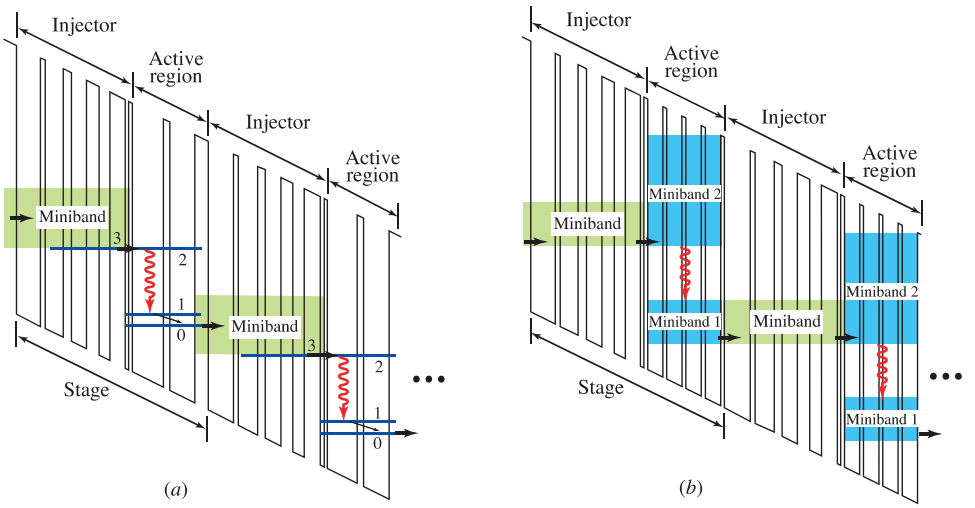


Figure 18.4-8 Schematic diagram of: (*a*) two stages of a QCL with a quantum-well active region, and (*b*) two stages of a QCL with a superlattice active region. QCLs usually contain between 10 and 100 stages, comprising hundreds to thousands of individual semiconductor layers.

Quantum cascade lasers can be fabricated with either quantum-well or superlattice active regions. As illustrated in Fig. 18.4-8(*a*), the quantum-well version consists of

a sequence of stages, each comprising an electron injector and a quantum-well active region. The injector contains a collection of wells of varying widths and thin barriers that form a superlattice, with an energy-level structure consisting of minibands separated by minigaps (Sec. 14.1D). In the presence of bias, the electrons are injected via resonant tunneling from the bottom (ground state) of a miniband, denoted level 3, into the upper laser level in the quantum-well active region, denoted level 2. A photon of frequency $\nu = E_{21}/h$ is emitted via stimulated emission on the $2 \rightarrow 1$ intersubband transition, as indicated in red in Fig. 18.4-8 (see also Sec. 17.2D). The electron then decays via phonon scattering to level 0, whereupon it enters the miniband in the next stage via resonant tunneling. The process is then repeated, resulting in the emission of another photon. A typical QCL contains between 10 and 100 stages, so that a substantial number of photons are generated for each electron that transits the device.

Because it makes use of transitions within a single band, the operation of a quantum-well QCL bears some resemblance to the operation of an electrically pumped gas laser. As is evident in Fig. 18.4-8(a), level 2 is not aligned with a miniband of the succeeding stage, so that it has a relatively long lifetime ($\tau_2 \approx 1$ ps) and therefore accumulates population. Level 1, in contrast, does not sustain population since decay to level 0 takes place via a fast nonradiative transition and subsequent tunneling into the succeeding stage ($\tau_1 \approx 0.1$ ps). At each stage of the QCL, therefore, the electron follows a path similar to that for a four-level laser system in which a population inversion is maintained on the $2 \rightarrow 1$ transition (Sec. 15.2B). As a result, the gain coefficient $\gamma_0(\nu)$ for a QCL is proportional to a narrow lineshape function $g(\nu)$, as for an atomic laser [see (16.1-1)], rather than to a broad joint density-of-states function $\varrho(\nu)$, as for an interband semiconductor laser [see (18.2-4)]. The QCL transition linewidth arises from lifetime broadening, intersubband scattering, and nonparabolicity of the subbands.

The superlattice QCL shown in Fig. 18.4-8(b) differs from the quantum-well version schematized in Fig. 18.4-8(a) in that stimulated emission takes place between the bottom and top of two minibands in the active region which, in this case, comprises a superlattice (Sec. 17.2D). The laser frequency is thus determined by the height of the minigap separating the two minibands. This structure is generally more suitable for generating coherent light with wavelengths longer than about 10 μm since the alignment between the injector and active region is less critical. Furthermore, higher drive currents are less deleterious and a population inversion is more readily achieved in this configuration because of very fast relaxation in the lower laser-level miniband.

Another widely used design for the QCL active region is the so-called *bound-to-continuum scheme*, where the laser action involves transitions from a discrete upper state to a superlattice miniband. This approach combines the efficient electron injection into the upper laser level of a quantum-well QCL with the fast depopulation of the lower laser level of a superlattice QCL, and thereby reduces the threshold and increases the output power.

Many other QCL designs have been developed. These include superlattice devices in which the injector region is eliminated; **heterogeneous QCLs** that generate broadband radiation or supercontinuum emission by incorporating multiple cascades that operate at different wavelengths within the active region; devices in which the light is guided in surface-plasmon modes so that long-wavelength operation can be achieved without thick dielectric waveguides; and Raman-laser devices that are pumped by a quantum cascade laser integrated into the same structure. Surface-emitting and ring-cavity configurations are also employed, and QCL arrays are readily fabricated.

QCLs can operate at wavelengths that stretch from the near to the far infrared, and beyond to the THz region, all with the same heterostructure configuration. The shortest wavelength at which a QCL can operate is determined by the heterostructure conduction-band offset, which governs the largest available photon energy for intersubband transitions — large conduction-band offsets allow lasing at shorter wavelengths. Single-mode operation can be achieved by incorporating a distributed-feedback (DFB)

element into the device. External-cavity feedback, in conjunction with a rotatable grating (Sec. 18.3C), offers single-frequency operation with high spectral purity and wavelength tuning over a range of about 10% of the center wavelength. Fine wavelength tuning extending over a range of about 1% of the center wavelength may be achieved by changing the injection current and/or temperature, which modify the effective refractive index of the material; this in turn changes the optical pathlength of the cavity and thus the emission wavelength. Sampled-grating DFB (SGDFB) QCLs, and arrays thereof, extend this tunability substantially by incorporating a pair of reflection gratings and imposing slightly different periodic spatial modulations on each. This in turn gives rise to two sequences of periodic reflection maxima in the wavelength domain with slightly different spacings. Single-frequency operation obtains since oscillation can only occur when two reflection maxima align, as illustrated in Figs. 16.2-14 and 16.2-15 for multiple-mirror resonators. There are several variations on this theme. Frequency-comb operation in the mid infrared is also readily achieved (Sec. 16.4E).

Mid-Infrared Quantum Cascade Lasers

Interest in the mid-infrared region of the electromagnetic spectrum was initially fostered by the two regions of atmospheric transparency that lie within it: the 3–5-μm and 8–14-μm windows, which are also known as the medium-wavelength infrared (MWIR) and long-wavelength infrared (LWIR) bands, respectively. Quantum cascade lasers operate over a range of wavelengths that encompasses the mid infrared ($2 \leq \lambda_o \leq 20$ μm), as well as portions of the far-infrared ($20 \leq \lambda_o \leq 300$ μm) and THz ($100 \leq \lambda_o \leq 1000$ μm) spectral regions (see Fig. 2.0-1).

This broad range of wavelengths offers extensive opportunities for scientific, industrial, and military applications, including infrared spectroscopy, infrared imaging and countermeasures, combustion diagnostics, rangefinding, and free-space optical communications. Moreover, since these bands encompass the *molecular-fingerprint region* that contains the vibrational–rotational transition wavelengths of many molecular species (Sec. 14.1C), the QCL has also engendered numerous remote-sensing applications, such as trace-gas analysis and sensing, chemical sensing and identification, and isotopic analysis. Together with room-temperature mid-infrared detectors, such as HgCdTe photovoltaic arrays and VOx microbolometer arrays (Sec. 19.5), QCLs provide unparalleled access to the mid infrared.

As discussed above, the operating wavelength of a QCL is determined by the widths of its quantum wells and barriers, which in turn establish the subband and miniband energy separations. In principle, QCLs can thus be constructed using a wide variety of semiconductor materials although the intersubband gain coefficient depends on the well and barrier effective masses, which in turn are governed by the choice of materials. Superior performance has been obtained by using MBE or MOCVD to fabricate the following material systems:

- InGaAs/InAlAs quantum wells on an InP substrate
- GaAs/AlGaAs quantum wells on a GaAs substrate
- InAs/AlSb quantum wells on an InAs substrate
- InGaAs/AlInAsSb, InGaAs/GaAsSb, or InGaAs/AlInGaAs on InP
- GaN/AlGaN quantum wells on a GaN substrate

As with interband lasers, strained-layer QCLs (e.g., compressively strained InGaAs and tensilely strained InAlAs) offer improved performance. QCLs are often fabricated in a buried-heterostructure configuration that contains between 10 and 100 stages, with overall lengths between $1/2$ and 10 mm and widths that range from 5 to 20 μm.

QCLs that operate CW at room temperature in the region $3 \leq \lambda_o \leq 25$ μm (which encompasses both the MWIR and LWIR bands) exhibit excellent performance. In particular, devices that operate CW at room temperature in the region $3 \leq \lambda_o \leq 12$ μm deliver output powers in excess of 5 W and power-conversion efficiencies greater than

25%. Moreover, such QCLs can be modulated at high rates and can be mode-locked to produce optical pulses of a few-picoseconds duration. Multi-wavelength QCLs can also incorporate integrated nonlinear mixing regions to foster parametric interactions such as difference-frequency and sum-frequency generation.

An important current frontier in QCL research is directed toward extending CW, room-temperature operation to wavelengths both shorter and longer than the mid-infrared, i.e., to the near-infrared and to the THz regions, respectively. Direct approaches to achieving this would make use of heterostructures with increased band offsets, and would mitigate the thermally activated relaxation between the upper and lower radiative states, respectively. QCLs have been operated in the frequency range $1.2 \leq \nu \leq 4.9$ THz, corresponding to the wavelength range $60 \leq \lambda_o \leq 250$ μm, but cryogenic cooling is required.

Comparison of QCLs with Other Mid-IR Sources

Finally, it is instructive to compare quantum cascade lasers with other commonly encountered sources of radiation in the mid infrared. As indicated below, QCLs are often superior because of a number of salutary features: 1) the ability to operate CW at room temperature, 2) access to a broad range of wavelengths, 3) multiwatt CW optical powers, 4) high power-conversion efficiencies, and 5) compact structure:

- Transition-ion-doped zinc chalcogenides such as Cr^{2+}:ZnS and Cr^{2+}:ZnSe (Sec. 16.3A) offer substantial optical power and are continuously tunable, but only over the limited wavelength range $1.9 \leq \lambda_o \leq 3.0$ μm.
- CO and CO_2 gas lasers (Sec. 16.3E) offer output powers many orders of magnitude greater than those attainable with QCLs but they are bulky and fragile, and suffer from limited ranges of accessible wavelengths.
- Difference-frequency generation (Sec. 22.2C) offers a broad range of accessible wavelengths but it is complex and challenging to implement in a CW configuration.
- Lead-salt (IV–VI) interband laser diodes, such as those fabricated from PbSnTe and PbSnSe (Sec. 17.1B), can be compositionally tuned over a broad range of wavelengths (4 to 30 μm), which encompasses much of the mid-infrared spectrum. However, these devices suffer from nonradiative recombination, low thermal conductivity, and small band offsets that necessitate cryogenic cooling for CW operation. Also, optical power levels are limited to the milliwatt range and power-conversion efficiencies are low.
- III–antimonide interband laser diodes (usually InGaAsSb/AlGaAsSb quantum wells on a GaSb substrate) offer CW, room-temperature operation with optical powers in excess of 1.5 W and power-conversion efficiencies greater than 15%, but only for wavelengths shorter than $\approx$ 2.2 μm. The use of *quinternary* AlInGaAsSb barriers allows operation to be extended to about 4 μm, but growing and using quinternary materials is a complex enterprise and output powers are limited to tens of milliwatts in any case.
- GaSb-based *interband cascade lasers* (ICLs) operate CW at room temperature and generate hundreds of milliwatts of optical power with power-conversion efficiencies of about 15%. However, ICLs operate only in the shorter wavelength reaches of the mid-infrared spectrum, namely in the range $3 \leq \lambda_o \leq 6$ μm.

18.5 MICROCAVITY LASERS

The quantum confinement considered in Sec. 18.4 relates to the confinement of *carriers* to a spatial region of the order of the *de Broglie wavelength* of an electron (for a thermalized electron in GaAs, $\lambda_{dB} \approx 50$ nm). The **microcavity lasers** considered in

this section, in contrast, involve the confinement of *photons* to a spatial region of the order of the *optical wavelength* ($\lambda_o \approx 1\ \mu\text{m} \gg \lambda_{\text{dB}}$). *Microresonators* are resonators in which one or more of the spatial dimensions is the size of a few wavelengths of light or smaller, $d \approx \lambda$. *Microcavities* are usually thought of as having small dimensions in all spatial directions; however, these two terms have come to be used interchangeably. Microcavity lasers are also called **microlasers**.

Photon confinement and carrier confinement are independent features of photonic devices. It is therefore possible to have a *microcavity* laser whose active region *is not subject* to quantum confinement (e.g., a microcavity containing a simple p–n homojunction active region), or a *large-resonator* laser whose active region *is subject* to quantum confinement (e.g., a quantum cascade laser). In practice, however, most microcavity lasers make use of quantum-confined structures for their active regions.

Microresonator lasers in which the light is confined to wavelength-size regions in various dimensions are exemplified by the micropillar, microdisk, and microsphere structures illustrated in Fig. 18.5-1. These, and other, microresonators have been described in Sec. 11.4.

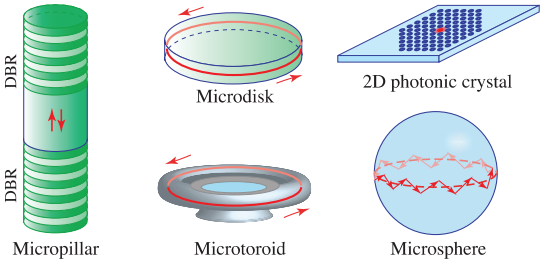


Figure 18.5-1 Microresonator (or microcavity) lasers, sometimes called microlasers for short, confine light within wavelength-size regions in various dimensions. The defect in the 2D photonic crystal creates a cavity that traps the light. Analogous quantum-confined structures are the quantum well, quantum wire, and quantum dot.

In laser diodes with large resonators ($d \gg \lambda$), the modes exhibit small spacings in all directions of k-space and the density of allowed resonance frequencies $M(\nu)$ can be determined via a continuous approximation (Sec. 11.3). The overall spontaneous emission probability density (s^{-1}) depends on the modal density $M(\nu)$ of the frequency space into which photons can be emitted, as specified by (14.3-11). In large-resonator lasers, as in free space, the modal density assumes the quadratic form $M(\nu) = 8\pi\nu^2/c^3$, in accordance with (11.3-10). This offers a large number of modes for spontaneous emission. After stimulated emission is initiated in a particular mode, however, spontaneous emission into modes other than the laser mode represents wasted energy. Indeed, for a conventional laser diode, the fraction of spontaneous emission that contributes to a given laser mode is generally very small. The current injected into a large-resonator laser at threshold is thus principally replenishing the wasted spontaneous emission rather than contributing to the stimulated emission.

However, the modal density $M(\nu)$ can be substantially reduced by making use of a microcavity, as discussed in Sec. 11.4. The allowed modes of microresonators can exhibit large spacings in one or more directions of k-space, so that modes can be absent over extended spectral bands. The reduction is most dramatic in microcavities that have large spacings in all directions of k-space, which results in a discrete collection of modes (Fig. 11.4-1). The opportunity to alter the modal environment is important in connection with spontaneous emission. Placing a source in this environment inhibits spontaneous emission since it is directed away from modes that do not exist and toward modes that are available. Moreover, the emission of light into particular modes of a high-Q, small-volume microcavity can be enhanced relative to emission into ordinary optical modes via the Purcell effect, as described in Sec. 14.3E. Microcavity lasers are designed to take maximum advantage of opportunities for both spontaneous-emission inhibition and enhancement.

Summary

Microcavity lasers offer a number of desirable features in comparison with their conventional counterparts:

- Reduced size
- Reduced laser threshold
- Reduced spectral width
- Reduced spatial width
- Increased efficiency

However, reduced size generally signifies reduced output power.

We consider three classes of microcavity lasers in turn in Secs. 18.5A, 18.5B, and 18.5C: vertical-cavity surface-emitting lasers (VCSELs), microdisk and microring lasers, and photonic-crystal lasers. Though microcavity lasers comprising semiconductor active media are most prevalent, gain media such as organic dyes, rare-earth-doped silica, and organic polymers are also used.

A. Vertical-Cavity Surface-Emitting Lasers

Vertical-cavity surface-emitting lasers (VCSELs) are designed so that the light emerges from the top face of a planar Fabry–Perot resonator, much like the surface-emitting LED displayed in Fig. 18.1-18. VCSELs, which can be realized using conventional or organic semiconductors, usually operate in the visible and near IR. They can be fabricated with a broad range of diameters, stretching from ≈ 1 μm, where they resemble micropillar lasers, to $\approx$ 1 mm. Small-area VCSELs have threshold currents in the μA region, output powers in the mW range, and power-conversion efficiencies in the vicinity of 70%. The output beams are circular and therefore easily coupled to optical fibers. VCSELs find use in a broad variety of applications that range from optical mice to short-haul optical fiber communications (Sec. 25.1B).

Large-area VCSEL. An example of a large-area VCSEL is shown in Fig. 18.5-2. This device has a multiquantum-well GaAs/InGaAs active region with an emission

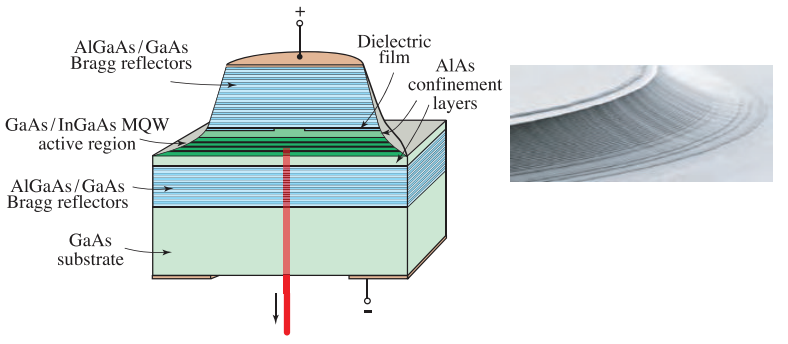


Figure 18.5-2 (a) Schematic diagram of a large-area (320-μm diameter) multiquantum-well GaAs/InGaAs VCSEL that operates at a wavelength of 995 nm. (b) Etched mesa showing the p contact, p-type DBR, and active region. (Adapted from M. Miller, M. Grabherr, R. King, R. Jäger, R. Michalzik, and K. J. Ebeling, Improved Output Performance of High-Power VCSELs, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 7, pp. 210–216, Fig. 2 ©2001 IEEE.)

wavelength centered at 995 nm. Because the thickness of the active region is only tens of nm, the single-pass gain is typically small (a fraction of 1%) and the light must be repeatedly reflected through the active region. Typical distributed Bragg reflector (DBR) mirrors contain dozens of layers to enhance reflectance at the operating wavelength. VCSELs often make use of dielectric films to localize carrier injection and thereby to laterally confine the optical mode. The spectral intensity, optical power, and angular emission distribution generated by the laser portrayed in Fig. 18.5-2 are shown in Fig. 18.5-3.

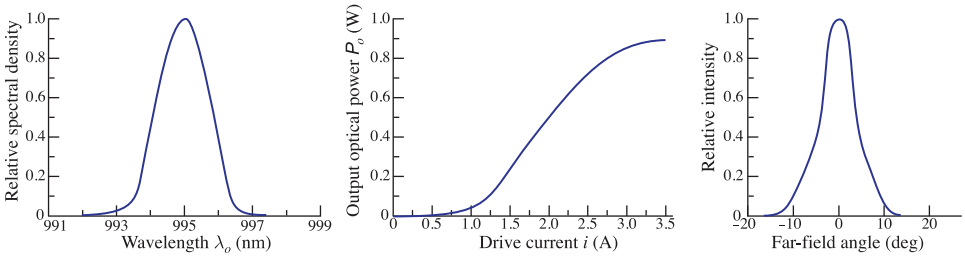


Figure 18.5-3 Spectral intensity, optical power, and angular emission distribution of the multiquantum-well GaAs/InGaAs VCSEL displayed in Fig. 18.5-2. The threshold current $i_t = 1.1$ A for this large-area device. The maximum optical power, about 1 W, is similar to that provided by the edge-emitting buried-heterostructure DFB laser depicted in Fig. 18.4-5. (Adapted from M. Miller, M. Grabherr, R. King, R. Jäger, R. Michalzik, and K. J. Ebeling, Improved Output Performance of High-Power VCSELs, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 7, pp. 210–216, Figs. 8, 5, and 9 ©2001 IEEE.)

Multiquantum-dot VCSEL. Though the active regions of VCSELs are usually multiquantum wells, they are also fabricated with multiquantum dots, as illustrated in Fig. 18.5-4. As with other structures that make use of quantum-dot emitters, the promise of reduced threshold and reduced temperature sensitivity, together with enhanced modulation bandwidth, is appealing.

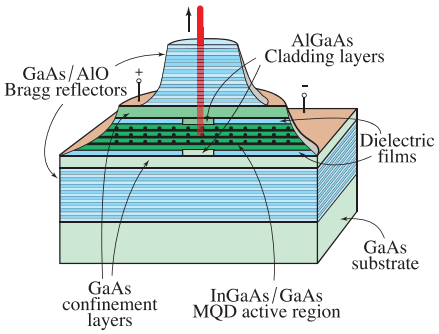


Figure 18.5-4 VCSEL with a multiquantum-dot active region.

Variations on the VCSEL theme. VCSELs assume an enormous variety of forms, and can incorporate auxiliary features such as photonic crystals for lateral mode control, coupled cavities, and integrated modulators that extend modulation bandwidths, as portrayed in Fig. 18.5-5.

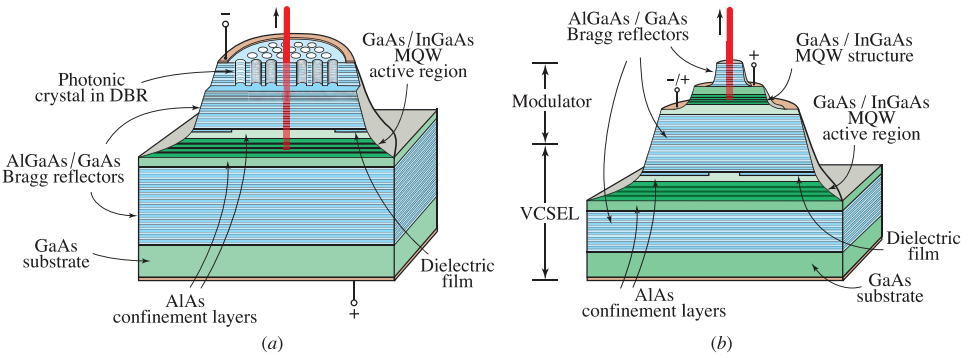


Figure 18.5-5 Variations on the theme of VCSELs. (*a*) VCSEL with photonic crystal for lateral mode control. (*b*) VCSEL with monolithically integrated electroabsorption modulator.

VCSEL Arrays

A salutary feature of VCSELs is that they offer high packing densities on a wafer scale and are readily fabricated in the form of dense arrays. As an early example, an array of about 1 million electrically pumped, tiny, vertical-cavity cylindrical InGaAs quantum-well VCSELs (diameter $\approx 2\ \mu$m, height $\approx 5.5\ \mu$m), with lasing wavelengths in the vicinity of 970 nm, was fabricated on a single 1-cm^2 chip of GaAs. These particular devices had thresholds $i_t \approx 100\ \mu$A, and operated CW at room temperature. A scanning electron micrograph of a tiny portion of this array is displayed in Fig. 18.5-6. VCSEL arrays can be fabricated with elements that have a prespecified distribution of diameters and laser frequencies.

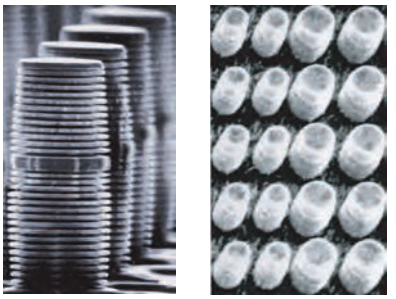


Figure 18.5-6 Scanning electron micrographs of an early array of electrically pumped vertical-cavity surface-emitting $In_{0.2}Ga_{0.8}As$ quantum-well lasers with diameters between 1 and 5 μm on a GaAs chip. The microresonators comprise AlAs/GaAs Bragg reflectors. (*a*) AlAs has been preferentially etched away from the Bragg reflectors in these devices, highlighting the GaAs disks (courtesy J. L. Jewell). (*b*) Top view of a small portion of the array. (Adapted from J. L. Jewell, A. Scherer, S. L. McCall, Y. H. Lee, S. Walker, J. P. Harbison, and L. T. Florez, Low-Threshold Electrically Pumped Vertical-Cavity Surface-Emitting Microlasers, *Optics News*, vol. 15, no. 12, pp. 10–11, 1989, Fig. 1.)

Two-dimensional arrays of incoherent emitters, as described above, offer broad-area power scaling. However, coherently combining the light emitted by a collection of individual VCSELs can, by virtue of the interferences among the individual emissions, achieve a divergence below that of a single emitter as well as increased on-axis intensity. Features such as these are important for applications such as imaging, targeting, sensing, and optical communications. Techniques for achieving coherent beam combination include **injection locking** and **lateral coupling** among neighboring incoherent emitters. In injection locking, part of the output from a single master laser, which may be separate from the array or one of the devices within it, is distributed and reflected back to seed all of the slave lasers, thereby causing them to oscillate coherently. Lateral coupling can be implemented via diffractive or antiguided (leaky-mode) interactions; evanescent interactions have proved difficult to marshal. Typical coherent power levels from arrays of hundreds of VCSELs are in the range of hundreds of mW.

Vertical External-Cavity Surface-Emitting Lasers (VECSELs)

As with external-cavity laser diodes (Sec. 18.3C), the VCSEL structure can be endowed with an external cavity to form a wavelength-tunable, single-mode **vertical external-cavity surface-emitting laser (VECSEL)**, sometimes called a **semiconductor disk laser (SDL)**. In one configuration, used in optical coherence tomography, tuning is provided by a movable microelectromechanical system (MEMS) that serves as the upper mirror. As with diode-pumped solid-state (DPSS) lasers (Sec. 16.4D), the use of an intracavity semiconductor saturable-absorber mirror (SESAM) can provide passive mode locking; VECSELs can attain pulse durations $\tau_{\text{pulse}} < 100$ fs and pulse repetition rates $1/T_F \approx 50$ GHz. Much as with VCSELs, VECSELs (along with QCLs and DFB lasers) can be fabricated in the form of arrays.

B. Microdisk and Microring Lasers

The march toward device miniaturization has facilitated the integration of lasers and other optical components in compact configurations and photonic integrated circuits (PICs). **Microdisk lasers** are configured with active regions comprising organic or conventional semiconductors, and most rely on quantum-well or quantum-dot structures. III–V and III–nitride devices operate over broad spectral ranges, stretching from the infrared to the ultraviolet. Because of their small size, low power consumption, low threshold current, ease of out-coupling, and CW room-temperature operation, dielectric microdisk lasers have experienced robust development since their invention in the early 1990s. They are often grown on their native substrates, but they can also be epitaxially grown on silicon — in some cases with equivalent performance. Electric current and mechanical support are provided by n- and p-type structures below and above the disk. Whereas VCSELs operate on the basis of Fabry–Perot modes and emit light vertically, microdisk lasers rely on high-Q whispering-gallery modes (Sec. 11.4B) and emit in the plane of the semiconductor substrate; light is extracted via in-plane coupling waveguides.

The modal structure of **microring lasers**, sometimes called microtoroid lasers, is quite similar to that of microdisk lasers of the same radius; indeed, a microring device can be considered to be a limiting case of a microdisk device in which the inner wall simply limits the number of modes in the radial direction. The overall dimensions of these devices typically lie in the multi-μm range, larger than the wavelength of the emitted light since they confine light by means of whispering-gallery modes. Applications of microdisk and microring lasers range from short-reach interconnects such as on-chip data communications to the detection of viruses and nanoparticles.

Silicon Photonics and Group-IV Devices

In the domains of silicon photonics and group-IV photonics, respectively, microcavity lasers can be fabricated via the direct heteroepitaxial growth of III–V quantum dots on Si and by the growth of GeSn on Si.

III–V quantum-dot-on-Si microring laser. Microring lasers containing active regions that consist of seven InAs/InGaAs quantum-dot-in-well (DWELL) layers have been epitaxially grown on planar or V-groove-patterned Si substrates.[†] The ring outer radii stretch from 5 to 50 μm and the ring widths range from 2 to 7 μm. These electrically pumped devices operate CW at room temperature. A microring device with a 5-μm outer radius and a 3-μm ring width exhibits a laser threshold of 0.6 mA and emits 8 μW near $\lambda_o \approx 1.3$ μm at a drive current of 2 mA. In comparison with

[†] See Y. Wan, J. Norman, Q. Li, M. J. Kennedy, D. Liang, C. Zhang, D. Huang, Z. Zhang, A. Y. Liu, A. Torres, D. Jung, A. C. Gossard, E. L. Hu, K. M. Lau, and J. E. Bowers, 1.3 μm Submilliamp Threshold Quantum Dot Micro-Lasers on Si, *Optica*, vol. 4, pp. 940–944, 2017.

quantum-well devices, quantum-dot lasers offer substantially lower thresholds and reduced sensitivity to temperature, as well as other salutary features, as discussed in Sec. 18.4C.

GeSn-on-Si microdisk laser. Direct-bandgap lasers can be fabricated by growing an alloy of Ge and α-Sn on a Ge-buffered Si substrate. An alloy of composition $Ge_{0.915}Sn_{0.085}$ contains just enough Sn to equalize the energies of the direct-bandgap (Γ) valley and the energetically lowest indirect-bandgap (L) valley in the band structure of Ge, thereby creating a direct-bandgap group-IV material. An 8-μm-diameter $Ge_{0.875}Sn_{0.125}$ microdisk laser, pumped with a pulsed Nd:YAG laser operated at $\lambda_o = 1.064$ μm, lases at $\lambda_o \approx 2.5$ μm.[†] In the current state of its development, this optically pumped, pulsed microdisk laser requires cooling but double-heterostructure $Si_xGe_{1-x-y}Sn_y/Ge_{1-y}Sn_y$ devices, as well as multiquantum-well structures, promise electrically pumped, CW operation at room temperature.

Optical Vortex-Beam Lasers

Light beams carrying orbital angular momentum (OAM) possess helical wavefronts which, by virtue of this additional degree of freedom, are appealing for use in specialized applications. An example of such an optical vortex is the Laguerre–Gaussian beam (Sec. 3.4); beams with spiral wavefronts can be produced with the help of holographic optical elements (Example 4.5-3). An optical vortex beam can also be generated by injecting light into a whispering-gallery mode microcavity fitted with an embedded refractive-index grating structure in the azimuthal direction. Recently, a microcavity laser that directly generates a single-mode OAM vortex beam, with a topological charge that can be chosen at will, has been developed.[‡] Unidirectional lasing in a microring resonator that supports whispering-gallery modes with large values of OAM is induced by selectively modulating the refractive index and gain/loss, which breaks the rotational symmetry of the lasing process. The microring sidewalls are designed in such a way that scattering causes the vortex beam to emerge vertically from the plane of the device, mimicking the output of a VCSEL.

C. Photonic-Crystal Lasers

Microcavities consisting of defects in photonic crystals (Sec. 11.4D), together with miniature quantum-confined emission sources such as quantum wells or quantum dots, can serve as wavelength-size lasers and laser arrays. **Photonic-crystal lasers**, with active volumes smaller than those of VCSELs and microdisk lasers, offer ultralow thresholds, ultralow power consumption, and high direct-modulation rates. They are suitable for sending information over distances of centimeters or millimeters, and promise substantially reduced power consumption in venues such as datacenters, where on-chip and rack-to-rack communications has long been mediated electrically (Sec. 24.1D).

2D Photonic-Crystal Lasers

Examples of individual 2D photonic-crystal devices, as well as coherently coupled arrays of such devices, are illustrated in Figs. 18.5-7(a) and (b), respectively. The device displayed in Fig. 18.5-7(a) is a single-mode 2D photonic-bandgap laser that operates at room temperature. It is electrically pumped via a sub-micron-size post and

[†] See D. Stange, S. Wirths, R. Geiger, C. Schulte-Braucks, B. Marzban, N. von den Driesch, G. Mussler, T. Zabel, T. Stoica, J.-M. Hartmann, S. Mantl, Z. Ikonic, D. Grützmacher, H. Sigg, J. Witzens, and D. Buca, Optically Pumped GeSn Microdisk Lasers on Si, *ACS Photonics*, vol. 3, pp. 1279–1285, 2016.

[‡] See P. Miao, Z. Zhang, J. Sun, W. Walasik, S. Longhi, L. M. Lichinitser, and L. Feng, Orbital Angular Momentum Microlaser, *Science*, vol. 353, pp. 464–467, 2016.

has a threshold current of 260 μA.* The active region comprises six strained InGaAsP quantum wells and lasing occurs at $\lambda_o = 1520$ nm. The structure produces 2 nW of power at a current of $^1/_2$ mA and has a differential responsivity R $\approx 10^{-5}$. The quality factor and modal volume are $Q \approx 2500$ and $V \approx 0.06\ \mu\text{m}^3$, respectively. Since the emission linewidth $\Delta\nu$ is smaller than the width of an electromagnetic mode in the device $\delta\nu$, and the quality factor Q of the microcavity is high, spontaneous emission is enhanced via the Purcell effect (Sec. 14.3E and Fig. 14.3-12). The Purcell factor for this device is $F_P = (3/4\pi^2)(\lambda^3/V)Q \approx 400$.

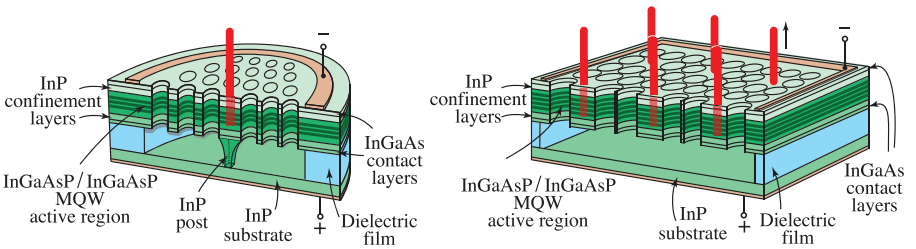


Figure 18.5-7 (a) InGaAsP/InGaAsP 2D multiquantum-well photonic-crystal laser. The InP post has a height of 1 μm and serves as an electrical contact. (b) Array of coherently coupled 2D quantum-well photonic-crystal lasers.

The nW-level output power of an individual device may be substantially enhanced by constructing a coherently coupled **microcavity-array laser**, as portrayed in Fig. 18.5-7(b). This particular array comprises 81 cavities and four InGaAsP/InP quantum wells, with an overall array area $\approx 15\ \mu\text{m}^2$ and output power $\approx 12\ \mu$W. A distributed-feedback version of a 2D photonic-crystal laser, in which feedback takes place over the entire 2D plane, achieves what is known as **band-edge lasing**. 2D photonic-crystal lasers rely on conventional reflection for out-of-plane confinement.

3D Photonic-Crystal Lasers

Lasing may also be attained in a 3D photonic-crystal microcavity with a complete photonic bandgap (Sec. 7.3B). Such devices, which operate at room temperature and often make use of active regions comprising electrically pumped stacks of quantum-dot layers, can attain values of $Q \approx 40\,000$ and modulation rates of 10 Gb/s. Both electrons and photons are confined in three dimensions in such devices.†

EXAMPLE 18.5-1. ***Thresholdless Quantum-Dot Photonic-Crystal Laser.*** In a cavity of wavelength size with dimensions such that only one optical mode is permitted to exist, photon emission must be into that unique mode, whether the emission is via spontaneous or stimulated emission. Since there is then no distinction between the two processes, the kink in the light–current (L–i) curve that characterizes the transition from spontaneous to stimulated emission in a conventional laser, and thus determines the "laser threshold," disappears (i.e., occurs at zero current). Constructing a thresholdless laser thus involves imposing strict control on the modal structure of the resonator. The fraction of spontaneous emission that contributes to a given laser mode is specified by the

* See H.-G. Park, S.-H. Kim, S.-H. Kwon, Y.-G. Ju, J.-K. Yang, J.-H. Baek, S.-B. Kim, and Y.-H. Lee, Electrically Driven Single-Cell Photonic Crystal Laser, *Science*, vol. 305, pp. 1444–1447, 2004.

† See, e.g., K. Takeda, T. Sato, A. Shinya, K. Nozaki, W. Kobayashi, H. Taniyama, M. Notomi, K. Hasebe, T. Kakitsuka, and S. Matsuo, Few-fJ/Bit Data Transmissions Using Directly Modulated Lambda-Scale Embedded Active Region Photonic-Crystal Lasers, *Nature Photonics*, vol. 7, pp. 569–575, 2013.

spontaneous-emission coupling coefficient β. This parameter is typically minuscule for conventional laser diodes; for an edge-emitting device it turns out to be $\beta \approx 10^{-5}$. As indicated earlier, microcavity lasers are designed to optimize the inhibition and enhancement of spontaneous emission. Indeed, the modification of the modal density provided by microcavities (and nanocavities) can increase β by many orders of magnitude. Implementing such a modification has the concomitant benefit of reducing the laser-diode threshold current i_t by a commensurate amount.

An optically pumped laser is characterized by a **light–light (L–L) curve** analogous to the light–current (L–i) curve of an electrically injected laser diode. A nearly thresholdless optically pumped laser ($\beta = 0.85$), with a threshold intensity $I_t < 1$ μW, has been fabricated by embedding InAsSb quantum dots in a photonic-crystal microcavity. The device is pumped by a CW laser diode at 785 nm. It emits light at $\lambda_o \approx 1.3$ μm and operates at room temperature.[†]

18.6 NANOCAVITY LASERS

The lasers considered in Secs. 18.4 and 18.5 rely on refractive-index differences between dielectrics to confine light in their resonators. The physical and modal sizes of such lasers in any given direction are therefore greater than the wavelength of the emitted light. In this section, we consider lasers that rely on metal-based nonplasmonic and plasmonic resonators that confine light to subwavelength dimensions while operating at optical frequencies. The active media in such **nanocavity lasers**, or **nanolasers**, are often semiconductors in the form of bulk, quantum-well, or quantum-dot structures, but organic semiconductors and dye solutions are used as well. As a consequence of dissipation, lasers with metal-based resonators have substantially reduced values of the quality factor Q that must be overcome by increased gain. Both bottom-up and top-down approaches are used in fabrication.

Both non-plasmonic and plasmonic nanolasers have been developed. Non-plasmonic devices, consisting of subwavelength-size metal cavities that encapsulate high-gain materials, support non-evanescent electromagnetic modes. The metal cladding limits the transverse size of the device, which enhances modal confinement and diminishes the volume of the active region. These devices can be electrically pumped, operated at room temperature, and operated at timescales that stretch from femtoseconds to CW.

Plasmonic nanolasers can be constructed with overall physical and modal sizes of the order of tens to hundreds of nanometers — well below the micrometer sizes associated with microcavity lasers. Plasmonic devices operate on the basis of propagating surface plasmon polaritons (SPPs) or nonpropagating localized surface plasmons (LSPs). Both versions make use of metal–dielectric or metal–semiconductor interfaces to confine light to subwavelength dimensions. Surface plasmon polariton nanolasers use resonators such as metal–dielectric nanodisks (Sec. 11.4E), or plasmonic waveguides such as metal–insulator–metal (MIM), metal–insulator–semiconductor (MIS), or metal-slab structures (Sec. 9.6). Gain is provided by replacing the dielectric insulating material in these resonators with an active gain medium, such as a direct-bandgap semiconductor or a dye. Feedback is provided in the same manner as for conventional lasers, namely via a Fabry–Perot cavity, a DFB grating, or a circular disk or ring structure that supports whispering-gallery modes. Propagating SPPs do not rely on resonances, so this class of devices offers broadband operation. Localized surface plasmon nanolasers, on the other hand, make use of plasmonic resonators such as the metallic nanosphere (Fig. 8.2-6 and Sec. 11.4E). The gain is then provided by replacing the surrounding dielectric material with an active gain medium such as a dye-impregnated shell.

[†] See I. Prieto, J. M. Llorens, L. E. Muñoz-Camúñez, A. G. Taboada, J. Canet-Ferrer, J. M. Ripalda, C. Robles, G. Muñoz-Matutano, J. P. Martínez-Pastor, and P. A. Postigo, Near Thresholdless Laser Operation at Room Temperature, *Optica*, vol. 2, pp. 66–69, 2015.

Diminishing the size of a laser to the nanoscale has manifold benefits for certain applications. These include:

- Reduced laser threshold
- Increased differential power-conversion efficiency (slope efficiency)
- Increased modulation bandwidth

Nanometer-size devices can be used in chip-scale optical communications and data-processing, display technology, and near-field photolithography. They can also be deployed in wireless sensor networks. Moreover, nanolasers can be implanted or injected into biological materials to carry out tasks such as imaging, sensing, spectroscopy, and therapeutics.

Representative examples of these three types of nanolasers are set forth below: 1) a metal-nanocavity laser; 2) a SPP nanoring laser; and 3) a LSP nanosphere laser.

Non-Plasmonic Nanocavity Lasers

Non-plasmonic nanolasers make use of subwavelength-size external metal cavities that support non-evanescent electromagnetic modes and encapsulate gain media such as semiconductors.

Ag-clad InGaAs metal-nanocavity laser. We consider a subwavelength metal-cavity nanolaser that takes the form of a rectangular pillar of InP/InGaAs/InP, with a SiN insulating layer, encapsulated in a rectangular silver shell that serves as a metallic Fabry–Perot resonator.[†] The light is mostly confined within the rectangular bulk-InGaAs active region (refractive index $n = 3.4$), which is sandwiched between two rectangular InP confining regions of lower refractive index ($n = 3.1$). The device operates CW at room temperature and is electrically pumped (threshold current $i_{\mathrm{t}} = 1.1$ mA). It lases at $\lambda_o = 1.59$ μm with a linewidth of 0.5 nm. The cavity volume is $0.67\,\lambda_o^3$ and the quality factor corresponding to the observed linewidth is $Q = 3182$. Configured in a flip-chip configuration, the emitted light passes through the substrate and emerges vertically. Despite the presence of the lossy metal comprising the cavity, the properties and performance of this semiconductor nanocavity laser are comparable to those of a conventional semiconductor microcavity laser, thereby demonstrating that metallic dissipation can be vanquished by semiconductor gain. Replacing the bulk-semiconductor active region with quantum wells or quantum dots should improve performance, as should improved heat dissipation.

Plasmonic Nanocavity Lasers

Plasmonic nanolasers incorporate metals within subwavelength-size internal structures that support evanescent modes and are juxtaposed with gain media such as semiconductors or dyes. Plasmonic devices function either via surface plasmon polariton (SPP) traveling waves (Sec. 8.2B) or via localized surface plasmon (LSP) oscillations (Sec. 8.2C) at their metal–dielectric boundaries. We first consider an example of an SPP plasmonic nanolaser and follow this with an example of an LSP plasmonic nanolaser (spaser).

MIS-on-Si surface-plasmon-polariton nanoring laser. Subwavelength nanocavity lasers based on the amplification of surface plasmon polaritons (SPPs) have been developed in a number of configurations. One didactic version makes use of a Ag-Al_2O_3-AlInGaP metal–insulator–semiconductor (MIS) structure that supports

[†] See K. Ding, M. T. Hill, Z. C. Liu, L. J. Yin, P. J. van Veldhoven, and C. Z. Ning, Record Performance of Electrical Injection Sub-Wavelength Metallic-Cavity Semiconductor Lasers at Room Temperature, *Optics Express*, vol. 21, pp. 4728–4733, 2013.

surface plasmon polariton whispering-gallery modes in a nanoring cavity.[‡] The lithographically defined ring has inner and outer diameters of 0.79 and 1.09 μm, respectively, so that the ring width is 150 nm. The nanolaser sits on a Si substrate and operates at room temperature. The AlInGaP heterostructure gain medium has an overall thickness of 110 nm and provides broadband gain at room temperature. The 570-nm pump laser supplies 4-ps pulses at a repetition rate of 1 kHz. The onset of lasing from a pair of 2D whispering gallery modes, at 610 and 634 nm, occurs at a threshold pump energy density of $\approx$ 2 mJ/cm^2 per pulse. Similar devices have been fabricated with other cavity geometries, including waveguides, 3D disks and squares, and arrays.

Dye-clad-Au nanosphere localized-surface-plasmon spaser. A subwavelength nanocavity laser based on the amplification of localized surface plasmons (LSPs) is often referred to as a **spaser**, an acronym for **S**urface **P**lasmon **A**mplification by **S**timulated **E**mission of **R**adiation, a term coined by Bergman and Stockman in 2003. In analogy with a laser, a spaser generates stimulated emission of localized surface plasmons in resonant metallic nanostructures juxtaposed with a gain medium. The first spaser-based nanolaser, which dates to 2009, consisted of a collection of 15-nm-diameter gold-nanosphere resonators surrounded by silica shells doped with green dye to provide gain and compensate for metallic dissipative losses. Pump pulses were provided by an optical parametric oscillator operated at 488 nm. A collection of these 44-nm-diameter core–shell nanoparticles supported surface plasmon oscillations that out-coupled to photonic modes at a wavelength of 531 nm.[†] Subsequently, a spaser with tunable output wavelength was implemented by replacing the Au nanospheres with Au nanorods and by making use of different organic dyes at various doping levels as gain media.

READING LIST

LEDs and OLEDs

See also the reading list in Chapter 17.

T.-Y. Seong, J. Han, H. Amano, and H. Morkoç, eds., *III-Nitride Based Light Emitting Diodes and Applications*, Springer-Verlag, 2nd ed. 2017.

D. J. Gaspar and E. Polikarpov, eds., *OLED Fundamentals: Materials, Devices, and Processing of Organic Light-Emitting Diodes*, CRC Press/Taylor & Francis, 2015.

M. H. Crawford, J. J. Wierer, A. J. Fischer, G. T. Wang, D. D. Koleske, G. S. Subramania, M. E. Coltrin, J. Y. Tsao, and R. F. Karlicek, Jr., Solid-State Lighting: Toward Smart and Ultra-Efficient Materials, Devices, Lamps and Systems, in D. L. Andrews, ed., *Photonics: Scientific Foundations, Technology and Applications*, Volume III, *Photonics Technology and Instrumentation*, Wiley–Science Wise, 2015.

V. K. Khanna, *Fundamentals of Solid-State Lighting: LEDs, OLEDs, and Their Applications in Illumination and Displays*, CRC Press/Taylor & Francis, 2014.

J.-J. Huang, H.-C. Kuo, and S.-C. Shen, eds., *Nitride Semiconductor Light-Emitting Diodes (LEDs): Materials, Technologies and Applications*, Woodhead, 2014.

A. Buckley, ed., *Organic Light-Emitting Diodes (OLEDs): Materials, Devices and Applications*, Woodhead, 2013.

S. Nakamura and M. R. Krames, History of Gallium-Nitride-Based Light-Emitting Diodes for Illumination, *Proceedings of the IEEE*, vol. 101, pp. 2211–2220, 2013.

E. F. Schubert, *Light-Emitting Diodes*, Cambridge University Press, 2nd ed. 2006.

[‡] See N. Liu, A. Gocalinska, J. Justice, F. Gity, I. Povey, B. McCarthy, M. Pemble, E. Pelucchi, H. Wei, C. Silien, H. Xu, and B. Corbett, Lithographically Defined, Room Temperature Low Threshold Subwavelength Red-Emitting Hybrid Plasmonic Lasers, *Nano Letters*, vol. 16, pp. 7822–7828, 2016.

[†] See M. A. Noginov, G. Zhu, A. M. Belgrave, R. Bakker, V. M. Shalaev, E. E, Narimanov, S. Stout, E. Herz, T. Suteewong, and U. Wiesner, Demonstration of a Spaser-Based Nanolaser, *Nature*, vol. 460, pp. 1110–1113, 2009.

J. H. Burroughes, D. D. C. Bradley, A. R. Brown, R. N. Marks, K. Mackay, R. H. Friend, P. L. Burn, and A. B. Holmes, Light-Emitting Diodes Based on Conjugated Polymers, *Nature*, vol. 347, pp. 539-541, 1990.

C. W. Tang and S. A. VanSlyke, Organic Electroluminescent Diodes, *Applied Physics Letters*, vol. 51, pp. 913–915, 1987.

H. J. Round, A Note on Carborundum, *Electrical World*, vol. 49, p. 309, 1907.

Silicon Photonics

See also the reading list on photonic integrated circuits in Chapter 25.

D. Van Thourhout, Z. Wang, and G. Roelkens, III–V-on-Silicon Integration, *Optics & Photonics News*, vol. 28, no. 3, pp. 32–39, 2017.

D. Thomson, A. Zilkie, J. E. Bowers, T. Komljenovic, G. T. Reed, L. Vivien, D. Marris-Morini, E. Cassan, L. Virot, J.-M. Fédéli, J.-M. Hartmann, J. H. Schmid, D.-X. Xu, F. Boeuf, P. O'Brien, G. Z. Mashanovich, and M. Nedeljkovic, Roadmap on Silicon Photonics, *Journal of Optics*, vol. 18, 073003, 2016.

J. W. Silverstone, D. Bonneau, J. L. O'Brien, and M. G. Thompson, Silicon Quantum Photonics, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 22, 6700113, 2016.

L. Chrostowski and M. Hochberg, *Silicon Photonics Design: From Devices to Systems*, Cambridge University Press, 2015.

A. Y. Liu, S. Srinivasan, J. Norman, A. C. Gossard, and J. E. Bowers, Quantum Dot Lasers for Silicon Photonics, *Photonics Research*, vol. 3, no. 5, pp. B1–B9, 2015.

Z. Zhou, B. Yin, and J. Michel, On-Chip Light Sources for Silicon Photonics, *Light: Science & Applications* (2015) **4**, e358; doi:10.1038/lsa.2015.131

W. Bogaerts, Silicon Photonics, in D. L. Andrews, ed., *Photonics: Scientific Foundations, Technology and Applications*, Volume II, *Nanophotonic Structures and Materials*, Wiley–Science Wise, 2015.

B. R. Koch, S. Srinivasan, and J. E. Bowers, Hybrid Silicon Lasers, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

L. Vivien and L. Pavesi, eds., *Handbook of Silicon Photonics*, CRC Press/Taylor & Francis, 2013.

M. J. Deen and P. K. Basu, *Silicon Photonics: Fundamentals and Devices*, Wiley, 2012.

D. J. Lockwood and L. Pavesi, eds., *Silicon Photonics II: Components and Integration*, Springer-Verlag, 2011.

H. Rong, S. Xu, Y.-H. Kuo, V. Sih, O. Cohen, O. Raday, and M. Paniccia, Low-Threshold Continuous-Wave Raman Silicon Laser, *Nature Photonics*, vol. 1, pp. 232–237, 2007.

H. Rong, R. Jones, A. Liu, O. Cohen, D. Hak, A. Fang, and M. Paniccia, A Continuous-Wave Raman Silicon Laser, *Nature*, vol. 433, pp. 725–728, 2005.

Semiconductor Optical Amplifiers and Laser Diodes

See also the reading lists in Chapters 15–17.

J. Ohtsubo, *Semiconductor Lasers: Stability, Instability and Chaos*, Springer-Verlag, 4th ed. 2017.

P. K. Basu, B. Mukhopadhyay, and R. Basu, *Semiconductor Laser Theory*, CRC Press/Taylor & Francis, 2016.

T. Numai, *Fundamentals of Semiconductor Lasers*, Springer-Verlag, 2nd ed. 2015.

N. K. Dutta and Q. Wang, *Semiconductor Optical Amplifiers*, World Scientific, 2nd ed. 2013.

L. A. Coldren, S. W. Corzine, and M. L. Mašanović, *Diode Lasers and Photonic Integrated Circuits*, Wiley, 2nd ed. 2012.

S. Nakamura, S. Pearton, and G. Fasol, *The Blue Laser Diode: The Complete Story*, Springer-Verlag, 2nd ed. 2000.

J. J. Coleman, ed., *Selected Papers on Semiconductor Diode Lasers*, SPIE Optical Engineering Press (Milestone Series Volume 50), 1992.

H. Amano, T. Asahi, and I. Akasaki, Stimulated Emission Near Ultraviolet at Room Temperature from a GaN Film Grown on Sapphire by MOVPE Using an AlN Buffer Layer, *Japanese Journal of Applied Physics*, vol. 29, pp. L205–L206, 1990.

Zh. I. Alferov, V. M. Andreev, E. L. Portnoi, and M. K. Trukan, AlAs-GaAs Heterojunction Injection Lasers with a Low Room-Temperature Threshold, *Soviet Physics–Semiconductors*, vol. 3, pp. 1107–1110, 1970 [*Fizika i Tekhnika Poluprovodnikov*, vol. 3, pp. 1328–1332, 1969].

Quantum-Confined and Microcavity Lasers

See also the reading list on quantum-confined materials and nanostructures in Chapter 17.

A. Rahimi-Iman, Recent Advances in VECSELs, *Journal of Optics*, vol. 18, 093003, 2016.

P. Blood, *Quantum Confined Laser Devices: Optical Gain and Recombination in Semiconductors*, Oxford University Press, 2015.

J. Wu, S. Chen, A. Seeds, and H. Liu, Quantum Dot Optoelectronic Devices: Lasers, Photodetectors and Solar Cells, *Journal of Physics D: Applied Physics*, vol. 48, 363001, 2015.

Y. Zhang, X. Zhang, K. H. Li, Y. F. Cheung, C. Feng, and H. W. Choi, Advances in III–Nitride Semiconductor Microdisk Lasers, *Physica Status Solidi A*, vol. 212, pp. 960–973, 2015.

R. De La Rue, S. Yu, and J.-M. Lourtioz, eds., *Compact Semiconductor Lasers*, Wiley–VCH, 2014.

M. T. Hill and M. C. Gather, Advances in Small Lasers, *Nature Photonics*, vol. 8, pp. 908–918, 2014.

R. Michalzik, ed., *VCSELs: Fundamentals, Technology and Applications of Vertical-Cavity Surface-Emitting Lasers*, Springer-Verlag, 2013.

K. Takeda, T. Sato, A. Shinya, K. Nozaki, W. Kobayashi, H. Taniyama, M. Notomi, K. Hasebe, T. Kakitsuka, and S. Matsuo, Few-fJ/Bit Data Transmissions Using Directly Modulated Lambda-Scale Embedded Active Region Photonic-Crystal Lasers, *Nature Photonics*, vol. 7, pp. 569–575, 2013.

C. Jagadish and E. R. Weber, eds., *Semiconductors and Semimetals*, J. J. Coleman, A. C. Bryce, and C. Jagadish, eds., Volume 86, *Advances in Semiconductor Lasers*, Academic Press/Elsevier, 2012.

X. Cai, J. Wang, M. J. Strain, B. Johnson-Morris, J. Zhu, M. Sorel, J. L. O'Brien, M. G. Thompson, and S. Yu, Integrated Compact Optical Vortex Beam Emitters, *Science*, vol. 338, pp. 363–366, 2012.

O. G. Okhotnikov, ed., *Semiconductor Disk Lasers: Physics and Technology*, Wiley–VCH, 2010.

O. Painter, R. K. Lee, A. Scherer, A. Yariv, J. D. O'Brien, P. D. Dapkus, and I. Kim, Two-Dimensional Photonic Band-Gap Defect Mode Laser, *Science*, vol. 284, pp. 1819–1821, 1999.

N. Kirstaedter, N. N. Ledentsov, M. Grundmann, D. Bimberg, V. M. Ustinov, S. S. Ruvimov, M. V. Maximov, P. S. Kop'ev, Zh. I. Alferov, U. Richter, P. Werner, U. Gösele, and J. Heydenreich, Low Threshold, Large T_0 Injection Laser Emission from (InGa)As Quantum Dots, *Electronics Letters*, vol. 30, pp. 1416–1417, 1994.

P. S. Zory, Jr., ed., *Quantum Well Lasers*, Academic Press, 1993.

A. F. J. Levi, R. E. Slusher, S. L. McCall, T, Tanbun-Ek, D. L. Coblentz, and S. J. Pearton, Room Temperature Operation of Microdisc Lasers with Submilliamp Threshold Current. *Electronics Letters*, vol. 28, pp. 1010–1012, 1992.

J. L. Jewell, A. Scherer, S. L. McCall, Y. H. Lee, S. Walker, J. P. Harbison, and L. T. Florez, Low-Threshold Electrically Pumped Vertical-Cavity Surface-Emitting Microlasers, *Electronics Letters*, vol. 25, pp. 1123–1124, 1989.

R. D. Dupuis, P. D. Dapkus, R. Chin, N. Holonyak, Jr., and S. W. Kirchoefer, Continuous 300° K Laser Operation of Single- Quantum-Well $Al_xGa_{1-x}As$–GaAs Heterostructure Diodes Grown by Metalorganic Chemical Vapor Deposition, *Applied Physics Letters*, vol. 34, pp. 265–267, 1979.

Mid-Infrared Interband and Quantum Cascade Lasers

W. Zhou, N. Bandyopadhyay, D. Wu, R. McClintock, and M. Razeghi, Monolithically, Widely Tunable Quantum Cascade Lasers Based on a Heterogeneous Active Region Design, *Scientific Reports*, vol. 6, 25213, 2016.

M. S. Vitiello, G. Scalari, B. Williams, and P. De Natale, Quantum Cascade Lasers: 20 Years of Challenges, *Optics Express*, vol. 23, pp. 5167–5182, 2015.

J. Faist, *Quantum Cascade Lasers*, Oxford University Press, 2013.

M. Troccoli, A. Lyakh, J. Fan, X. Wang, R. Maulini, A. G. Tsekoun, R. Go, and C. K. N. Patel, Long-Wave IR Quantum Cascade Lasers for Emission in the $\lambda = 8$–12 μm Spectral Region, *Optical Materials Express*, vol. 3, 1546–1560, 2013.

Y. Yao, A. J. Hoffman, and C. F. Gmachl, Mid-Infrared Quantum Cascade Lasers, *Nature Photonics*, vol. 6, pp. 432–439, 2012.

E. Tournié and A. N. Baranov, Mid-Infrared Semiconductor Lasers: A Review, in *Semiconductors and Semimetals*, C. Jagadish and E. R. Weber, eds., Volume 86, *Advances in Semiconductor Lasers*, Academic Press/Elsevier, 2012, pp. 183–226.

A. Hugi, R. Maulini, and J. Faist, External Cavity Quantum Cascade Laser, *Semiconductor Science and Technology*, vol. 25, 083001, 2010.

B. S. Williams, Terahertz Quantum-Cascade Lasers, *Nature Photonics*, vol. 1, pp. 517–525, 2007.

R. Paiella, ed., *Intersubband Transitions in Quantum Structures*, McGraw–Hill, 2006.

J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, Quantum Cascade Laser, *Science*, vol. 264, pp. 553–556, 1994.

R. F. Kazarinov and R. A. Suris, Amplification of Electromagnetic Waves in a Semiconductor Superlattice, *Soviet Physics–Semiconductors*, vol. 5, pp. 707–709, 1971 [*Fizika i Tekhnika Poluprovodnikov*, vol. 5, pp. 797–800, 1971].

L. Esaki and R. Tsu, Superlattice and Negative Differential Conductivity in Semiconductors, *IBM Journal of Research and Development*, vol. 14, pp. 61–65, 1970.

Nanolasers

Q. Gu and Y. Fainman, *Semiconductor Nanolasers*, Cambridge University Press, 2017.

M. Karl, C. P. Dietrich, M. Schubert, I. D. W. Samuel, G. A. Turnbull, and M. C. Gather, Single Cell Induced Optical Confinement in Biological Lasers, *Journal of Physics D: Applied Physics*, vol. 50, 084005, 2017.

S. Gwo and C.-K. Shih, Semiconductor Plasmonic Nanolasers: Current Status and Perspectives, *Reports on Progress in Physics*, vol. 79, 086501, 2016.

M. I. Stockman, Quantum Nanoplasmonics, in D. L. Andrews, ed., *Photonics: Scientific Foundations, Technology and Applications*, Volume II, *Nanophotonic Structures and Materials*, Wiley–Science Wise, 2015.

M. T. Hill and M. C. Gather, Advances in Small Lasers, *Nature Photonics*, vol. 8, pp. 908–918, 2014.

S. Noda, Fundamentals of Photonic Crystals for Telecom Applications — Photonic Crystal Lasers, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

R.-M. Ma, R. F. Oulton, V. J. Sorger, and X. Zhang, Plasmon Lasers: Coherent Light Source at Molecular Scales, *Laser & Photonics Review*, vol. 7, pp. 1–21, 2013.

Y. Ma, X. Guo, X. Wu, L. Dai, and L. Tong, Semiconductor Nanowire Lasers, *Advances in Optics and Photonics*, vol. 5, pp. 216–273, 2013.

R. F. Oulton, V. J. Sorger, T. Zentgraf, R.-M. Ma, C. Gladden, L. Dai, G. Bartal, and X. Zhang, Plasmon Lasers at Deep Subwavelength Scale, *Nature*, vol. 461, pp. 629–632, 2009.

M. A. Noginov, G. Zhu, A. M. Belgrave, R. Bakker, V. M. Shalaev, E. E, Narimanov, S. Stout, E. Herz, T. Suteewong, and U. Wiesner, Demonstration of a Spaser-Based Nanolaser, *Nature*, vol. 460, pp. 1110-1113, 2009.

M. T. Hill, Y.-S. Oei, B. Smalbrugge, Y. Zhu, T. de Vries, P. J. van Veldhoven, F. W. M. van Otten, T. J. Eijkemans, J. P. Turkiewicz, H. de Waardt, E. J. Geluk, S.-H. Kwon, Y.-H. Lee, R. Nötzel, and M. K. Smit, Lasing in Metallic-Coated Nanocavities, *Nature Photonics*, vol. 1, pp. 589–594, 2007.

D. J. Bergman and M. I. Stockman, Surface Plasmon Amplification by Stimulated Emission of Radiation: Quantum Generation of Coherent Surface Plasmons in Nanosystems, *Physical Review Letters*, vol. 90, 027402, 2003.

Historical Articles

See also the reading list on historical articles and reprint collections in Chapter 16.

I. Akasaki, Fascinated Journeys into Blue Light; H. Amano, Growth of GaN on Sapphire by Low Temperature Deposited Buffer Layer and Realization of P-Type GaN by Mg-Doping Followed by LEEBI Treatment; S. Nakamura, Background Story of the Invention of Efficient Blue InGaN Light Emitting Diodes (Nobel Lectures in Physics, 2014).

Scientific Background on the Nobel Prize in Physics 2014: Efficient Blue Light-Emitting Diodes Leading to Bright and Energy-Saving White Light Sources, *Kungliga Vetenskapsakademien*, Compiled by the Class for Physics of the Royal Swedish Academy of Sciences, pp. 1–9, 2014.

R. D. Dupuis and M. R. Krames, History, Development, and Applications of High-Brightness Visible Light-Emitting Diodes, *Journal of Lightwave Technology*, vol. 26, pp. 1154–1171, 2008.

Zh. I. Alferov, Double Heterostructure Concept and its Applications in Physics, Electronics and Technology (Nobel Lecture in Physics, 2000), in G. Ekspong, ed., *Nobel Lectures in Physics*

1996–2000, World Scientific, 2003.

H. Kroemer, Quasi-Electric Fields and Band Offsets: Teaching Electrons New Tricks (Nobel Lecture in Physics, 2000), in G. Ekspong, ed., *Nobel Lectures, Physics 1996–2000*, World Scientific, 2003.

Millennium issue, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 6, no. 6, 2000.

E. E. Loebner, Subhistories of the Light Emitting Diode, *IEEE Transactions on Electron Devices*, vol. ED-23, pp. 675–699, 1976.

N. G. Basov, Semiconductor Lasers (Nobel Lecture in Physics, 1964), in *Nobel Lectures in Physics, 1963–1970*, World Scientific, 1998, pp. 89–105.

T. M. Quist, R. H. Rediker, R. J. Keyes, W. E. Krag, B. Lax, A. L. McWhorter, and H. J. Zeiger, Semiconductor Maser of GaAs, *Applied Physics Letters*, vol. 1, pp. 91–92, 1962.

N. Holonyak, Jr. and S. F. Bevacqua, Coherent (Visible) Light Emission from Ga($As_{1-x}P_x$) Junctions, *Applied Physics Letters*, vol. 1, pp. 82–83, 1962.

M. I. Nathan, W. P. Dumke, G. Burns, F. H. Dill, Jr., and G. Lasher, Stimulated Emission of Radiation from GaAs p–n Junctions, *Applied Physics Letters*, vol. 1, pp. 62–64, 1962.

R. N. Hall, G. E. Fenner, J. D. Kingsley, T. J. Soltys, and R. O. Carlson, Coherent Light Emission from GaAs Junctions, *Physical Review Letters*, vol. 9, pp. 366–368, 1962.

J. I. Pankove and J. E. Berkeyheiser, A Light Source Modulated at Microwave Frequencies, *Proceedings of the IRE*, vol. 50, pp. 1976–1977, 1962.

J. I. Pankove, Tunneling-Assisted Photon Emission in Gallium Arsenide pn Junctions, *Physical Review Letters*, vol. 9, pp. 283–285, 1962.

R. J. Keyes and T. M. Quist, Recombination Radiation Emitted by Gallium Arsenide, *Proceedings of the IRE*, vol. 50, pp. 1822–1823, 1962.

J. R. Biard and G. E. Pittman, Semiconductor Radiant Diode, *U.S. Patent 3,293,513*, Patented December 20, 1966 (Filed August 8, 1962).

N. G. Basov, O. N. Krokhin, and Yu. M. Popov, Production of Negative-Temperature States in pn Junctions of Degenerate Semiconductors, *Soviet Physics–JETP*, vol. 13, pp. 1320–1321, 1961 [*Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki*, vol. 40, pp. 1879–1880, 1961].

M. G. A. Bernard and G. Duraffourg, Laser Conditions in Semiconductors, *Physica Status Solidi*, vol. 1, pp. 699–703, 1961.

John von Neumann, in unpublished calculations sent to Edward Teller in September 1953, showed that it was possible, in principle, to upset the equilibrium concentration of carriers in a semiconductor and thereby to obtain light amplification by stimulated emission, e.g., via the recombination of electrons and holes injected into a p–n junction [see J. von Neumann, Notes on the Photon-Disequilibrium-Amplification Scheme (JvN), Sept. 16, 1953, *IEEE Journal of Quantum Electronics*, vol. QE-23, pp. 659–673, 1987].

PROBLEMS

18.1-5 **LED Spectral Widths.** Consider seven LED spectra drawn from Figs. 18.1-14 and P18.1-5, namely those centered at $\lambda_o = 0.37$, 0.53, 0.64, 0.91, 1.30, 1.93, and 2.25 μm. Graphically estimate their spectral widths (FWHM) in units of nm, Hz, and eV. Compare your estimates with the results calculated from the formulas provided in Exercise 18.1-3. Estimate the alloy broadening in the LED spectrum centered at $\lambda_o = 0.53$ μm in units of nm, Hz, and eV.

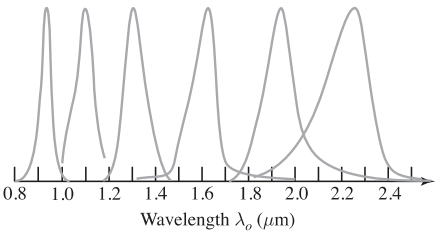


Figure P18.1-5 Spectral intensities versus wavelength for InGaAsP LEDs operating in the near-infrared region of the spectrum. The peak intensities are all normalized to the same value. The spectral width generally increases as λ_p^2, in accordance with (18.1-30).

18.1-6 **Extraction Efficiency for an LED.** Derive an expression for η_e, the efficiency for the extraction of internal unpolarized light from an LED, that includes the angular dependence of Fresnel reflection at the semiconductor–air boundary (refer to Sec. 6.2).

18.1-7 **Coupling Light from an LED into an Optical Fiber.** Calculate the fraction of optical power emitted from an LED that is accepted by a step-index optical fiber of numerical aperture $\mathrm{NA} = 0.1$ in air and core refractive index 1.46 (refer to Sec. 10.1). Assume that the LED has a planar surface, a refractive index $n = 3.6$, and an angular dependence of optical power that is proportional to $\cos^4(\theta)$. Assume further that the LED is bonded to the core of the fiber and that the emission area is smaller than the fiber core.

18.2-1 **Bandwidth of a Semiconductor Optical Amplifier.** Use the data in Fig. 18.2-3(a) to plot the full bandwidth of the InGaAsP SOA amplifier as a function of the injected-carrier concentration $\Delta\mathrm{n}$. Determine an approximate linear formula for this bandwidth as a function of $\Delta\mathrm{n}$ and, using the data in Fig. 18.2-3(b), plot the peak gain coefficient versus bandwidth.

18.2-2 **Peak Gain Coefficient of a Semiconductor Optical Amplifier at $T = 0°$ K.**

(a) Show that the peak value γ_p of the gain coefficient $\gamma_0(\nu)$ at $T = 0°$ K is located at $\nu = (E_{fc} - E_{fv})/h$.

(b) Obtain an analytical expression for the peak gain coefficient γ_p as a function of the injected-carrier concentration $\Delta\mathrm{n}$ at $T = 0°$ K.

(c) Plot γ_p versus $\Delta\mathrm{n}$ for an InGaAsP amplifier ($\lambda_o = 1300$ nm, $n = 3.5$, $\tau_\mathrm{r} = 2.5$ ns, $m_c = 0.06\,m_0$, $m_v = 0.4\,m_0$) for values of $\Delta\mathrm{n}$ in the range 1×10^{18} to 2×10^{18} cm^{-3}.

(d) Compare these results with the data provided in Fig. 18.2-3(b).

*18.2-3 **Gain Coefficient of a GaAs Semiconductor Optical Amplifier.** A room-temperature ($T = 300°$ K) p-type GaAs SOA ($E_g \approx 1.40$ eV, $m_c = 0.07\,m_0$, $m_v = 0.50\,m_0$), with refractive index $n = 3.6$, is doped ($\mathrm{p}_0 = 1.2 \times 10^{18}$) such that the radiative recombination lifetime $\tau_\mathrm{r} \approx 2$ ns.

(a) Given the steady-state injected-carrier concentration $\Delta\mathrm{n}$ (which is controlled by the injection rate R and the overall recombination time τ), use (18.2-2)–(18.2-4) to compute the gain coefficient $\gamma_0(\nu)$ versus the photon energy $h\nu$, assuming that $T = 0°$ K.

(b) Carry out the same calculation numerically, assuming that $T = 300°$ K.

(c) Plot the peak gain coefficient as a function of $\Delta\mathrm{n}$ for both cases.

(d) Determine the loss coefficient α and the transparency concentration $\Delta\mathrm{n}_\mathrm{T}$ using the linear approximation model.

(e) Plot the full amplifier bandwidth (in Hz, nm, and eV) as a function of $\Delta\mathrm{n}$ for both cases.

(f) Compare your results with the gain coefficient and peak gain coefficient curves shown in Fig. P18.2-3.

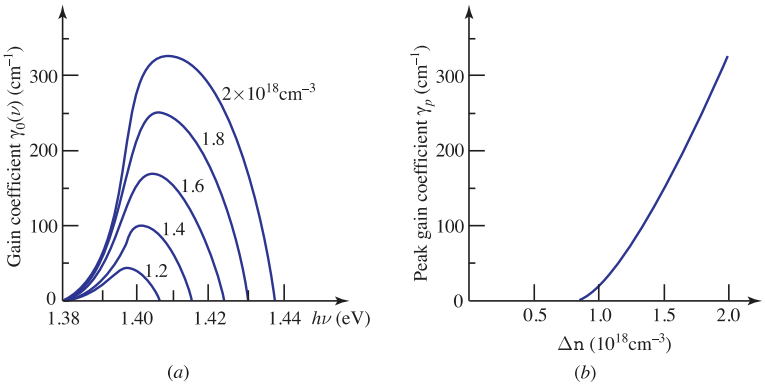


Figure P18.2-3 Gain coefficient and peak gain coefficient for a GaAs SOA. (Adapted from M. B. Panish, Heterostructure Injection Lasers, *Proceedings of the IEEE*, vol. 64, pp. 1512–1540, Fig. 4 ©1976 IEEE.)

18.2-4 **Bandgap Reduction Arising from Band-Tail States.** The bandgap reduction ΔE_g arising from band-tail states in InGaAsP and GaAs can be empirically expressed as

$$\Delta E_g(\text{eV}) \approx (-1.6 \times 10^{-8}) \left(\mathrm{p}^{1/3} + \mathrm{n}^{1/3}\right),$$

where n and p are the carrier concentrations (cm^{-3}) provided by doping or carrier injection or both.

(a) For p-type InGaAsP and GaAs, determine the concentration p that reduces the bandgap by approximately 0.02 eV.

(b) For undoped InGaAsP and GaAs, determine the injected-carrier density $\Delta\mathrm{n}$ that reduces the bandgap by approximately 0.02 eV. Assume that $\mathrm{n_i}$ is negligible.

(c) Compute $E_g + \Delta E_g$ and compare the result with the energy at which the gain coefficient in Fig. P18.2-3(a) is zero on the low-frequency side.

18.2-5 **Amplifier Gain and Bandwidth.** GaAs has an intrinsic carrier concentration $\mathrm{n_i} = 1.8 \times 10^6$ cm^{-3}, a recombination lifetime $\tau = 50$ ns, a bandgap energy $E_g = 1.42$ eV, an effective electron mass $m_c = 0.07\, m_0$, and an effective hole mass $m_v = 0.50\, m_0$. Assume that $T = 0°$ K.

(a) Determine the center frequency, bandwidth, and peak net gain within the bandwidth for a GaAs amplifier of length $d = 200$ μm, width $w = 10$ μm, and thickness $l = 2$ μm, when 1 mA of current is passed through the device.

(b) Determine the number of voice messages that can be supported by the bandwidth determined above, given that each message occupies a bandwidth of 4 kHz.

(c) Determine the bit rate that can be passed through the amplifier given that each voice channel requires 64 kb/s.

18.2-6 **Transition Cross Section.** Determine the transition cross section $\sigma(\nu)$ for GaAs as a function of $\Delta\mathrm{n}$ at $T = 0°$ K. The probability density for stimulated emission or absorption is $\phi\sigma(\nu)$, where ϕ is the photon-flux density. Why is the transition cross section less useful for semiconductor optical amplifiers than for other laser amplifiers?

*18.2-7 **Gain Profile.** Consider a 1550-nm InGaAsP amplifier ($n = 3.5$) of the configuration shown in Fig. 18.2-6, with identical antireflection coatings on its input and output facets. Calculate the maximum reflectance of each of the facets that can be tolerated if it is desired to maintain the variations in the gain profile arising from the frequency dependence of the Fabry–Perot transmittance to less than 10% [refer to (7.1-30)].

18.3-1 **Dependence of Laser-Diode Output Power on Refractive Index.** Identify the terms in the output photon flux Φ_o provided in (18.3-10) that depend on the refractive index of the crystal.

18.3-2 **Number of Longitudinal Modes.** A current is injected into an InGaAsP diode of bandgap energy $E_g = 0.91$ eV and refractive index $n = 3.5$ such that the difference in Fermi levels is $E_{fc} - E_{fv} = 0.96$ eV. If the resonator is of length $d = 250$ μm and has no losses, determine the maximum number of longitudinal modes that can oscillate.

18.3-3 **Minimum Gain Required for Lasing.** A 500-μm-long InGaAsP crystal operates at a wavelength where its refractive index $n = 3.5$. Neglecting scattering and other losses, determine the gain coefficient required to barely compensate for reflection losses at the crystal boundaries.

*18.3-4 **Modal Spacings with a Wavelength-Dependent Refractive Index.** The frequency separation of the modes of a laser diode is complicated by the fact that the refractive index is wavelength dependent, i.e., $n = n(\lambda_o)$. A laser diode of length 430 μm oscillates at a central wavelength $\lambda_c = 650$ nm. Within the emission bandwidth, $n(\lambda_o)$ may be assumed to be linearly dependent on λ_o [i.e., $n(\lambda_o) = n_0 - a(\lambda_o - \lambda_c)$, where $n_0 = n(\lambda_c) = 3.4$ and $a = dn/d\lambda_o$].

(a) The separation between the laser modes with wavelength near λ_c is observed to be $\Delta\lambda \approx 0.12$ nm. Explain why this does not correspond to the usual modal spacing $\nu_F = c/2d$.

(b) Obtain an estimate for a.

(c) Explain the phenomenon of mode pulling in a gas laser and compare it with the effect described above in semiconductor lasers.

CHAPTER

19

PHOTODETECTORS

19.1 PHOTODETECTORS 873
- A. External and Internal Photoeffects
- B. General Properties

19.2 PHOTOCONDUCTORS 883
- A. Intrinsic Photoconductors
- B. Extrinsic Photoconductors
- C. Heterostructure Photoconductors

19.3 PHOTODIODES 887
- A. The *p–n* Photodiode
- B. The *p–i–n* Photodiode
- C. Heterostructure Photodiodes

19.4 AVALANCHE PHOTODIODES 895
- A. Conventional Avalanche Photodiodes
- B. History- and Position-Dependent Parameters
- C. Single-Photon and Photon-Number-Resolving Detectors

19.5 ARRAY DETECTORS 907

19.6 NOISE IN PHOTODETECTORS 909
- A. Photoelectron Noise
- B. Gain Noise
- C. Circuit Noise
- D. Signal-to-Noise Ratio and Analog Receiver Sensitivity
- E. Bit Error Rate and Digital Receiver Sensitivity

Heinrich Hertz (1857–1894) discovered the photoelectric effect in 1887; its origin was explained by Einstein in 1905.

Siméon Denis Poisson (1781–1840) developed the fundamental probability distribution that describes photodetector noise.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

A photodetector is a device whose electrical characteristics (e.g., current, voltage, resistance) vary when exposed to light. By converting the energy of the absorbed photons into a measurable form, it can be used to determine the photon flux (or optical power) of a light beam. It can also be used to display temporal and/or spatial interference between incident optical beams. Two principal classes of photodetectors are in common use, **photoelectric detectors** and **thermal detectors**:

1. The operation of *photoelectric detectors* is based on the **photoelectric effect**, also called the **photoeffect**. The absorption of photons by a material causes electrons to transition to higher energy levels, resulting in mobile charge carriers. Under the effect of an electric field, these carriers move and produce a measurable electric current. The photoeffect takes two forms: external and internal. The **external photoeffect** involves **photoelectric emission**, also called **photoemission**, in which the photogenerated electrons escape from the material as free electrons. The **internal photoeffect** involves **photoconductivity**, in which the excited carriers remain within the material and serve to increase its conductivity.
2. *Thermal detectors* operate by converting photon energy into heat energy via any of a number of effects. **Bolometers** and **microbolometers** operate on the basis of temperature-induced changes in the resistance of a material while **thermocouples** and **thermopiles** rely on the **thermoelectric effect**, which is associated with the direct conversion of a temperature difference into a voltage difference across two dissimilar juxtaposed metals. **Pyroelectric detectors**, which are responsive to the rate of change of temperature, develop a surface voltage difference when heated. This is caused by a modification of the atomic positions within the crystal, which alter its polarization density. Finally, **Golay cells** are enclosures that contain an infrared absorbing gas and a flexible membrane. Incident infrared radiation heats the gas, which increases its pressure and deforms the membrane. The level of an auxiliary source of light that reflects from the membrane registers its motion and thus reveals the infrared power incident on the cell. Thermal detectors have long been considered to be inefficient and slow in comparison with photoelectric detectors because of the time required to effect a temperature change. Yet recent advances in manufacturing and miniaturization have dramatically improved the performance of thermal array detectors so that they are now viable contenders for imaging applications in the mid-infrared spectral region.

This Chapter

This chapter is devoted principally to a study of various photoelectric detectors that find use in photonics. We begin in Sec. 19.1 with a discussion of the external and internal photoelectric effects and set forth several important general properties of photodetectors, including quantum efficiency, responsivity, and response time. We then direct our attention to three types of semiconductor photodetectors that rely on the internal photoeffect: photoconductors, photodiodes, and avalanche photodiodes, which are considered in Secs. 19.2, 19.3, and 19.4, respectively. Array detectors, which produce electronic versions of optical images, are discussed in Sec. 19.5.

To assess the performance of photodetectors in various applications, it is important to understand their noise properties, and these are set forth in Sec. 19.6. Noise in the output circuit of a photoelectric detector arises from several sources: the photon character of the light itself (photon noise), the conversion of photons to photocarriers (photoelectron noise), the generation of secondary carriers by internal amplification mechanisms (gain noise), and receiver circuit noise. The chapter closes with a discussion of the performance of analog and digital optical receivers.

19.1 PHOTODETECTORS

A. External and Internal Photoeffects

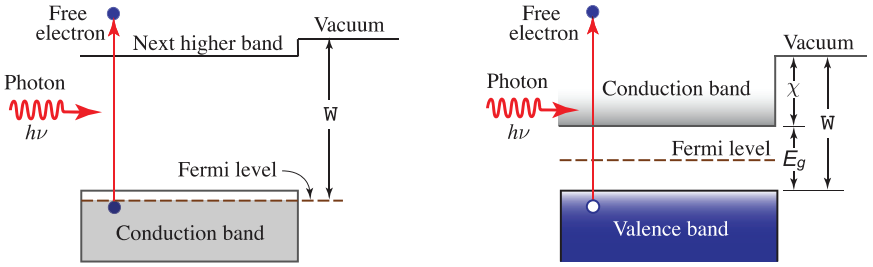
Photoelectric Emission

If the energy of a photon illuminating a material in vacuum is sufficiently large, the excited electron can escape over the potential barrier of the surface of the material and be liberated into the vacuum as a free electron. This process, called **photoelectric emission** or **photoemission**, is illustrated in Fig. 19.1-1(a) for a metal. An incident photon of energy $h\nu$ releases a free electron from within the partially filled conduction band. A brief delay in the emission is incurred by the interaction of the outgoing electron with the remaining ion, as well as by transport, screening, and scattering effects. The delay time depends on the particular metal and on the photon energy, but it is roughly in the range of 100 attoseconds.

Photoemission from a metal. Energy conservation requires that electrons emitted from below the Fermi level, where they are plentiful, have a maximum kinetic energy

$$E_{\max} = h\nu - \mathbb{W}, \tag{19.1-1}$$

where the **photoelectric work function** $\mathbb{W}$ is the energy difference between the vacuum level and the Fermi level of the metal. Equation (19.1-1) is known as the *Einstein photoemission equation*. Only if the electron initially lies right at the Fermi level can it receive the maximum kinetic energy specified in (19.1-1); the removal of a deeper-lying electron requires additional energy to transport it to the Fermi level, thereby reducing the kinetic energy of the liberated electron. The lowest work function for a metal (Cs) is about 2 eV, so that optical detectors based on the external photoeffect from pure metals are useful in the visible and ultraviolet regions of the spectrum, but not in the infrared.



(a) Metal (b) Semiconductor

Figure 19.1-1 Photoelectric emission (a) from a metal, and (b) from an intrinsic semiconductor. The bandgap energy and electron affinity of the material are denoted E_g and χ, respectively, and $\mathbb{W}$ is the photoelectric work function. All three of these quantities are usually specified in eV.

Photoemission from a semiconductor. Photoelectric emission from an intrinsic semiconductor is portrayed schematically in Fig. 19.1-1(b). Photoelectrons are usually released from the valence band, where electrons are plentiful. The formula analogous to (19.1-1) is

$$E_{\max} = h\nu - \mathbb{W} = h\nu - (E_g + \chi), \tag{19.1-2}$$

where E_g is the bandgap energy and χ is the electron affinity of the material, i.e., the energy difference between the vacuum level and the bottom of the conduction band. The energy $E_g + \chi$ can be as small as 1.4 eV for certain materials (e.g., the multialkali compound NaKCsSb, which forms the basis for the so-called S–20-type photocathode), so that semiconductor photoemissive detectors can operate in the near infrared, as well as in the visible and ultraviolet.

Negative-electron-affinity materials. Furthermore, **negative-electron-affinity** (NEA) semiconductors have been developed in which the conduction-band edge lies above the vacuum level so that $h\nu$ need only exceed E_g for photoemission to occur. This is achieved by depositing a thin n-type or metallic layer on p-type material, which causes the bands to bend at the surface of the material. NEA detectors, such as Cs-coated GaAs, are therefore responsive to slightly longer near-infrared wavelengths, and also exhibit improved quantum efficiency and reduced dark current. Photocathodes constructed from inhomogeneous materials or oxides, such as the S–1-type photocathode, can also be used in the near infrared, but only for wavelengths $\lambda_o \lesssim 1\ \mu$m.

Vacuum photodiodes and photomultiplier tubes. In their simplest form, photodetectors based on photoelectric emission take the form of vacuum tubes called **vacuum photodiodes** or **phototubes**. Electrons are emitted from the surface of a photoemissive material called the **photocathode** and travel to an electrode (anode), which is maintained at a higher electrical potential. The photocathode can be opaque and operate in reflection mode [Fig. 19.1-2(a)], or semitransparent and operate in transmission mode. As a result of the electron transport between the cathode and anode, a current proportional to the photon flux, known as the **photocurrent**, is created in the circuit. The photoemitted electrons may also create a cascade of electrons via the process of **secondary emission** [Fig. 19.1-2(b)]. This occurs when the photoelectrons emitted from the photocathode impact other specially placed cesiated-oxide or semiconductor surfaces in the tube, called **dynodes**, which are maintained at successively higher potentials. A device such as this, known as a **photomultiplier tube (PMT)**, offers low-noise amplification of the generated photocurrent with gains as high as 10^8 (Example 19.6-2).

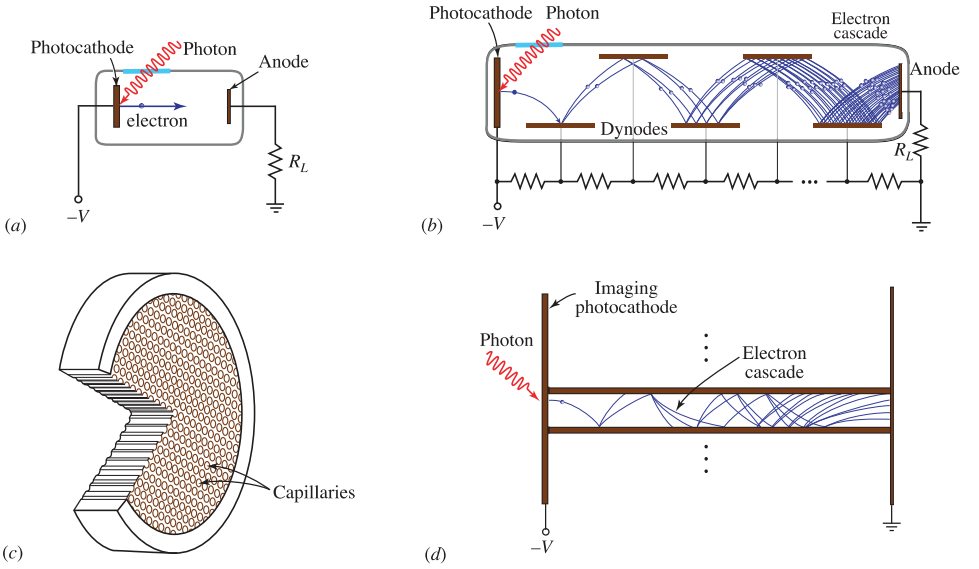


Figure 19.1-2 (a) Photon detection in a vacuum photodiode with a photocathode operated in reflection mode. (b) Photon detection and electron multiplication at the dynodes in a photomultiplier tube (PMT). (c) Cutaway view of a microchannel plate (MCP). (d) Photon detection and electron multiplication in a single capillary of an MCP endowed with a semitransparent imaging photocathode.

Applications of PMTs. Though PMTs usually have modest quantum efficiencies and require high voltages to operate, they find use in many venues. Their high gain, low dark current, low noise, and fast response time endows them with the ability to detect

individual optical photons (Sec. 19.4C). With diameters ranging from millimeters to half a meter, PMTs are used in applications as diverse as oil-well logging (they can operate at high temperatures) and gamma cameras. Gamma and beta rays are readily detected with PMTs by making use of scintillator materials that convert these high-energy particles into visible photons via radioluminescence and betaluminescence, respectively (Sec. 14.5A). The *Super-Kamiokande neutrino-detection experiment* in Japan makes use of more than 11 000 PMTs, each with a diameter of $^1/_2$ m, that carpet the interior walls of an underground tank containing 50 000 tons of water. Neutrinos interact with the constituent atoms of the water to produce charged particles that travel faster than the speed of light in water, thereby generating blue Čerenkov radiation pulses that are detected by the PMTs. At the other end of the size spectrum, μ-PMTs measuring roughly 1-cm^2 in area by 2-mm high, and weighing only about $^1/_2$ g, offer gains approaching 10^7. These devices find use in portable photosensing instruments for applications such as biomedical point-of-care testing, biochemical micro-total analysis, and environmental monitoring.

Microchannel plates. A compact imaging device that makes use of the secondary-emission principle is the **microchannel plate** (**MCP**) displayed in Fig. 19.1-2(c). It consists of an array of millions of capillaries (of internal diameter $\approx$ 10 μm) created in a glass plate (of thickness $\approx$ 1 mm). Both faces of the plate are coated with thin metallic films that act as electrodes, across which a voltage is applied. The interior wall of each capillary is coated with a material that facilitates electron secondary emission so it behaves as a continuous dynode, multiplying the photocurrent generated at that lateral position in the MCP [Fig. 19.1-2(d)]. This allows the local photon flux of a faint image to be converted to a substantial electron-flux image that can be directly measured. Moreover, if desired the electron-flux image can be reconverted into an (amplified) photon-flux image by applying a phosphor coating to the rear electrode that then produces light via cathodoluminescence (Sec. 14.5A); this combination is known as an **image intensifier**.

Photoconductivity

Most modern photodetectors operate on the basis of the internal photoeffect, in which the photoexcited carriers (electrons and holes) remain within the sample. **Photoconductive detectors** rely directly on the light-induced increase in the electrical conductivity of a material. The absorption of a photon by an intrinsic semiconductor, for example, results in the *generation* of a free electron excited from the valence band to the conduction band (Fig. 19.1-3). Concurrently, a hole is generated in the valence band. The application of an electric field to the material results in the *transport* of both electrons and holes through the material and, as a consequence, the production of an electric current in the electrical circuit.

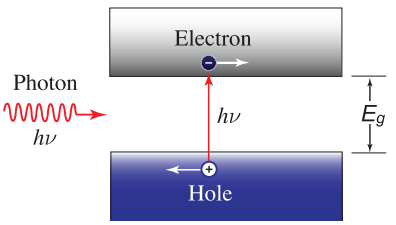


Figure 19.1-3 Electron–hole pair photogeneration in an intrinsic semiconductor.

Semiconductor photodiodes are p–n junction structures that are also based on the internal photoeffect. Photons absorbed in the depletion layer of the device *generate* electrons and holes, which are subjected to the local electric field within that layer. The two carriers drift in opposite directions. This *transport* process induces an electric current in the external circuit.

Some photodetectors also incorporate an internal gain mechanism so that the photocurrent is amplified, thereby making the signal more easily detectable. **Avalanche photodiodes (APDs)** are devices in which an internal *amplification* process takes place via carrier multiplication within the detector. If the depletion-layer electric field in a photodiode is increased sufficiently by applying a large reverse-bias voltage across the junction, the electrons and holes generated may themselves acquire sufficient energy to liberate additional electrons and holes by a process called **impact ionization**, which is the inverse of Auger recombination (Fig. 17.1-18). An APD can be used as an alternative to, or in conjunction with, a laser preamplifier [see, e.g., Fig. 25.1-5(*c*)]. Each of these amplification mechanisms carries its own form of noise, however.

Semiconductor photoelectric detectors with gain therefore involve the following three basic processes:

1. *Generation*: Absorbed photons generate free carriers.
2. *Transport*: An applied electric field causes these carriers to move, resulting in a circuit current.
3. *Gain*: In an avalanche photodiode, a large applied electric fields imparts sufficient energy to the carriers so that they in turn free additional carriers by impact ionization; this internal amplification process enhances the responsivity of the detector, but also introduces noise.

Organic photodetectors. Organic semiconductors are usually either small organic molecules or conjugated polymer chains. The lowest-unoccupied molecular orbital (LUMO) and the highest-occupied molecular orbital (HOMO) may be viewed as analogous to the conduction- and valence-band edges of inorganic semiconductors, respectively (Sec. 17.1B). Photon absorption leads to the generation of charge carriers. Organic photodetectors (OPDs) can be configured as photoconductors, photodiodes, or phototransistors and can operate over a broad range of wavelengths that stretches from the near infrared to the ultraviolet. They are often fabricated as heterojunctions comprising conjugated organic semiconductors with different electron affinities.

Much as with OLEDs (Sec. 18.1E), OPDs offer a number of salutary features: they can be thin, lightweight, mechanically flexible, semitransparent, responsive across large portions of the optical spectrum, and easy to fabricate in large sizes. Printed OPDs can convert flexible substrates (e.g., paper, plastic, or glass) into smart surfaces and can be fashioned into wearable biomedical devices. In the current state of their development, however, OPD parameters such as dark current, responsivity, and life span are generally inferior to those of silicon-based photodetectors, although not substantially so.

B. General Properties

Certain general features are associated with all photodetectors. Before considering the details of specific photoelectric detectors of interest in photonics, we examine three such general features: *quantum efficiency*, *responsivity*, and *response time*. Photodetectors and semiconductor sources are inverse devices and these three features also have their counterparts in the domain of semiconductor sources (Secs. 18.1B and 18.3B). Indeed, the same materials are often used to fabricate semiconductor photodetectors and semiconductor sources (see, e.g., Figs. 17.1-7, 17.1-8, and 18.1-16).

Quantum Efficiency

The **quantum efficiency** η $(0 \leq \eta \leq 1)$ of a semiconductor photodetector is the probability that a single photon incident on the device generates a photocarrier pair that contributes to the detector current. When many photons are incident, as is usually the

case, η becomes the flux of generated electron–hole pairs that contribute to the detector current divided by the flux of incident photons. For non-semiconductor photodetectors, the **photon detection efficiency (PDE)** η is defined as the probability that a single photon incident on the device generates a detectable electrical current in the output circuit. For detectors that operate on the basis of the external photoeffect, such as the vacuum photodiode and the photomultiplier tube, the quantum efficiency η is the probability that a single incident photon generates a free photoelectron.

Not all incident photons produce electron–hole pairs in a semiconductor photodetector because not all of them reach the photosensitive region and are absorbed. As illustrated in Fig. 19.1-4, some of the photons are reflected at the surface of the detector via Fresnel reflection while others fail to be absorbed because the photosensitive material has insufficient thickness (the rate of photon absorption in a semiconductor material was considered in Sec. 17.2C). Furthermore, some electron–hole pairs produced near the photodetector surface quickly recombine because of the abundance of recombination centers at surfaces, and are therefore not available to contribute to the detector current.

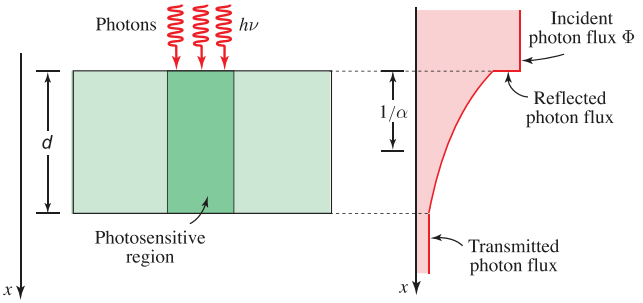


Figure 19.1-4 Effect of surface reflection and incomplete absorption (arising from an insufficient thickness of photosensitive material) on the detector quantum efficiency η.

For semiconductor photoelectric detectors, the quantum efficiency η can therefore be written as

$$\eta = (1 - \mathcal{R}) \cdot \zeta \cdot [1 - \exp(-\alpha d)] , \tag{19.1-3}$$

Photoelectric-Detector Quantum Efficiency

where $\mathcal{R}$ is the power reflectance at the surface of the photodetector, ζ is the fraction of electron–hole pairs that successfully contributes to the detector photocurrent, α is the absorption coefficient of the photosensitive material (cm^{-1}) discussed in Sec. 17.2C, and d is the thickness of the photosensitive region. Equation (19.1-3) is thus a product of three factors:

- The first factor, $\mathcal{T} = (1 - \mathcal{R})$, represents the power transmittance at the surface of the device. The transmittance can be increased, for example, by the use of antireflection coatings. Some definitions of the quantum efficiency η exclude the effects of reflection at the surface, which must then be considered separately.
- The second factor ζ is the fraction of electron–hole pairs that successfully avoid recombination at the material surface so they can potentially contribute to the useful photocurrent. Surface recombination can be reduced by careful material growth and device design.
- The third factor, $\int_0^d e^{-\alpha x}\, dx / \int_0^\infty e^{-\alpha x}\, dx = [1 - \exp(-\alpha d)]$, represents the fraction of the photon flux absorbed in the bulk of the photosensitive material. The device should have a value of d that is sufficiently large so this factor is maximized, subject to other constraints.

Additional loss is also incurred if the light is not properly focused onto the photosensitive region of the detector.

Dependence of quantum efficiency on wavelength. The quantum efficiency η is a function of wavelength principally because the absorption coefficient α is wavelength dependent (Fig. 17.2-3). The characteristics of the semiconductor material thus determine the spectral window within which η is large. For sufficiently large values of the free-space wavelength λ_o, η is small because absorption cannot occur for $\lambda_o \geq \lambda_g = hc_o/E_g$ (the photon energy is then smaller than the bandgap energy and the material is transparent). The bandgap wavelength λ_g is thus the **long-wavelength photodetection limit** for a semiconductor material. Representative values of E_g and λ_g are presented in Table 17.1-2 and displayed in Figs. 17.1-7 and 17.1-8 for representative semiconductor materials of interest in photonics. For sufficiently small values of λ_o, η also decreases because most photons are then absorbed near the surface of the device (for $\alpha = 10^4$ cm^{-1}, for example, most of the light is absorbed within a distance $1/\alpha = 1$ μm). The recombination lifetime is quite short near the surface, so the photocarriers recombine before being collected.

Resonant-cavity photodetectors. The quantum efficiency η may be enhanced by constructing a detector configuration in which the light can interact with the photosensitive material on multiple passes. This is equivalent to increasing the photodetector depth d, which increases the absorption and reduces the transmitted photon flux. This may be achieved in practice by placing the photodetector inside a resonant cavity, which traps the light and thus increases the quantum efficiency, but it does so at the expense of restricting the bandwidth and extending the response time.

Plasmonic photodetectors. Another approach for augmenting the quantum efficiency η of a semiconductor photodetector relies on endowing the photosensitive material with metallic nanostructures that have the ability to scatter, concentrate, and guide light at the nanoscale, as discussed in Chapter 8:

- The presence of metallic nanoparticles at the upper surface of the semiconductor material can result in enhanced trapping of the incident light in the photosensitive layer via high-angle, multiple scattering. Increasing the effective optical path length in this manner is especially useful for structures such as thin-film solar cells. The scattering of light by metallic nanoparticles is detailed in Sec. 8.2C.
- Metallic nanoparticles embedded in the interior of the semiconductor material can serve as resonant optical antennas that trap and concentrate light in the form of localized surface plasmons. Such modes generate near-field radiation, such as optical dipole waves, that in turn can produce electron–hole pairs. The properties of optical antennas are discussed in Sec. 8.2D.
- A corrugated metallic surface placed at the lower surface of the semiconductor material can trap light in the form of surface plasmon polaritons (SPPs) that propagate in the plane of the photosensitive layer. SPPs at a metal–dielectric boundary are considered in Sec. 8.2B.

Responsivity

The **responsivity** of a photodetector relates the electric current i_p flowing in the device circuit to the optical power P incident on it. If every photon were to generate a photocarrier pair in the device, a photon flux Φ (photons per second) would produce an electron flux (electrons per second) in the photodetector circuit that corresponds to a short-circuit electric current $i_p = e\Phi$. Thus, an optical power $P = h\nu\Phi$ (watts) at optical frequency ν would give rise to an electric current $i_p = eP/h\nu$.

However, since the fraction of photons that produces detected electrons is η rather than unity, the electric current is

$$i_p = \eta e\Phi = \frac{\eta eP}{h\nu} \equiv \mathsf{R}P. \tag{19.1-4}$$

The proportionality factor between the electric current and the optical power, $\mathrm{R} = i_p/P$, has units of A/W and is called the photodetector responsivity:

$$\boxed{\mathrm{R} = \frac{\eta e}{h\nu} = \eta \frac{\lambda_o}{1.24}\,.} \qquad (19.1\text{-}5)$$

Photodetector Responsivity
(A/W; λ_o in μm)

It is important to distinguish the photodetector responsivity (A/W) from the LED responsivity (W/A) defined in (18.1-29).

The responsivity is linearly proportional to both the quantum efficiency η and the free-space wavelength λ_o, as is evident from (19.1-5) and Fig. 19.1-5. An appreciation for the order of magnitude of the responsivity is gained by setting $\eta = 1$ and $\lambda_o = 1.24$ μm in (19.1-5), whereupon $\mathrm{R} = 1$ A/W $= 1$ nA/nW.

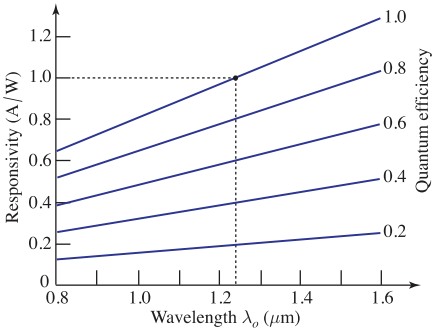


Figure 19.1-5 Responsivity R (A/W) versus wavelength λ_o, with the quantum efficiency η as a parameter. For $\eta = 1$, the responsivity is $\mathrm{R} = 1$ A/W at $\lambda_o = 1.24$ μm.

The proportionality of R to λ_o is a consequence of the fact that the responsivity is defined on the basis of optical power, whereas most photodetectors generate currents proportional to the photon flux Φ. For a given photon flux $\Phi = P/h\nu = P\lambda_o/hc_o$ (corresponding to a given photodetector current i_p), the product $P\lambda_o$ is fixed so that an increase in λ_o requires a commensurate decrease in P, thereby leading to an increase in the responsivity. Indeed, some thermal detectors are responsive to optical power rather than to photon flux, causing R to be independent of λ_o.

The region over which R increases with λ_o is limited, however, inasmuch as the wavelength dependence of η comes into play at both long and short wavelengths, as discussed earlier. The responsivity can also be degraded if the detector is presented with an excessively large optical power. This condition, known as **detector saturation**, limits the **linear dynamic range** of the detector, which is the range over which it responds to the incident optical power in a linear fashion.

Devices with gain. The formulas presented above are predicated on the assumption that each photocarrier pair produces a charge e in the photodetector circuit. However, devices that exhibit gain can produce a charge q in the circuit that differs from e. The gain G is defined as the number of circuit electrons generated per photocarrier pair,

$$G \equiv q/e. \qquad (19.1\text{-}6)$$

The gain can be either greater than or less than unity, as will be seen subsequently.

In the presence of gain, the formulas for the photocurrent and responsivity presented in (19.1-4) and (19.1-5), respectively, must be modified. Substituting $q = Ge$ for e in

these equations, respectively, yields

$$i_p = \eta q\Phi = \eta G e\Phi = \frac{\eta G e P}{h\nu} \tag{19.1-7}$$

Photocurrent with Gain

and

$$\mathsf{R} = \frac{\eta G e}{h\nu} = \eta G \frac{\lambda_o}{1.24}\,. \tag{19.1-8}$$

Responsivity with Gain (A/W; λ_o in μm)

The gain of the device G is to be distinguished from the photodetector quantum efficiency η, which is the probability that an incident photon produces a detectable photocarrier pair. Other useful measures of photodetector behavior, such as signal-to-noise ratio and receiver sensitivity, await discussion of detector noise properties, presented in Sec. 19.6.

Response Time

Transit-time spread. A constant electric field E applied to a semiconductor (or metal) causes its free charge carriers to accelerate. In the course of doing so, they encounter frequent collisions with lattice ions moving about their equilibrium positions via thermal motion, as well as imperfections in the crystal lattice associated with impurity ions. These collisions cause the carriers to suffer random decelerations; the result is motion at an average velocity rather than at a constant acceleration. The mean velocity of a carrier is given by $\mathsf{v} = \mathtt{a}\tau_{\rm col}$, where $\mathtt{a} = eE/m$ is the acceleration imparted by the electric field and $\tau_{\rm col}$ is the mean time between collisions, which serves as a relaxation time. The net result is that the carrier drifts in the direction of the electric field with a mean **drift velocity** $\mathsf{v} = e\tau_{\rm col}E/m$, which is conventionally written in the form

$$\mathsf{v} = \mu E, \tag{19.1-9}$$

where $\mu = e\tau_{\rm col}/m$ is known as the carrier **mobility**.

Ramo's theorem. The carrier motion in the photodetector creates a current in its external circuit. To determine the magnitude of the current $i(t)$, consider an electron–hole pair generated (by photon absorption, for example) at an arbitrary position x in a semiconductor material of length w, to which a voltage V is applied, as shown in Fig. 19.1-6(a). We restrict our attention to motion in the x direction and invoke an energy argument. If a carrier of charge Q (a hole of charge $Q = e$ or an electron of charge $Q = -e$) moves a distance dx in the time dt under the influence of an electric field of magnitude $E = V/\mathsf{w}$, the work done is $-QE\,dx = -Q(V/\mathsf{w})\,dx$. This work must equal the energy provided by the external circuit, $i(t)V\,dt$. Thus, $i(t)V\,dt = -Q(V/\mathsf{w})\,dx$, which leads to $i(t) = -(Q/\mathsf{w})(dx/dt) = -(Q/\mathsf{w})\mathsf{v}(t)$. A carrier moving with drift velocity $\mathsf{v}(t)$ in the x direction therefore creates a current in the external circuit characterized by **Ramo's theorem**:

$$i(t) = -\frac{Q}{\mathsf{w}}\mathsf{v}(t). \tag{19.1-10}$$

Ramo's Theorem

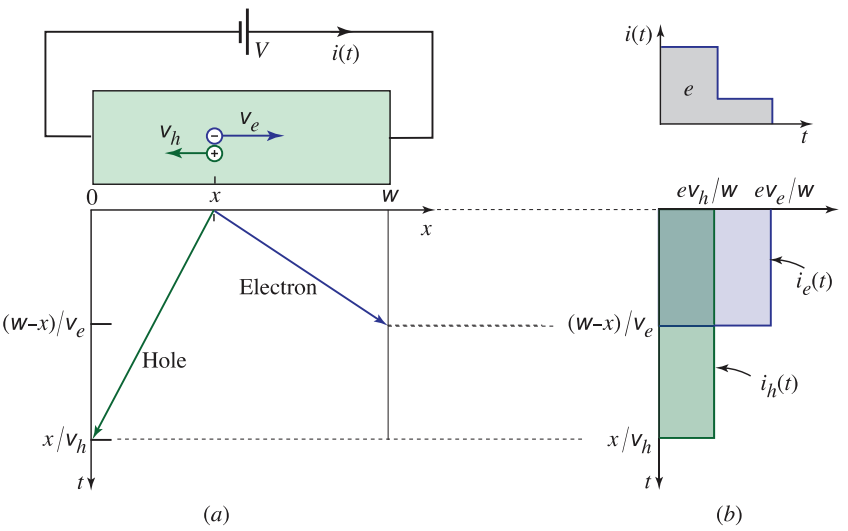


Figure 19.1-6 (a) An electron–hole pair is generated at the position x. The hole drifts to the left with constant velocity v_h and the electron drifts to the right with constant velocity v_e. The process terminates when the carriers reach the edges of the material. (b) The hole current $i_h(t)$, electron current $i_e(t)$, and total current $i(t)$ induced in the circuit. The total charge induced in the circuit per carrier pair is e.

Assuming that the hole moves with constant velocity v_h to the left, and the electron moves with constant velocity v_e to the right, (19.1-10) provides that the hole current is $i_h = -(e)(-v_h)/w = ev_h/w$ and the electron current is $i_e = -(-e)v_e/w = ev_e/w$, as illustrated in Fig. 19.1-6(b). Each carrier contributes to the current as long as it is moving. If the carriers continue their motion until they reach the edges of the material, the hole moves for a duration x/v_h while the electron moves for a duration $(w-x)/v_e$ [Fig. 19.1-6(a)]. In semiconductors, v_e is generally larger than v_h so that the overall duration of the response is x/v_h. The finite duration of the current is known as the **transit-time spread**; it is an important limiting factor for the speed of operation (bandwidth) of all semiconductor photodetectors.

Charge generated in external circuit. One might be inclined to argue that the charge generated in an external circuit should be $2e$ when a photon generates an electron–hole pair in a photodetector material, since there are two charge carriers. In fact, the charge generated is e, as is demonstrated by calculating the total charge q induced in the external circuit from the sum of the areas under i_e and i_h:

$$q = e\frac{v_h}{w}\frac{x}{v_h} + e\frac{v_e}{w}\frac{w-x}{v_e} = e\left(\frac{x}{w} + \frac{w-x}{w}\right) = e. \tag{19.1-11}$$

This result is independent of the position x at which the electron–hole pair was created.

Uniform generation of carrier pairs. The transit-time spread is far more severe if the electron–hole pairs are generated uniformly throughout the material rather than at a single point x, as can be understood from Fig. 19.1-7 and Prob. 19.1-4. For $v_h < v_e$, the full width of the transit-time spread is then w/v_h rather than x/v_h. This occurs because uniform illumination produces carrier pairs at all locations, including at $x = w$, the location at which the holes have the farthest to travel before being able to recombine at $x = 0$.

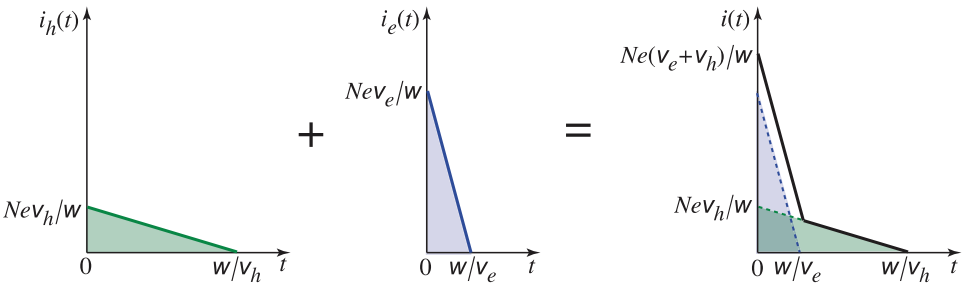


Figure 19.1-7 Hole current $i_h(t)$, electron current $i_e(t)$, and total current $i(t)$ induced in the photodetector circuit for electron–hole generation by N photons uniformly distributed between 0 and w (Prob. 19.1-4). The tail in the total current results from the motion of the slow holes. The total current $i(t)$ can be viewed as the impulse response function (see Appendix B, Sec. B.1) of a uniformly illuminated detector subject to transit-time spread.

Summary

Ramo's theorem demonstrates that the charge delivered to the external circuit of a semiconductor photodetector by carrier motion within the photodetector material is not provided instantaneously, but rather occupies an extended time. It is as if the motion of the charge carriers in the material pulls charge slowly from the wire on one side of the device and pushes it slowly into the wire on the other side, so that each charge passing through the external circuit is spread out in time.

Ohm's law. In the presence of a uniform charge density ϱ, rather than a single point charge Q, the total charge in the photodetector material is ϱAw, where A is the cross-sectional area [Fig. 19.1-6(a)]. From (19.1-10), the current density in the x direction is then $J(t) = i(t)/A = -(\varrho Aw/Aw)v(t) = -\varrho v(t)$. The well-known vector form of this equation is

$$\boxed{\mathcal{J} = \varrho\,\mathbf{v}.} \qquad \begin{matrix}(19.1\text{-}12)\\ \text{Current Density}\end{matrix}$$

Combining (19.1-12) with (19.1-9) yields $J = \sigma E$, where σ is the **conductivity** of the medium,

$$\sigma = \varrho\,\mu = e\varrho\,\tau_{\text{col}}/m = Ne^2\,\tau_{\text{col}}/m; \qquad (19.1\text{-}13)$$

N is the number of carriers per unit volume [see (8.2-3) and (8.2-17)].

More generally, **Ohm's law** takes the form of a relationship between the current-density and electric-field vectors, $\mathcal{J}$ and $\mathcal{E}$, respectively, mediated by the second-rank **conductivity tensor** $\boldsymbol{\sigma}$:

$$\boxed{\mathcal{J} = \boldsymbol{\sigma}\mathcal{E}\,.} \qquad \begin{matrix}(19.1\text{-}14)\\ \text{Ohm's Law}\end{matrix}$$

For charge carried by a homogeneous conductive medium with cross-sectional area A and length w, $J = \sigma E$ can be written as $i = (\sigma A/w)Ew = (\sigma A/w)V = GV = V/R$, where G and R are the conductance and resistance of the material, respectively. In this configuration, Ohm's law takes its beloved form

$$V = iR\,. \qquad (19.1\text{-}15)$$

***RC* time constant.** The resistance R and capacitance C of the photodetector, along with that of its circuitry, give rise to another response time called the ***RC* time constant**, $\tau_{RC} = RC$. The resistance/capacitance combination serves to integrate the current at the photodetector output, thereby increasing the duration of the impulse response function. In the presence of transit-time *and* an RC time-constant, the overall impulse response function is determined by convolving the current $i(t)$ displayed in Fig. 19.1-7 with the exponential function $(1/RC)\exp(-t/RC)$ (Sec. B.1). It is worthy of note that photodetectors of different types may exhibit other specific limitations on their bandwidths, which require consideration on a case-by-case basis. As a final point, we mention that photodetectors fabricated with a given material and structure sometimes exhibit a fixed gain–bandwidth product, in which case increasing the gain results in a decrease of the bandwidth and *vice versa*. This tradeoff between gain and frequency response is associated with the time required for the gain process to take place.

19.2 PHOTOCONDUCTORS

When photons are absorbed in a semiconductor, mobile charge carriers are *generated* (ideally an electron–hole pair for every absorbed photon). The electrical conductivity of the material σ increases in proportion to the photon flux Φ. An electric field applied to the material by an external voltage source causes the electrons and holes to be *transported*. This in turn results in a measurable electric current in the circuit, as illustrated in Fig. 19.2-1(a). **Photoconductive detectors** operate by registering either the photocurrent i_p, which is proportional to the photon flux Φ, or the voltage drop across a load resistor R placed in series with the circuit.

A. Intrinsic Photoconductors

If the photon energy is greater than the bandgap of the semiconductor, photons are absorbed via interband transitions [Fig. 17.2-4(a)]. A photoconductive device may take the form of a slab or a thin film. The anode and cathode contacts are often interdigitated on the same surface of the device to minimize the transit time [Fig. 19.2-1(b)]. At the same time, the photon flux reaching the photosensitive material can be maximized by directing the light to the opposite surface if the insulating substrate has a sufficiently large bandgap so that it is not absorptive.

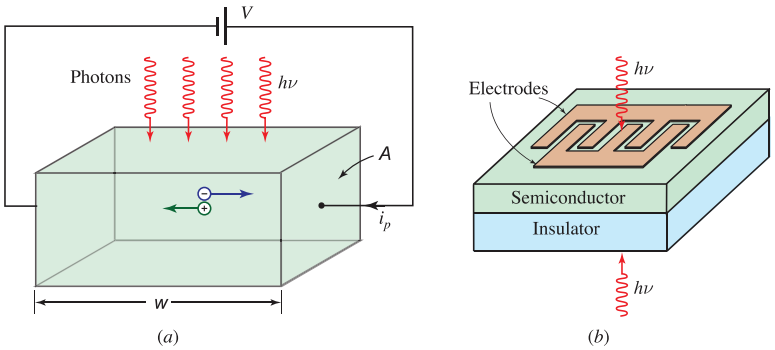


Figure 19.2-1 The photoconductive detector. (a) Photogenerated carrier pairs move in response to the applied voltage V, generating a photocurrent i_p proportional to the incident photon flux Φ. (b) The interdigitated electrode structure of this device is designed to minimize the carrier transit time (and thereby maximize the bandwidth) while maximizing the light reaching the photosensitive material.

The increase in conductivity arising from a photon flux Φ (photons per second) illuminating a semiconductor volume wA [Fig. 19.2-1(a)] is calculated as follows. A fraction η of the incident photon flux is absorbed and gives rise to excess electron–hole pairs. The pair-production rate R (per unit volume per unit time) is thus $R = \eta\Phi/wA$. If τ is the excess-carrier recombination lifetime, in accordance with (17.1-23) electrons are lost at the rate $\Delta\mathrm{n}/\tau$ where $\Delta\mathrm{n}$ is the electron concentration. Under steady-state conditions both rates are equal so that $R = \Delta\mathrm{n}/\tau$ and $\Delta\mathrm{n} = \eta\tau\Phi/wA$. The increase in the carrier concentration $\Delta\mathrm{n}$ is accompanied by an increase in the charge density $\Delta\varrho = e\Delta\mathrm{n}$, and thus, in accordance with (19.1-13), by an increase in the conductivity $\Delta\sigma = \mu\Delta\varrho = e\mu\Delta\mathrm{n}$, so that

$$\Delta\sigma = \frac{\eta e\tau(\mu_e + \mu_h)}{wA}\,\Phi, \tag{19.2-1}$$

where μ_e and μ_h are the electron and hole mobilities, respectively. In accordance with (19.2-1), the increase in conductivity is proportional to the photon flux, as expected.

Ohm's law (19.1-14) dictates that the photogenerated current density is given by $J_p = \Delta\sigma\, E$. Combining this with (19.2-1) and (19.1-9), which provides $v_e = \mu_e E$ and $v_h = \mu_h E$, gives $J_p = [\eta e\tau(v_e + v_h)/wA]\,\Phi$, which corresponds to an electric current $i_p = J_p A = [\eta e\tau(v_e + v_h)/w]\,\Phi$. If $v_h \ll v_e$, and if the formula is cast in terms of the electron transit time across the sample $\tau_e = w/v_e$, we finally obtain

$$i_p \approx \eta(\tau/\tau_e)e\Phi. \tag{19.2-2}$$

Comparison with (19.1-7) shows that the ratio τ/τ_e in (19.2-2) corresponds to the detector gain G, for reasons we now proceed to elucidate.

Gain. The responsivity of a photoconductor with gain is given by (19.1-8). Simply viewed, the device exhibits internal gain because the excess-carrier recombination lifetime τ and the transit time τ_e differ in general. Suppose that electrons travel faster than holes and that τ is very long. As the electron and hole are transported to opposite sides of the photoconductor (Fig. 19.2-1), the electron completes its trip sooner than the hole. The requirement of current continuity then forces the external circuit to immediately provide another electron, which enters the device from the wire at the left. This new electron moves quickly toward the right, again completing its trip before the hole reaches the left edge (or is released from a trap). This process continues until the electron recombines with the hole, which requires a long time τ.

A single photon absorption can therefore result in an electron passing through the external circuit many times. The expected number of trips that the electron makes before the process terminates is

$$G = \tau/\tau_e, \tag{19.2-3}$$

where τ is the excess-carrier recombination lifetime and $\tau_e = w/v_e$ is the electron transit time across the sample. The charge delivered to the circuit by a single electron–hole pair is then $q = Ge > e$ so that the device exhibits gain.

At the other extreme, the recombination lifetime may be sufficiently short such that the carriers recombine before reaching the edge of the material. This can occur if there is a ready availability of carriers of the opposite type for recombination to take place. In that case $\tau < \tau_e$ and the gain is less than unity so that, on average, each carrier pair contributes only a fraction of the electronic charge e to the circuit. Charge is, of course, conserved so that the many carrier pairs present deliver an integral number of electronic charges to the circuit.

The photoconductor gain $G = \tau/\tau_e$ can therefore be interpreted as the fraction of the sample length traversed by the average excited carrier before it undergoes recombination. The transit time τ_e is determined from the length of the device and the applied

voltage via (19.1-9) and $\tau_e = w/v_e$; typical values of $w = 1$ mm and $v_e = 10^7$ cm/s yield $\tau_e \approx 10^{-8}$ s. The recombination lifetime τ can range from 10^{-13} s to many seconds, depending on the photoconductor material and doping [see (17.1-24)]. Thus, the gain G can assume a broad range of values, stretching from below unity to well above unity, depending on the parameters of the material, the size of the device, and the applied voltage. However, the gain of a photoconductor generally cannot exceed 10^6 because of the restrictions imposed by space-charge-limited current flow, impact ionization, and dielectric breakdown.

The process of electron–hole recombination is actually random so that the implicit assumption of deterministic photoconductor gain invoked above can be too simplistic. In those circumstances, a more realistic model must be used (Prob. 19.6-4).

Spectral response. The spectral sensitivity of a photoconductive detector is governed principally by the wavelength dependence of the quantum efficiency η, as discussed in Sec. 19.1B. Different semiconductors have different long-wavelength photodetection limits λ_g (Table 17.1-2). Interband transitions in elemental, binary, and ternary semiconductor photoconductors allow operation into the mid infrared (in contrast to photoemissive detectors) but their use at wavelengths beyond about 2 μm generally requires cooling to minimize the thermal generation of electron–hole pairs.

Response time. The response time of a photoconductive detector is constrained by the transit-time and RC time-constant considerations discussed in Sec. 19.1B. The carrier-transport response time is approximately equal to the recombination time τ, so that the carrier-transport bandwidth B is inversely proportional to τ. Since the gain G is directly proportional to τ in accordance with (19.2-3), increasing τ serves to increase the gain, which is desirable, but concomitantly decreases the bandwidth, which is undesirable. The gain–bandwidth product GB thus turns out to be roughly independent of τ; values of GB can extend up to $\approx 10^{12}$.

B. Extrinsic Photoconductors

Photoconductivity can be achieved at wavelengths that extend to the far infrared by making use of doped semiconductors. Mobile charge carriers can be generated via photon absorption by dopants with energy levels lying within the forbidden gap. The process can occur in one of two ways: (1) an incident photon interacts with a bound electron at a donor site, frees it to the conduction band, and leaves behind a bound hole; or (2) an incident photon interacts with a bound hole at an acceptor site, frees it to the valence band, and leaves behind a bound electron, as illustrated in Fig. 17.2-1(b). Donor and acceptor levels in the bandgap of doped semiconductors can have very low activation energies E_A, and therefore quite extended long-wavelength limits $\lambda_A = hc_o/E_A$. These detectors must be cooled to avoid thermal excitation; liquid He at 4° K is often used. Representative values of E_A and λ_A are provided in Table 19.2-1 for a number of extrinsic photoconductive detectors.

Table 19.2-1 Selected extrinsic semiconductor materials with their activation energies E_A and long-wavelength limits $\lambda_A = hc_o/E_A$.

Semiconductor:Dopant	E_A (eV)	λ_A (μm)
Ge:Hg	0.088	14
Ge:Cu	0.041	30
Ge:Zn	0.033	38
Ge:Ga	0.010	115
Si:B	0.044	23

The relative responsivities of several extrinsic photoconductive detectors are illustrated in Fig. 19.2-2. For all of these materials, the responsivity increases approximately linearly with λ_o, in accordance with (19.1-8), peaks slightly below the long-wavelength limit λ_A, and falls off rapidly beyond it. The quantum efficiencies of these detectors can be substantial (e.g., $\eta \approx 0.5$ for Ge:Cu), although the gain can be low under usual operating conditions (e.g., $G \approx 0.03$ for Ge:Hg).

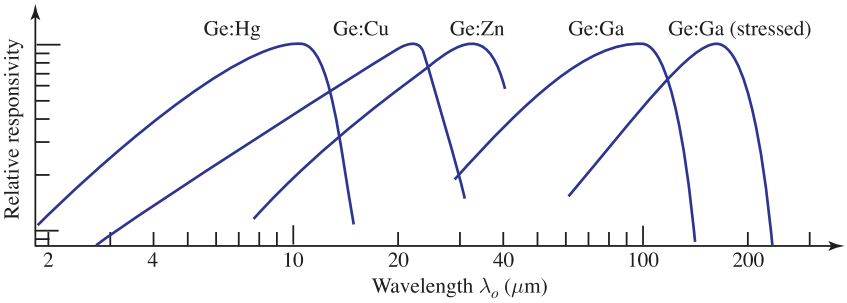


Figure 19.2-2 Relative responsivity vs. wavelength λ_o (μm) for a number of different doped-Ge extrinsic materials used as infrared photoconductive detectors.

C. Heterostructure Photoconductors

Properly configured heterostructures can serve as useful photoconductive detectors. An example is the **quantum-well infrared photodetector** (**QWIP**). An incident infrared photon releases the electron occupying a bound energy level in a quantum-well to the continuum, thereby creating a mobile charge carrier that increases the conductivity of the material (Fig. 19.2-3). In an alternate configuration, the quantum wells are situated between superlattice barriers and the electrons are swept out via a miniband transport channel lying below the continuum [Fig. 17.2-10(d)]. QWIPs typically have sharp spectral responses dictated by the narrowness of the quantized states and operate at cryogenic temperatures.

The **quantum-dot infrared photodetector** (**QDIP**), a variation on this theme, can also be used for multiwavelength infrared detection via intersubband transitions. The **dot-in-well QDIP** (**DWELL-QDIP**) offers a further improvement in performance by imposing constraints on the locations of the quantum dots.

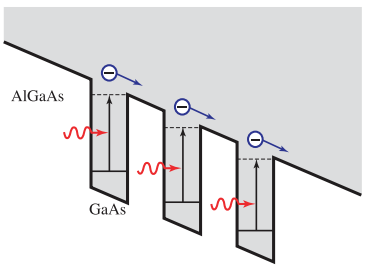


Figure 19.2-3 Generation of mobile charge carriers by absorption of photons in a QWIP. The detector illustrated comprises AlGaAs barriers and n-type GaAs quantum wells that provide the electrons occupying the energy levels. The device is configured in such a way that there is a single energy level in each well, whose parameters are adjusted to provide sensitivity at a particular central wavelength. Though they require cooling, QWIPs fabricated from III–V compound semiconductors offer high speeds and high responsivities from mid- to far-infrared wavelengths ($\lambda_o \approx 3$–20μm). They are often used in focal-plane arrays (Sec. 19.5).

19.3 PHOTODIODES

A. The *p–n* Photodiode

As with photoconductors, **photodiode detectors** rely on photogenerated charge carriers for their operation. A photodiode is a p–n junction (Sec. 17.1E) whose reverse current increases when it absorbs photons. Though p–n and p–i–n (**PIN**) photodiodes are generally faster than photoconductors, they do not exhibit gain.

Consider a reverse-biased p–n junction under illumination, as depicted in Fig. 19.3-1. Photons are absorbed everywhere with absorption coefficient α. Whenever a photon is absorbed, an electron–hole pair is *generated*. But only at locations where an electric field is present can the charge carriers be *transported* in a particular direction. Since a p–n junction can support an electric field only in the depletion layer, this is the region in which it is most desirable to generate photocarriers.

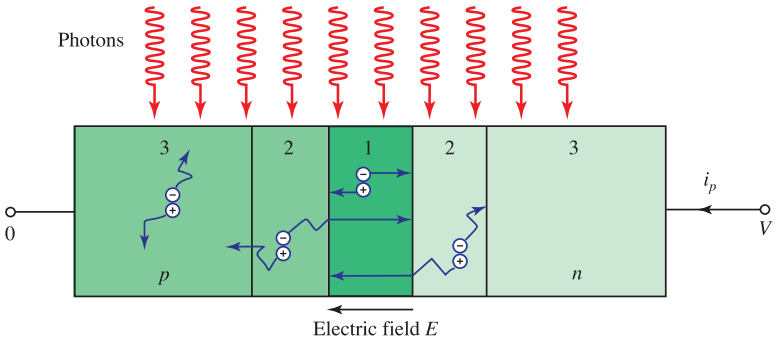


Figure 19.3-1 Photons illuminating an idealized reverse-biased p–n photodiode detector. The drift and diffusion regions are indicated by 1 and 2, respectively. Carriers generated beyond the diffusion region, in 3, fail to contribute to the photocurrent. Illumination can be directed parallel to the junction layer (edge illumination), as illustrated, or at normal incidence to the junction layer.

There are, nevertheless, three possible locations where electron–hole pairs can be generated:

- Electrons and holes generated in the depletion layer (region 1) quickly drift in opposite directions under the influence of the strong electric field. Since the electric field always points in the $n \rightarrow p$ direction, electrons move to the n side and holes to the p side. As a result, the photocurrent created in the external circuit is always in the reverse direction (from the n to the p region). Each carrier pair generates in the external circuit an electric current pulse of area e $(G = 1)$ since recombination does not take place in the depleted region.
- Electrons and holes generated well away from the depletion layer (region 3) cannot be transported because of the absence of an electric field. They wander randomly until they are annihilated by recombination. They do not contribute a signal to the external electric current.
- Electron–hole pairs generated outside the depletion layer, but in its vicinity (region 2), have a chance of entering the depletion layer by random diffusion. An electron coming from the p side is quickly transported across the junction and therefore contributes a charge e to the external circuit. A hole coming from the n side has a similar effect.

Photodiodes have been fabricated from many of the semiconductor materials listed in Table 17.1-2, as well as from binary, ternary, and quaternary compound semiconductors such as InGaAs, InGaAsP, SiC, and GeSn. Devices can be constructed so that

light is directed parallel to the junction layer via edge illumination, as illustrated in Fig. 19.3-1, or at normal incidence to the junction layer.

Response Time

The transit time of carriers drifting across the depletion layer (w_d/v_e for electrons and w_d/v_h for holes) and the RC time constant both play a role in the response time of photodiode detectors, as discussed in Sec. 19.1B. The resulting circuit current is displayed in Fig. 19.1-6(b) for an electron–hole pair generated at a given position x, and in Fig. 19.1-7 for uniform electron–hole pair generation.

In photodiodes there is an additional contribution to the response time arising from diffusion, which is a relatively slow process in comparison with drift. Carriers generated outside the depletion layer, but sufficiently close to it, take some time to diffuse into it, where they contribute to the current. The maximum times allowed for this process are the carrier lifetimes (τ_p for electrons in the p region and τ_n for holes in the n region). The deleterious effects of diffusion time can be diminished by making use of p–i–n photodiodes, as will be seen in Sec. 19.3B. Nevertheless, photodiodes are generally faster than photoconductors because of the large velocity of the photogenerated carriers imparted by the strong field in the depletion region. Furthermore, photodiodes are not affected by many of the trapping effects associated with photoconductors.

Modes of Operation

From the perspective of an electronic device, the photodiode i–V relation is given by

$$i = i_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right] - i_p\,, \tag{19.3-1}$$

as illustrated in Fig. 19.3-2. This is the usual i–V relation for a p–n junction provided in (17.1-32) with an added term for the photocurrent $-i_p$ that is proportional to the photon flux Φ.

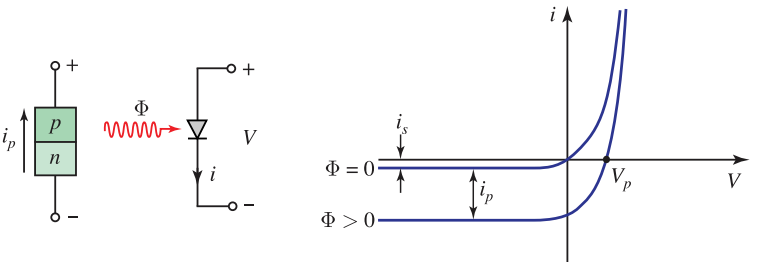


Figure 19.3-2 Generic photodiode and its i–V relation.

There are three classical modes in which photodiodes are operated: open-circuit (photovoltaic), short-circuit, and reverse-biased (photoconductive).

Open-circuit mode. In the open-circuit mode (Fig. 19.3-3), the light generates electron–hole pairs in the depletion region. The additional electrons freed on the n side of the layer recombine with holes on the p side, and *vice versa*. The net result is an increase in the electric field, which produces a photovoltage V_p across the device that increases with increasing photon flux Φ. Since it is operating as an open-circuit device ($i = 0$), the responsivity of a photodiode operating in **photovoltaic mode** is measured in V/W rather than A/W. This is the mode of operation used in solar cells.

Short-circuit mode. The short-circuit ($V = 0$) mode is illustrated in Fig. 19.3-4. The short-circuit current is simply the photocurrent $-i_p$.

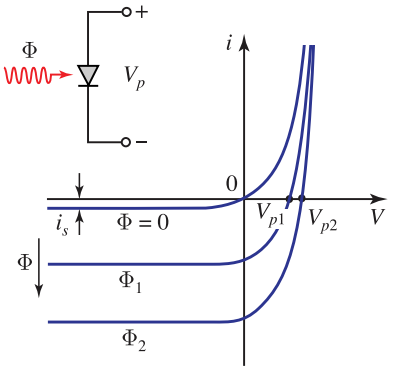


Figure 19.3-3 Photovoltaic (open-circuit) operation of a photodiode.

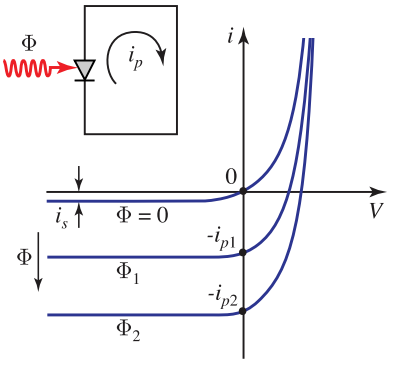


Figure 19.3-4 Short-circuit operation of a photodiode.

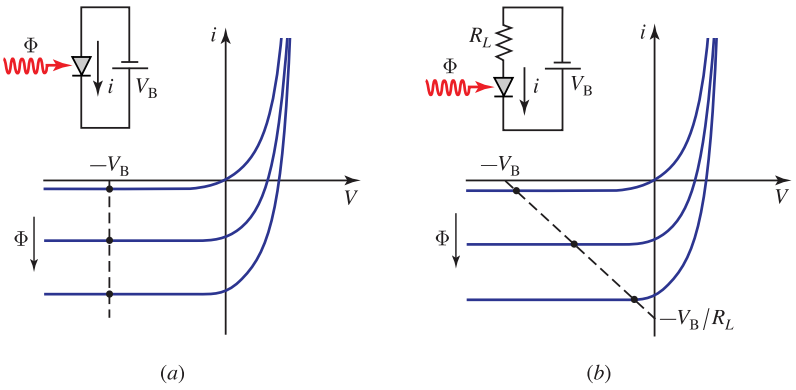


Figure 19.3-5 Reverse-biased operation of a photodiode: (*a*) without a load resistor and (*b*) with a load resistor. The operating point lies on the dashed line.

Reverse-biased mode. Lastly, a photodiode may be operated in its reverse-biased or "photoconductive" mode, as portrayed in Fig. 19.3-5(*a*). If a series load resistor is inserted in the circuit, the operating conditions are those shown in Fig. 19.3-5(*b*). Photodiodes are usually operated in a strongly reverse-biased mode for the following reasons:

- A strong reverse bias creates a strong electric field in the junction region that increases the drift velocity of the carriers, thereby reducing transit time.
- A strong reverse bias increases the width of the depletion layer, which reduces the junction capacitance and improves the response time.
- The increased width of the depletion layer offers a larger photosensitive area, facilitating the collection of more light.

B. The *p–i–n* Photodiode

As a detector, the *p–i–n* (PIN) photodiode has a number of advantages over the *p–n* photodiode. A *p–i–n* diode is a *p–n* junction with an intrinsic (often unintentionally or lightly doped) layer sandwiched between the *p* and *n* layers (Sec. 17.1E). It can be operated under the various bias conditions considered for the *p–n* photodiode in the preceding section. The energy-band diagram, charge distribution, and electric-

field distribution for a reverse-biased p–i–n diode are illustrated in Fig. 19.3-6. This structure serves to extend the width of the region supporting an electric field, in effect widening the depletion layer.

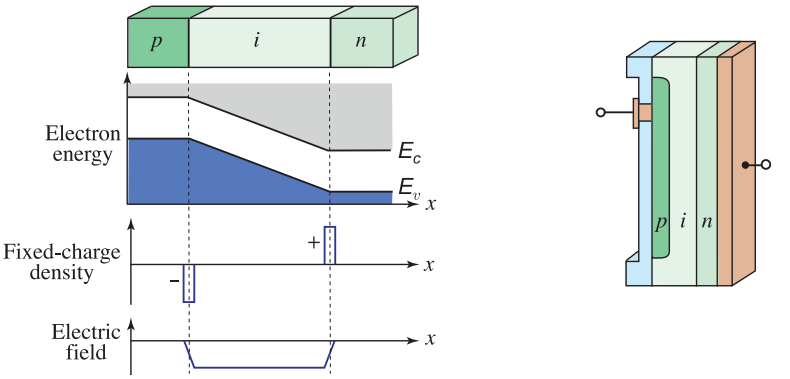


Figure 19.3-6 The p–i–n photodiode structure, energy-band diagram, charge distribution, and electric-field distribution. The incident light can be directed parallel to the junction layer (so-called edge illumination), i.e., vertically in the figure. Alternatively, the light can be directed at normal incidence to the junction layer, i.e., horizontally in the plane of the figure.

Photodiodes with a p–i–n structure offer the following advantages:

- Increasing the width of the depletion layer of the device (where the generated carriers can be transported by drift) increases the area available for capturing light.
- Increasing the width of the depletion layer reduces the junction capacitance and thereby the RC time constant. On the other hand, the transit time increases with increasing width of the depletion layer.
- Reducing the ratio between the diffusion length and the drift length of the device results in a greater proportion of the generated carriers undergoing the faster drift process.

Normal vs. edge illumination. For **normally illuminated photodiodes**, the carrier flow is parallel to that of the photons, which leads to a tradeoff between responsivity and bandwidth since attaining high responsivity requires a sufficiently thick absorption layer (considering $In_{0.53}Ga_{0.47}As$ with $\alpha \approx 10^4$ cm^{-1} as an example, a thickness of ≈ 2 μm is required to absorb 88% of the incident light). Yet, the thicker the absorption region the greater the transit time, and hence the narrower the bandwidth. For **edge-illuminated photodiodes**, on the other hand, light is coupled into the device in a direction perpendicular to the carrier transport, which permits the absorption layer to be much thinner than that for normal-incidence photodiodes. Thus, a salutary feature of the edge-illumination configuration is that it decouples the absorption of light (responsivity) and the carrier transit time (bandwidth).

Evanescent coupling. A common implementation for edge-illuminated photodiodes is to abut an optical fiber or a passive waveguide with the intrinsic region. In that case, the quantum efficiency (19.1-3) must accommodate the coupling loss and the appropriate optical power confinement factor Γ. **Evanescently coupled waveguide photodiodes** were developed to mitigate such losses. They typically consist of a photodiode located atop a passive waveguide and are well-suited to monolithic integration in photonic integrated circuits. Evanescent coupling also mitigates any lattice mismatch between the photodiode and waveguide materials and typically offers large bandwidths. Performance can be further enhanced by implementing a **traveling-wave configuration**.

Responsivity of Si PIN photodiodes. The responsivity of two commercially available Si p–i–n photodiodes is compared with that of an ideal device ($\eta = 1$) in Fig. 19.3-7. The responsivity maximum occurs at a wavelength shorter than the bandgap wavelength. This is because Si is an indirect-bandgap material. The photon-absorption transitions therefore typically take place from valence-band to conduction-band states that lie well above the conduction-band edge, as depicted in Fig. 17.2-6.

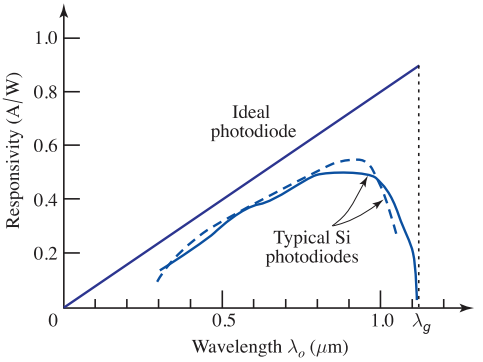


Figure 19.3-7 Responsivity (A/W) vs. wavelength (μm) for an ideal Si photodiode ($\eta = 1$) and for two commercially available Si p–i–n photodiodes. The quantum efficiency of a carefully constructed, antireflection-coated silicon device can approach unity.

Enhancing Si-photodiode performance with photon-trapping microstructures. Silicon p–i–n photodiodes that operate at visible wavelengths enjoy the benefit of a large absorption coefficient α. However, as the wavelength moves toward the near infrared, the absorption coefficient decreases, thereby necessitating the use of a thicker absorption region to maintain the external quantum efficiency. This in turn results in increased transit time and reduced device bandwidth. This undesirable effect can be mitigated by permeating the intrinsic absorption region with micro- and nanostructured holes that serve to efficiently trap the light, which substantially increases the effective absorption coefficient.[†] This in turn allows for devices with thinner absorption regions and larger bandwidths.

C. Heterostructure Photodiodes

Heterostructure photodiodes, comprising at least two semiconductor materials with different bandgaps, provide flexibility that can offer advantages over homojunctions fabricated from a single material. A heterojunction that incorporates a large-bandgap material ($E_g > h\nu$), for example, can make use of its transparency to minimize optical absorption outside the depletion region and hence to reduce surface recombination and maximize ζ in (19.1-3); the large-bandgap material is then said to be a **window layer**.

Two or more materials can also be fashioned into a structure that makes use of the best features of each. A Ge-on-Si waveguide structure, for example, combines the superior guiding properties of Si with the strong near-infrared absorption properties of Ge (Example 19.3-2). Several material systems are of particular interest (see, e.g., Figs. 17.1-7, 17.1-8, and 18.1-16):

- $Al_xGa_{1-x}As/GaAs$ (AlGaAs lattice matched to a GaAs substrate) is useful in the wavelength range 0.7 to 0.87 μm.

[†] See, e.g., Y. Gao, H. Cansizoglu, K. G. Polat, S. Ghandiparsi, A. Kaya, H. H. Mamtaz, A. S. Mayet, Y. Wang, X. Zhang, T. Yamada, E. P. Devine, A. F. Elrefaie, S.-Y. Wang, and M. S. Islam, Photon-Trapping Microstructures Enable High-Speed High-Efficiency Silicon Photodiodes, *Nature Photonics*, vol. 11, pp. 301–308, 2017.

- $In_xGa_{1-x}As/InP$ is a direct-bandgap material that can be lattice matched to an InP substrate. The bandgap wavelength of this material is compositionally tunable over the near infrared and a portion of the mid infrared: 0.873 μm (GaAs) $\leq \lambda_g \leq$ 3.44 μm (InAs) (Fig. 18.1-16). This range of wavelengths includes the 1.3–1.6-μm telecommunications band. A typical InGaAs p–i–n photodetector operating at 1550 nm has a quantum efficiency $\eta \approx 0.80$, a responsivity R ≈ 0.95 A/W, and a bandwidth ≈ 10 GHz.
- $Ge/Ge_{0.93}Sn_{0.07}/Ge$ CMOS-compatible double-heterostructures grown directly on Si operate at wavelengths as long as 2.2 μm. In the current state of the technology, a typical p–i–n photodiode operating at room temperature at 1550 nm exhibits a responsivity R $\approx$ 0.3 A/W at normal incidence under a reverse-bias voltage of 0.1 V.
- $Hg_xCd_{1-x}Te/CdTe$ finds extensive use in the mid infrared. This II–VI ternary semiconductor can be lattice matched to CdTe at nearly all compositions since CdTe and HgTe have nearly the same lattice parameter (Fig. 17.1-8). HgCdTe is compositionally tunable with a bandgap energy that extends from about 0.85 to 16 μm. Applications include night vision, thermal imaging, and long-wavelength optical communications.
- Quaternaries, such as $In_{1-x}Ga_xAs_{1-y}P_y/InP$, $In_{1-x}Ga_xAs_{1-y}Sb_y/GaSb$, and $Al_{1-x}Ga_xAs_{1-y}Sb_y/GaSb$, which are useful over wavelengths that extend from 0.92 to 5 μm, are of interest because the fourth element provides an additional degree of freedom that allows lattice matching to be achieved for different compositionally determined values of E_g.

EXAMPLE 19.3-1. ***III–V Multi-Junction Photovoltaic Solar Cell.*** Multi-junction solar cells are heterostructures comprising multiple thin semiconductors films of different bandgap energies stacked one atop the other. If the light enters from the top, the values of E_g successively decrease from the top to the bottom layer, so that each layer acts as a window layer for the one below it. This allows absorption to be optimized for the various spectral slices of the solar radiation that reaches earth. A heterostructure photodiode containing five layers of III–V materials (InGaP/GaAs/InGaAsNSb/GaSb/InGaAsSb) operating in the photovoltaic mode harvests $> 99\%$ of the available solar power and achieves a quantum efficiency $\eta > 41\%$.[†] The two smallest-bandgap layers at the bottom of the stack, GaSb and InGaAsSb, absorb light in the wavelength region between 1.7 and 2.5 μm in the infrared (Fig. 18.1-16).

EXAMPLE 19.3-2. ***Ge-on-Si Waveguide Photodiode.*** In the 1.3–1.6-μm telecommunications band, silicon is transparent ($\lambda_g = 1.11$ μm; see Table 17.1-2) and therefore not photosensitive. Efficient and high-speed silicon-photonics-based photodetectors have traditionally been fabricated via hybrid integration of photosensitive III–V materials, such as InGaAs on Si, but an alternative is to make use of CMOS-compatible Ge, which is sensitive over this range of wavelengths. However, accommodating the substantial lattice mismatch between Ge and Si ($\approx 4.2\%$) dictates using configurations in which Ge and Si are evanescently or butt coupled. In one particular design that relies on edge illumination, light emerging from a Si waveguide is butt-coupled to the intrinsic region of a lateral p–i–n Ge-on-Si photodiode integrated at the end of the Si waveguide.[‡] Operating at 1550 nm, this photodiode offers a responsivity of ≈ 1 A/W and a bandwidth > 50 GHz. It has performance comparable with that obtained using hybrid integration of InGaAs on Si and it can also be operated in photovoltaic mode.

[†] See M. P. Lumb, S. Mack, K. J. Schmieder, M. González, M. F. Bennett, D. Scheiman, M. Meitl, B. Fisher, S. Burroughs, K.-T. Lee, J. A. Rogers, and R. J. Walters, GaSb-Based Solar Cells for Full Solar Spectrum Energy Harvesting, *Advanced Energy Materials*, vol. 7, 1700345, 2017.

[‡] See L. Vivien, A. Polzer, D. Marris-Morini, J. Osmond, J. M. Hartmann, P. Crozat, E. Cassan, C. Kopp, H. Zimmermann, and J. M. Fédéli, Zero-Bias 40 Gb/s Germanium Waveguide Photodetector on Silicon, *Optics Express*, vol. 20, pp. 1096–1101, 2012.

Schottky-Barrier Photodiodes

Metal–semiconductor photodiodes (also called **Schottky-barrier photodiodes**) are formed from metal–semiconductor heterojunctions. A thin semitransparent metallic film is used in place of the p-type (or n-type) layer in the p–n junction photodiode. The thin film is sometimes fabricated from a metal–semiconductor alloy that behaves like a metal. The Schottky-barrier structure and its energy-band diagram are shown schematically in Fig. 19.3-8 for metal deposited on a lightly doped n-type semiconductor.

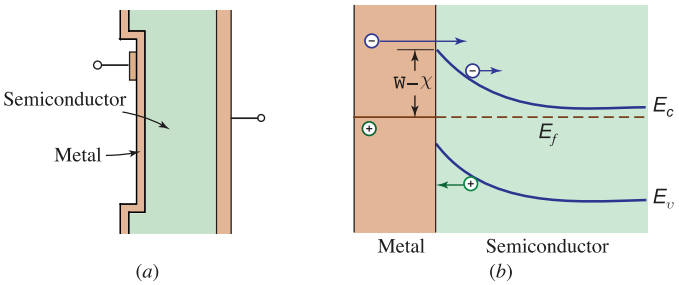


Figure 19.3-8 (a) Structure and (b) energy-band diagram of a Schottky-barrier photodiode formed by depositing a metal on an n-type semiconductor. At equilibrium, the Fermi levels in the two regions align. These photodetectors are responsive to photon energies greater than the Schottky barrier height, $h\nu > \mathbb{W} - \chi$. Schottky-barrier photodiodes can be fabricated from many materials, such as Au on n-type Si (which operates in the visible) and platinum silicide (PtSi) on p-type Si (which operates over a range of wavelengths that stretches from the ultraviolet to the infrared).

On contact, electrons flow from the semiconductor to the metal, bringing the Fermi levels of the two materials into alignment. This results in a region depleted of free electrons just inside the semiconductor interface. The accompanying fixed positive charges in the semiconductor cause its valence and conduction bands to bend upward at the interface. At equilibrium, the discontinuity in allowed energy states of the two materials gives rise to the Schottky barrier, which blocks the flow of electrons from the metal back to the semiconductor and is responsible for the rectifying nature of the device. The absorption of a photon results in current flow.

Schottky-barrier photodiodes are particularly useful in a number of circumstances:

- Not all semiconductors can be prepared in both p-type and n-type forms; Schottky-barrier devices can be used in these material systems.
- Semiconductors used for the detection of visible and ultraviolet light with a photon energy well above the bandgap energy have a large absorption coefficient. This gives rise to substantial surface recombination and a reduction of quantum efficiency. The depletion layer of the metal–semiconductor junction is present immediately at the surface, in contrast, thereby eliminating surface recombination.
- The response speed of p–n and p–i–n junction photodiodes is in part limited by the slow diffusion current associated with photocarriers generated close to, but outside of, the depletion layer. One way of decreasing this unwanted absorption is to decrease the thickness of one of the junction layers, but this should be implemented without substantially increasing the series resistance of the device, which increases the RC time constant. The Schottky-barrier structure achieves this because of the low resistance of the metal. Furthermore, Schottky-barrier structures are majority-carrier devices and therefore have inherently fast responses and large operating bandwidths. Response times of ps, corresponding to bandwidths of 100 GHz, are readily available.

Representative responsivity curves for several p–i–n and Schottky-barrier photodiodes are displayed in Fig. 19.3-9.

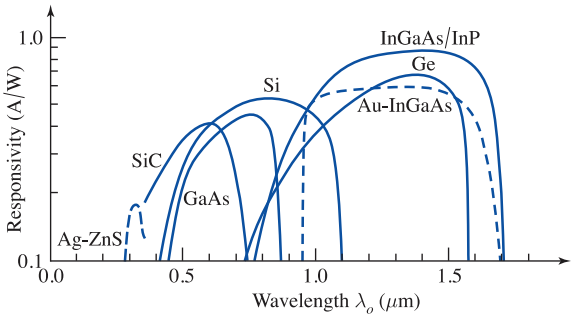


Figure 19.3-9 Responsivity R (A/W) versus wavelength λ_o (μm) for a number of p–i–n (solid) and Schottky-barrier (dashed) photodiodes. For ternary and quaternary devices, the wavelength of maximal response depends on composition. Response times in the tens of ps, corresponding to bandwidths $\approx$ 50 GHz, are generally available.

EXAMPLE 19.3-3. ***Graphene–Si Schottky-Barrier Photodiode.*** Graphene is a 2D crystal comprising a one-atom-thick layer of graphite whose atoms are arranged in a hexagonal honeycomb lattice (Sec. 17.1B). By virtue of graphene's high conductivity, low reflectance, high carrier mobility, and the broad spectrum over which it interacts with radiation, the junction between graphene and n-type silicon can serve as a high-speed, broadband Schottky-barrier photodiode. It is remarkable that a junction formed by a 2D/zero-bandgap material and a 3D/finite-bandgap material yields a practical operating device. Calculations show that the absorbance of graphene is $\mathcal{A} \approx \pi e^2/\hbar c \approx 2.3\%$ for photon energies below 3 eV. Since the reflectance of a single graphene layer is minuscule, $\mathcal{R} \approx 1.3 \times 10^{-4}$, the intensity transmittance is $\mathcal{T} \approx 1 - \mathcal{A} \approx 97.7\%$ (at normal incidence). The graphene thus acts as a nonreflecting, transparent electrode that serves to collect carriers while the optical absorption associated with the detection process takes place in the silicon. Schottky-barrier photodiodes have been fabricated by depositing graphene on lightly doped n-type silicon using CVD in a CMOS-compatible process.[†] The devices are sensitive over the visible and near-infrared spectral regions, exhibiting a cutoff at the bandgap wavelength of Si, $\lambda_g = 1.1$ μm. Illuminated at normal incidence, the responsivity of a typical device closely resembles the Si responsivity curve portrayed in Fig. 19.3-9. Using (19.1-5), together with the observed responsivity of R $\approx$ 0.3 A/W at 0.6 μm on the graphene–Si curve leads to a quantum efficiency $\eta \approx 0.6$. Devices such as these can also be configured to provide *internal gain* by using electrodes that direct the current flow laterally, along the width of the device rather than through its thickness, thereby co-opting the enormous contrast between the fast transit time of graphene, τ_h, and the slow recombination time in the graphene–silicon system, τ. In accordance with (19.2-3), this configuration leads to a gain given by $G = \tau/\tau_h$. The observed value of gain in these devices is $G \approx 3 \times 10^4$, corresponding to a responsivity R $\approx 10^4$ in accordance with (19.1-8). It is noteworthy that the Fermi level of graphene may be tuned by doping, or by the application of a bias voltage, which serves to modify the barrier height of the Schottky junction and thus the current–voltage relationship of the device. Other 2D materials have also been juxtaposed with various semiconductor structures to construct photodiodes that operate in both biased and photovoltaic modes.

[†] See F. Liu and S. Kar, Quantum Carrier Reinvestment-Induced Ultrahigh and Broadband Photocurrent Responses in Graphene–Silicon Junctions, *ACS Nano*, vol. 8, pp. 10270–10279, 2014.

19.4 AVALANCHE PHOTODIODES

An **avalanche photodiode (APD)** operates by converting each detected photon into a cascade of moving carrier pairs. Weak light is then able to elicit a current that is sufficiently large so that it can be readily detected by the electronics following the device. APDs are configured as strongly reverse-biased photodiodes that have large electric fields in the junction region, enabling charge carriers to acquire sufficient energy so they can excite new carriers via impact ionization. However, the multiplication process requires time to play out and introduces gain noise, which limits system bandwidth and performance. Avalanche photodiodes find extensive use in optical fiber communication receivers (Sec. 25.1D) and are used in applications involving imaging, scanning, and range finding.

A. Conventional Avalanche Photodiodes

The history of a typical electron–hole pair in the depletion region of a **conventional avalanche photodiode (CAPD)** is depicted in Fig. 19.4-1. A photon is absorbed at point 1, creating an electron–hole pair (an electron in the conduction band and a hole in the valence band). The electron accelerates under the influence of the strong electric field, thereby increasing its energy with respect to the bottom of the conduction band. The acceleration process is constantly interrupted by random collisions with the lattice in which the electron loses some of its acquired energy. These competing processes cause the electron to reach an average saturation velocity. Should the electron be lucky and acquire an energy larger than E_g at any time during the process, it has an opportunity to generate a second electron–hole pair by impact ionization (say at point 2). The two electrons then accelerate under the effect of the field, and each of them may be the source for a further impact ionization. The holes generated at points 1 and 2 also accelerate, moving toward the left. Each of these also has a chance of creating an impact ionization should they acquire sufficient energy, thereby generating a hole-initiated electron–hole pair (e.g., at point 3).

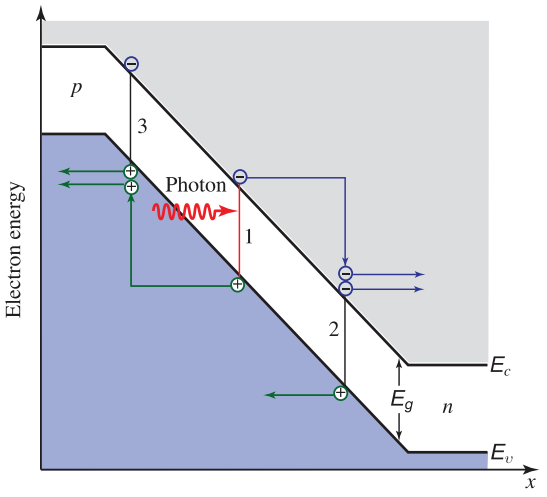


Figure 19.4-1 Schematic representation of the multiplication process in a conventional homojunction avalanche photodiode (CAPD).

Ionization Coefficients and Ionization Ratio

Ionization coefficients. The abilities of electrons and holes to impact ionize are characterized by the **ionization coefficients** α_e and α_h, respectively. These quantities represent ionization probabilities per unit length (cm^{-1}); the inverse coefficients $1/\alpha_e$

and $1/\alpha_h$ represent average distances between consecutive ionizations for electrons and holes, respectively. The ionization coefficients increase with electric-field strength in the depletion layer (since it provides the acceleration) and decrease with increasing device temperature (since the increased frequency of collisions diminishes the opportunity of a carrier gaining sufficient energy to cause an ionization). The simplified theory presented in this section assumes that α_e and α_h are constants. For purposes of noise reduction, however, it can be advantageous to design devices in which the ionization coefficients depend on carrier history and position in particular ways, as discussed in Sec. 19.4B.

Ionization ratio. An important parameter for characterizing the performance of an APD is the **ionization ratio**, which is defined as the ratio of the ionization coefficients,

$$\mathrm{k} = \alpha_h/\alpha_e\,. \tag{19.4-1}$$

When holes do not ionize appreciably (i.e., when $\alpha_h \ll \alpha_e$ so that $\mathrm{k} \ll 1$), most of the ionization is achieved by electrons. The avalanching process then proceeds principally from left to right in Fig. 19.4-1 (i.e., from the p side to the n side of the device), and terminates when all of the electrons arrive at the n side of the depletion layer. If electrons and holes both ionize appreciably ($\mathrm{k} \approx 1$), those holes that move to the left create electrons that move to the right, which in turn generate further holes moving to the left, possibly leading to an unending circulation. Though this feedback process increases the gain of the device (the total generated charge in the circuit per photocarrier pair, q/e), it is nevertheless undesirable for several reasons:

- It is time consuming and therefore reduces the device bandwidth.
- It is random and therefore increases the device noise.
- It can be unstable, thereby causing avalanche breakdown.

It is therefore desirable to fabricate APDs from **single-carrier-multiplication** materials, i.e., from materials that permit only one type of carrier (either electrons or holes) to impact ionize. For example, if electrons have the higher ionization coefficient, optimal behavior is attained by injecting the electron of a photocarrier pair at the p-type edge of the depletion layer, and by making use of a material whose value of k is as small as possible. If the material is such that holes have the higher ionization coefficient, the hole of a photocarrier pair should be injected at the n-type edge of the depletion layer and k should be as large as possible. Hence, the ideal case of single-carrier multiplication is achieved when $\mathrm{k} = 0$ or ∞.

Gain and Responsivity

Single-carrier-injection single-carrier-multiplication (SCISCM) devices. As a prelude to determining the gain of an APD in which both kinds of carriers can give rise to ionizations, we first consider the simpler problem of single-carrier (electron) multiplication ($\alpha_h = 0$, $\mathrm{k} = 0$) with single-carrier (electron) injection. Let $J_e(x)$ be the electric current density carried by electrons at location x, as illustrated in Fig. 19.4-2.

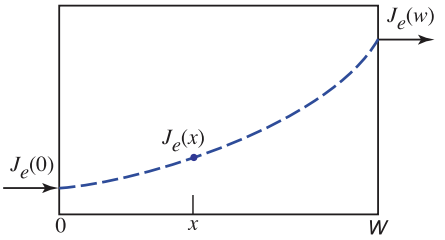


Figure 19.4-2 Exponential growth of the electric current density in a single-carrier-injection single-carrier-multiplication APD.

Within a distance dx, on average, the current is incremented by

$$dJ_e(x) = \alpha_e J_e(x)\, dx, \tag{19.4-2}$$

from which we obtain the differential equation

$$\frac{dJ_e}{dx} = \alpha_e J_e(x), \tag{19.4-3}$$

whose solution is the exponential function $J_e(x) = J_e(0)\exp(\alpha_e x)$. The gain $G = J_e(w)/J_e(0)$ is therefore

$$G = \exp(\alpha_e w). \tag{19.4-4}$$

The electric current density increases exponentially with the product of the ionization coefficient α_e and the multiplication layer width w. The result is analogous to that for gain in a laser amplifier [see (15.1-7)].

Single-carrier-injection double-carrier-multiplication (SCIDCM) devices. Solution of the double-carrier multiplication problem requires knowledge of both the electron current density $J_e(x)$ and the hole current density $J_h(x)$. We assume that only electrons are injected into the multiplication region (single-carrier injection). Since hole ionizations also produce electrons, however, the growth of $J_e(x)$ is governed by the differential equation

$$\frac{dJ_e}{dx} = \alpha_e J_e(x) + \alpha_h J_h(x). \tag{19.4-5}$$

As a result of charge neutrality, $dJ_e/dx = -dJ_h/dx$, and the sum $J_e(x) + J_h(x)$ must remain constant for all x under steady-state conditions. This is clear from the illustration provided in Fig. 19.4-3; the total number of charge carriers crossing any plane is the same regardless of the position x.

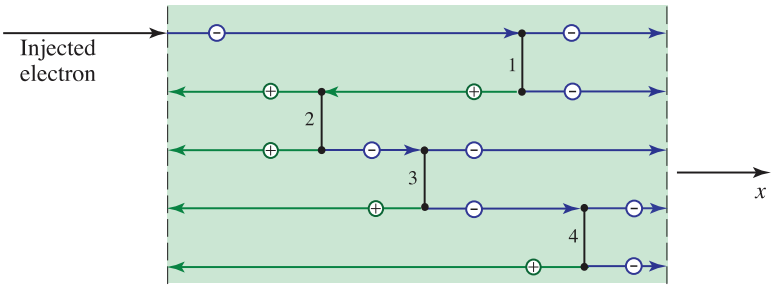


Figure 19.4-3 Constancy of the sum of the electron and hole current densities across a plane at any value of x. By way of illustration, a single injected electron gives rise to four impact ionizations, with four electrons plus four holes crossing every plane.

Since it is assumed that no holes are injected at $x = w$, we have $J_h(w) = 0$ so that

$$J_e(x) + J_h(x) = J_e(w), \tag{19.4-6}$$

as displayed in Fig. 19.4-4.

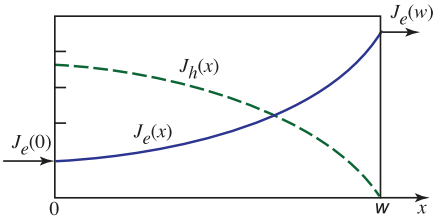


Figure 19.4-4 Electron and hole current densities in a double-carrier-multiplication APD with electron injection.

The hole current density $J_h(x)$ can therefore be eliminated in (19.4-5) to obtain

$$\frac{dJ_e}{dx} = (\alpha_e - \alpha_h)J_e(x) + \alpha_h J_e(w). \tag{19.4-7}$$

This first-order differential equation is readily solved for the gain $G = J_e(w)/J_e(0)$. For $\alpha_e \neq \alpha_h$, the result is $G = (\alpha_e - \alpha_h)/\{\alpha_e \exp[-(\alpha_e - \alpha_h)w] - \alpha_h\}$, from which we obtain

$$G = \frac{1 - \mathrm{k}}{\exp\left[-(1 - \mathrm{k})\alpha_e w\right] - \mathrm{k}}. \tag{19.4-8}$$

(19.4-8) APD Gain

The single-carrier multiplication result for the gain (19.4-4), with its simple exponential growth, is recovered when $\mathrm{k} = 0$. When $\mathrm{k} = \infty$ so that only holes multiply, the gain remains unity since only electrons are injected and electrons do not multiply. For $\mathrm{k} = 1$, (19.4-8) is indeterminate and the gain must be obtained directly from (19.4-7); the result is then $G = 1/(1 - \alpha_e w)$. An instability is reached when $\alpha_e w = 1$. The dependence of the gain on $\alpha_e w$ for several values of the ionization ratio k is illustrated in Fig. 19.4-5.

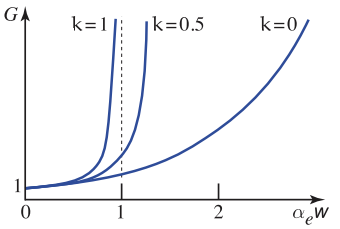


Figure 19.4-5 Growth of the gain G with multiplication-layer width w for several values of the ionization ratio k in a single-carrier-injection double-carrier-multiplication (SCIDCM) APD with pure electron injection.

The responsivity R of the SCIDCM APD is obtained by using (19.4-8) in conjunction with the general relation (19.1-8).

Device Structures

As with photodiodes, APDs should be configured with a geometry designed to maximize photon absorption (e.g., by emulating a p–i–n structure) and to minimize dark current. Concomitantly, the multiplication region should be designed with a sufficiently strong field to foster single-carrier impact ionization with minimal multiplication noise, and thin enough to minimize the possibility of uncontrolled localized avalanches associated with instabilities and microplasmas.

SAM APDs. These conflicting requirements call for an APD design in which the absorption and multiplication regions are spatially separated. Structures of this kind are known as **separate absorption and multiplication** (**SAM**) APDs.

Reach-through APDs. The operation of a SAM APD is most readily understood by considering a device fabricated from a material such as Si, for which $k \approx 0$ (Example 19.6-3). In the device portrayed in Fig. 19.4-6, known as a **reach-through APD**, photons are absorbed in an extended intrinsic (or lightly doped) π region. The photoelectrons drift across this region under the influence of a moderate electric field and then enter a thin p–n^+ junction where they experience an electric field sufficient to cause avalanching. The reverse-bias voltage applied across the device is large enough for the depletion region to *reach through* the p and π regions into the p^+ contact layer.

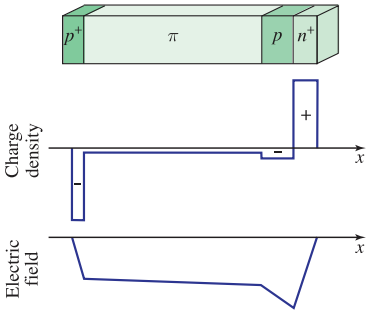


Figure 19.4-6 Reach-through p^+–π–p–n^+ APD structure. The π region is intrinsic or very lightly doped p-type material. The p^+ and n^+ regions are heavily doped. Avalanching occurs in the thin p–n^+ junction region where the electric field is strong.

SACM APDs. A variation on the SAM theme is the **separate absorption, charge, and multiplication (SACM)** APD, a heterostructure device that incorporates a *charge layer* in addition to the separate absorption and multiplication layers. The charge layer is designed to keep the electric field in the absorption region small (so that dark current arising from tunneling in this narrow-bandgap layer is minimized), while providing a large field in the wide-bandgap multiplication region (to facilitate impact ionizations). Use of a wide-bandgap multiplication layer minimizes tunneling and thermal effects, thereby minimizing the dark current generated in this region. The relative field strengths in the two regions are governed by the doping profile, much as with the reach-through structure shown in Fig. 19.4-6. Compositionally graded layers are often incorporated to avoid carriers from being trapped at the interfaces between layers.

Response Time

Avalanche buildup time. Aside from the usual transit, diffusion, and RC time constants that govern the response time of photodiodes, APDs are subject to an additional time constant known as the **avalanche buildup time**. It is the time required for the impact-ionization process to unfold and settle, and it places a limit on the speeds at which APD-based systems can operate. Systems operating at bit rates beyond those limits must make use of p–i–n photodiodes.

The response time of a separate absorption and multiplication (SAM) APD is illustrated in Fig. 19.4-7 by displaying the history of a photoelectron generated at the edge of the absorption region (point 1). The APD is assumed to operate via single-carrier-injection double-carrier-multiplication (SCIDCM). The electron drifts with a saturation velocity v_e, reaching the multiplication region (point 2) after a transit time w_d/v_e. Within the multiplication region the electron also travels with velocity v_e. Through impact ionization it creates electron–hole pairs, say at points 3 and 4, generating two additional electron–hole pairs. The holes travel in the opposite direction with their saturation velocity v_h. The holes can also cause impact ionizations, resulting in electron–hole pairs as shown, for example, at points 5 and 6. These carriers can themselves cause

impact ionizations, sustaining the feedback loop. The process terminates when the last hole (at point 7) leaves the multiplication region and crosses the absorption region to arrive at point 8.

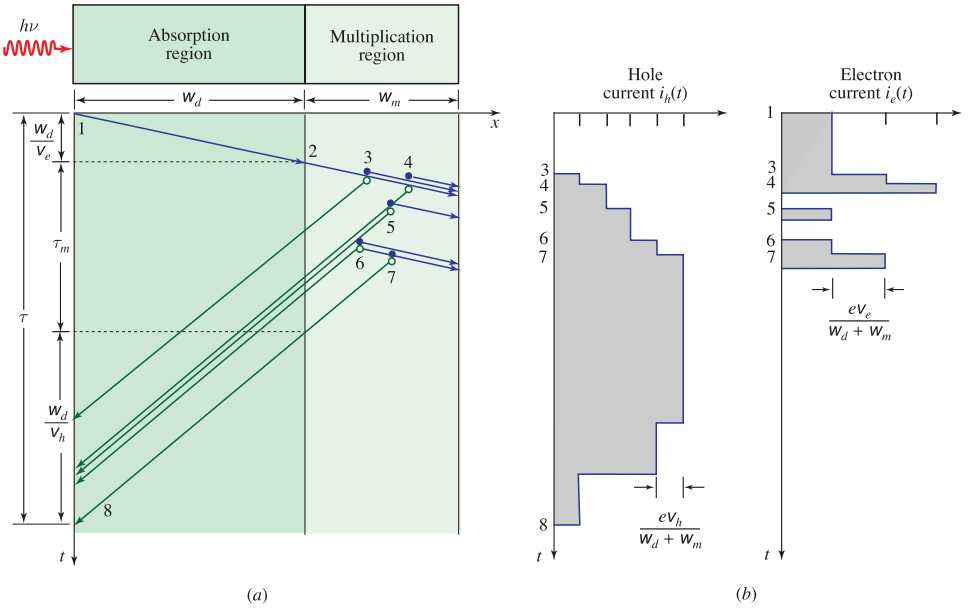


Figure 19.4-7 (a) Tracing the course of the avalanche buildup time in a SAM APD with the help of a position–time graph. The device is assumed to operate via single-carrier-injection and double-carrier multiplication. The blue lines represent electrons, and the green lines represent holes. Electrons move to the right with velocity v_e and holes move to the left with velocity v_h. Electron–hole pairs are produced in the multiplication region. The carriers cease moving when they reach the edge of the material. (b) Hole current $i_h(t)$ and electron current $i_e(t)$ induced in the circuit. Each carrier pair induces a charge e in the circuit. The total induced charge q, which is the area under the $i_e(t) + i_h(t)$ vs. t curve, is Ge. This figure is a generalization of Fig. 19.1-6, which applies for a single electron–hole pair.

The total time τ required for the entire process (between points 1 and 8) to unfold is the sum of the transit times (roughly from 1 to 2 and from 7 to 8) and the multiplication time denoted τ_m,

$$\tau \approx \frac{w_d}{v_e} + \frac{w_d}{v_h} + \tau_m \,. \tag{19.4-9}$$

Because of the randomness of the multiplication process, the multiplication time τ_m is random. In the special case when $\mathrm{k} = 0$ (no hole multiplication) the maximum value of τ_m is readily seen from Fig. 19.4-7 to be

$$\tau_m \approx \frac{w_m}{v_e} + \frac{w_m}{v_h} \,. \tag{19.4-10}$$

For large gain G, and for electron injection with $0 < \mathrm{k} < 1$, an order of magnitude estimate of the average value of τ_m is obtained by multiplying the first term of (19.4-10) by the factor $G\mathrm{k}$, so that

$$\tau_m \approx \frac{G\mathrm{k}w_m}{v_e} + \frac{w_m}{v_h} \,. \tag{19.4-11}$$

Reducing k shortens τ_m and τ and hence increases the speed at which the APD operates. The associated hole current $i_h(t)$ and electron current $i_e(t)$ are also displayed in Fig. 19.4-7. A more accurate theory is rather complex.

EXAMPLE 19.4-1. ***Avalanche Buildup Time in a Si APD.*** Consider a silicon APD with $w_d = 50\ \mu$m, $w_m = 0.5\ \mu$m, $v_e = 10^7$ cm/s, $v_h = 5 \times 10^6$ cm/s, $G = 100$, and $\mathrm{k} = 0.1$. Equation (19.4-10) yields $\tau_m = 5 + 10 = 15$ ps, so that (19.4-9) gives $\tau = 1020$ ps $= 1.02$ ns. On the other hand, (19.4-11) yields $\tau_m = 60$ ps, so that (19.4-9) provides $\tau = 1065$ ps $= 1.07$ ns. For a p–i–n photodiode with the same values of w_d, v_e, and v_h, the transit time is $w_d/v_e + w_d/v_h \approx 1$ ns. The results do not differ greatly because τ_m is quite low in a Si SAM device.

Materials

The materials used for the photosensitive absorption layers of APDs are closely related to those used for p–i–n photodiodes (Fig. 19.3-9). The materials used for the multiplication layers should have values of the ionization ratio k that are as low as possible for electron injection or as high as possible for hole injection; relatively large bandgap energies are also useful for minimizing dark current.

Silicon and AlInAsSb APDs, which are principally used in the wavelength regions 700–900 nm and 1.3–1.6 μm, respectively, offer ionization ratios as low as $\mathrm{k} \approx 0.01$–0.02, and hence nearly negligible gain noise. Low-k materials suitable for use in the ultraviolet include GaN and AlGaN, while HgCdTe finds use in the mid infrared.

EXAMPLE 19.4-2. ***InGaAs/InP SAM APD.*** InGaAs/InP SAM APDs, in which InGaAs and InP serve as the materials used in the absorption and multiplication regions, respectively, have traditionally been used for optical fiber communication systems that operate in the 1.3–1.6-μm telecommunications band (Sec. 25.1D). Because they are easily fabricated and offer high responsivities (Fig. 19.3-9), these devices continue to be used despite the somewhat unfavorable ionization ratio for InP ($1/\mathrm{k} \approx 0.3$). They are generally operated at voltages that lie between punchthrough and breakdown (Fig. 19.4-8); as the reverse-bias voltage increases, so too do the gain and dark current. Tens of volts of bias results in an electric field $\approx 10^5$ V/cm, which is sufficient to initiate the avalanche process. Typical values of the mean gain and bandwidth are $\overline{G} \approx 10$ and $B = 10$ GHz, respectively. In an SACM configuration, InGaAsP can serve as the charge layer. AlInAs is sometimes used in place of InP for the multiplication region because of its more favorable ionization ratio ($1/\mathrm{k} \approx 0.2$); in accordance with (19.4-11), this also provides higher speed.

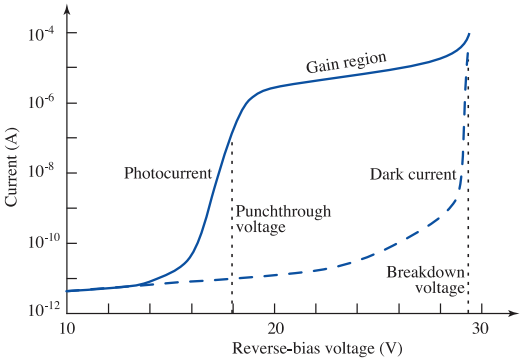


Figure 19.4-8 Current–voltage characteristic for an InGaAs/InP separate absorption and multiplication (SAM) APD. The device is operated at a reverse-bias voltage that lies between **punchthrough** (the voltage at which the depletion region penetrates the absorption region) and **breakdown** (the voltage at which uncontrolled avalanching occurs).

EXAMPLE 19.4-3. ***Ge-on-Si SACM APD.*** The Ge-on-Si SACM APD, a group-IV-photonics monolithic device, operates across most of the 1.3–1.6-μm telecommunications band. A principal merit of this device is its CMOS compatibility and its availability for on-chip integration. The normal-incidence device is fabricated such that the photons impinge on a Ge absorption layer grown atop a layer of Si, in which carrier multiplication takes place; this avoids the intrinsic noisiness of carrier multiplication in Ge (see Prob. 19.4-2). A Si charge layer maintains a low electric field in the absorption region to minimize dark current. With unintentionally doped Ge and Si layers of thicknesses 1 and ½ μm, respectively, and a 0.1-μm p-type Si charge layer, this APD exhibits the following properties:[†] mean gain $\overline{G} \approx 50$; gain–bandwidth product $GB \approx 350$ GHz (substantially exceeding that of an InGaAs/InP APD); responsivity R ≈ 5.9 A/W at $\lambda_o = 1.3$ μm; ionization ratio k ≈ 0.09; and operation at bit rates of 25 Gb/s. The principal limiting features are: 1) the steep decrease in the absorption coefficient of Ge for wavelengths greater than ≈ 1.55 μm, and 2) the relatively large dark current arising from deep-level traps associated with the lattice mismatch between Ge and Si. Much as with Ge-on-Si waveguide photodiodes (Example 19.3-2), waveguide-based Ge-on-Si SACM APDs decouple the light absorption and carrier collection, enabling these devices to offer both high quantum efficiency and high speed.

EXAMPLE 19.4-4. ***AlInAsSb/GaSb SACM APD.*** The AlInAsSb SACM APD is a III–V direct-bandgap device that operates across the 1.3–1.6-μm telecommunications band. This digital-alloy material system, which is lattice matched to GaSb, offers a high absorption coefficient and is suitable for designing complex structures. The $Al_xIn_{1-x}As_ySb_{1-y}$ APD illustrated in Fig. 19.4-9 makes use of absorption and multiplication layers with low and high Al content, respectively, corresponding to small and large energy bandgaps.[‡] It has a quantum efficiency $\eta \approx 0.4$ and a dark current that is somewhat greater than that of the InGaAs/AlInAs APD, but substantially lower than that of the Ge-on-Si APD. The ionization ratio k, which is comparable to that of Si, is approximately 0.01 at a mean gain $\overline{G} = 10$ (Example 19.6-4). Digital alloys of AlInAs also prove useful in fabricating low-noise APDs.

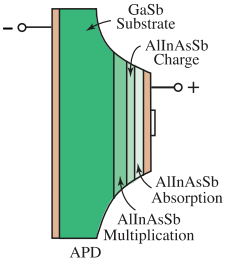


Figure 19.4-9 Structure of an $Al_xIn_{1-x}As_ySb_{1-y}$ separate absorption, multiplication, and charge (SACM) APD with a lattice-matched GaSb substrate. The device has a 1-μm-thick, n^--type AlInAsSb absorption layer ($x = 0.4$, $y = 0.3$) surrounded by a pair of 100-nm-thick, p^+-type compositionally graded AlInAsSb layers (not shown). A 150-nm-thick, p-type AlInAsSb charge layer ($x = 0.7$, $y = 0.3$) separates the absorption layer from the 1-μm-thick, n^--type AlInAsSb multiplication layer ($x = 0.7$, $y = 0.3$). The substrate is n^+-type GaSb.

EXAMPLE 19.4-5. ***HgCdTe SAM APD.*** $Hg_xCd_{1-x}Te$ is a II–VI direct-bandgap material whose bandgap wavelength can be compositionally tuned from 0.85 to 16 μm (Fig. 17.1-8). This material is useful for fabricating SAM APDs with cutoff wavelengths extending from 2 to 11 μm in the mid infrared.[*] HgCdTe APDs offer quantum efficiencies as high as 0.9, high gain ($\overline{G} > 1000$), and large gain–bandwidth products ($GB > 1$ THz), but they require cryogenic cooling. They exhibit single-carrier-injection single-carrier-multiplication (SCISCM) behavior (electron multiplication prevails for $0.2 < x < 0.6$) with an ionization ratio k ≈ 0, so they are very low-noise devices. HgCdTe APDs are readily incorporated into focal plane arrays that are useful in a whole host of low-flux and high-speed infrared applications, including imaging and lidar.

[†] See D. Dai, M. Piels, and J. E. Bowers, Monolithic Germanium/Silicon Photodetectors with Decoupled Structures: Resonant APDs and UTC Photodiodes, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, 3802214, 2014.

[‡] See S. R. Bank, J. C. Campbell, S. J. Maddox, M. Ren, A.-K. Rockwell, M. E. Woodson, and S. D. March, Avalanche Photodiodes Based on the AlInAsSb Materials System, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, 3800407, 2018.

[*] See J. Beck, C. Wan, M. Kinch, J. Robinson, P. Mitra, R. Scritchfield, F. Ma, and J. Campbell, The HgCdTe Electron Avalanche Photodiode, *Journal of Electronic Materials*, vol. 35, pp. 1166–1173, 2006.

B. History- and Position-Dependent Parameters

The theory for the conventional APD (CAPD) set forth in Sec. 19.4A is predicated on the assumption that the probability that a carrier will effect an impact ionization is independent of both its ionization history and the location at which the ionization occurs. The ionization rate for the CAPD is thus taken to be the same at all times and locations within the multiplication region.

These simplifying assumptions are not always applicable, however, and other conditions can prevail:

- To garner sufficient energy so it can initiate an impact ionization, a newly generated carrier may need to travel some distance in the multiplication region, indicating that the probability of impact ionization depends on the carrier's ionization history.
- The multiplication region of an APD can be bandgap-engineered in such a way that it contains quantum wells, graded bandgaps, and/or other features that render the probability of impact ionization dependent on the carrier's location within the device.

We proceed to consider APDs with **history-dependent parameters** and **position-dependent parameters** in turn. Combinations of these features can be deliberately incorporated into APD designs to reduce gain noise and improve performance.

APDs with History-Dependent Parameters

In certain cases, a newly generated carrier in the multiplication region of an APD can cause an impact ionization only after traveling a certain distance that enables it to accumulate sufficient energy from the electric field. The carrier ionization probability is then zero immediately following its generation, and it remains zero over a distance known as the **dead space**. A more accurate portrayal considers the carrier ionization probability to slowly recover following its generation, in which case the recovery distance is more properly termed **sick space**. In either case, the ionization coefficient is clearly dependent on the carrier's ionization history.

Specially designed multilayer APD structures can offer high-gain, low-noise, and low-dark-current by appropriately tailoring their history-dependent parameters:

- *Dead space.* Dead space, which is inherent in the process of impact ionization, regularizes the locations at which impact ionizations occur, thereby enhancing the orderliness of the carrier-generation process, which reduces the gain noise. The effect of dead space is particularly pronounced when the multiplication region is thin and the number of multiplications is small.
- *Initial-energy effects.* Carriers traversing an appropriately designed field gradient before entering the multiplication region can garner substantial kinetic energy, thereby reducing the *initial* dead space in the multiplication region and thus further regularizing the impact ionizations and reducing the gain noise.

The theory of gain noise for APDs with history-dependent parameters is outlined in Sec. 19.6B.

APDs with Position-Dependent Parameters

Multilayer avalanche photodiodes can be fabricated with bandgaps and ionization coefficients that have arbitrary positional dependencies within each device layer. Special cases of such multilayer devices are conventional, separate absorption and multiplication, multiquantum-well, and superlattice APDs.

Specially designed multilayer APD structures can offer high gain, low noise, and low dark current by appropriately tailoring their position-dependent parameters:

- *Position-dependent field gradients.* A multilayer device can be designed such that a carrier traversing it encounters energy-band discontinuities that accelerate carriers at specific locations within the structure. The additional kinetic energy suddenly imparted to a carrier can then selectively enhance the probability of impact ionization.
- *Position-dependent ionization thresholds.* A multilayer device can be designed such that a carrier traversing it encounters a sudden decrease in the ionization threshold energy as it crosses from a layer of one material into a layer of another. A carrier with insufficient energy in the first layer is then more likely to cause an ionization upon entering the second layer.

The theory of gain noise for superlattice APDs (SAPDs), with their attendant position-dependent parameters, is provided in Sec. 19.6B.

C. Single-Photon and Photon-Number-Resolving Detectors

Single-photon detectors are able to detect individual photons. This capability is important in a broad variety of applications that include imaging, lidar, remote sensing, communications, astronomy, quantum optics, and quantum information. The performance of a single-photon detector is assessed on the basis of a number of parameters, including spectral sensitivity, detection efficiency, dark-count rate, timing jitter, maximum count rate, active area, operating temperature, and photon-number-resolving capability.

We consider in turn three solid-state detectors suitable for detecting and counting individual photons: single-photon avalanche diodes (SPADs), silicon photomultipliers (SiPMs), and superconducting single-photon detectors (SSPDs).

It is perhaps worth mentioning that there are several emerging technologies that may (or may not) prove useful for achieving single-photon and photon-number-resolving detection: 1) nonlinear-optical frequency up-conversion of IR photons to visible wavelengths where single-photon detection is more efficient; 2) SCISCM Si:As SAM devices operated at cryogenic temperatures; 3) superconducting nanowire single-photon detectors (SNSPDs) operated at cryogenic temperatures; and 4) quantum-dot and defect-based single-photon detectors.

Single-Photon Avalanche Diodes (SPADs)

Photomultiplier tubes (PMTs) have long been the workhorses of single-photon and photon-counting systems (Sec. 19.1A). Visible and near-infrared PMTs exhibit spectral sensitivities that extend from 150 to 1000 nm, detection efficiencies as high as 40%, dark-count rates as low as 100 counts/s, timing jitter of 300 ps, maximum count rates of 10 MHz, diameters ranging from mm to $1/2$ m, room-temperature operation, and limited photon-number-resolving capability.

As a convenient solid-state alternative to the PMT, single-photon detection can be achieved by making use of a **single-photon avalanche diode (SPAD)**, which is also known as a **Geiger-mode avalanche photodiode** since its operation is analogous to that of the Geiger counter used to detect ionizing radiation. The SPAD is an APD that is biased slightly above its avalanche breakdown voltage so that a single electron–hole pair generated by the absorption of a photon is sufficient to precipitate avalanche breakdown, creating a large current pulse that signifies the arrival of the photon. The detector response is binarized in this mode of operation, which serves to mitigate gain noise and circuit noise. The photon detection efficiency (PDE) η is the probability that an incident photon generates a detectable electrical current in the output circuit. Thermal, tunneling, and trapping processes in the semiconductor material also result in the generation of electron–hole pairs that precipitate avalanches, and these events give rise to a finite dark-count rate.

Each avalanche must be quenched to prepare the SPAD for the arrival of the next photon. This quenching may be carried out by passive or active means:

- *Passive quenching* is most simply achieved by incorporating a high-resistance series resistor in the SPAD circuit that develops a substantial voltage drop in response to the avalanche current pulse, which reduces the voltage across the diode and quenches the avalanche.
- *Active quenching* is more complex but allows for a substantially greater maximum photon count rate. A fast discriminator senses the steep onset of the avalanche current pulse across a series resistor and triggers circuitry that reduces the voltage across the diode to below breakdown, which quenches the avalanche. The voltage is then restored to a value above breakdown, and, after a **recovery time** of about 100 ns, the device is ready for the next photon. In practice, the recovery time depends on the characteristics of the device as well as on its ancillary circuitry. It can be abrupt or gradual, in which case it is represented by a **dead time** or **sick time**, respectively. The dead time limits the maximum count rate of the detector and is sometimes deliberately extended to suppress afterpulsing.

A number of materials are useful for fabricating SPADs in different wavelength regions:

- *Si SPADs* operate in the visible and near infrared ($350 < \lambda_o < 900$ nm). They offer photon detection efficiencies as high as 65%, dark-count rates as low as 25 counts/s, timing jitter of 400 ps, maximum count rates of 10 MHz, diameters of 100–200 μm, and room-temperature operation, but they do not offer photon-number-resolving capability.
- *InGaAs/InP SPADs*, which operate in the near infrared ($900 < \lambda_o < 1700$ nm) are the devices of choice in the optical fiber telecommunications band, but their performance is not nearly as good as that of Si SPADs. Because these devices have inordinately high dark-count rates (typically $> 20\,000$ counts/s), they are often cryogenically cooled and operated in a **gated Geiger mode**, which is suitable when the photon delivery time is known. They then offer photon detection efficiencies of 10%, gated dark-count rates of 100 counts/s, timing jitter of 400 ps, maximum count rates of 10 kHz, and diameters of 40 μm. *Ge-on-Si SPADs* are also sometimes used in this wavelength region but their performance is inferior to that of InGaAs/InP devices.
- *HgCdTe SPADs* are useful in the mid-infrared region while *GaN* and *SiC SPADs* have found use in the ultraviolet. SiC has the particular merit that it can tolerate high temperatures and hostile environments.

When operated individually, SPADs are typically only able to distinguish between the detection of *zero* photons and *one-or-more* photons, thus obviating their use for photon-number-resolving applications. This is because the detection of one photon leads to an avalanche breakdown indistinguishable from that initiated by more than one photon.

Silicon Photomultipliers (SiPMs)

Though a stand-alone SPAD cannot distinguish between the detection of a single photon and multiple photons, photon-number-resolving capability can be achieved by making use of multiple SPADs. In one configuration, a cascade of beamsplitters and suitable delays can be used to splay the multiphoton pulse out in time so that the constituent photons can be separately detected. Or, as with a microchannel plate (MCP), the multiphoton pulse can be made to broadly illuminate an array of SPADs so that the constituent photons are spatially splayed out among the individual SPADs.

The **silicon photomultiplier (SiPM)** is a Si SPAD spatial array in which a quenching resistor is inserted in series with each SPAD and the summed output is fed to an

amplifier. The *modus operandi* of this device mimics that of human scotopic vision. The SPADs that comprise the SiPM are analogous to the retinal rods in a receptive field, and the SiPM collection circuit is analogous to the associated retinal ganglion cell. A photon-number-resolving scheme such as this requires a sufficient number of SPADs to mitigate the possibility of two or more photons being absorbed in any one SPAD. Systems based on this architecture have the following salutary features:

- Immunity to dead time in the individual elements
- Enlarged photosensitive area
- Photon-number-resolving capability
- High sensitivity and gain

A typical SiPM is a two-terminal Si device with an area in the range of 1–50 mm^2 that contains between hundred and thousands of Si SPADs, each of area 10^2–10^4 μm^2. Typical devices exhibit gains in the range $\overline{G} \approx 10^6$ and overall decay times $\tau \approx 100$ ns. SiPMs find applications in medical imaging, quantum optics, astrophysics, and high-energy physics.

Comparison of SiPMs and PMTs. Since silicon photomultipliers and photomultiplier tubes serve similar functions, it is useful to consider their relative advantages:

Advantages of SiPMs over PMTs:

- Lower operating voltage (tens of volts)
- Higher spatial resolution
- Superior photon-number-resolving capability
- Smaller size
- More rugged construction
- Lower cost
- Immunity to magnetic fields (can be used in MRI and PET scanners)
- Absence of hysteresis
- Compatibility with semiconductor technology

Disadvantages of SiPMs with respect to PMTs:

- Operation restricted to the 300–900-nm spectral region
- Lower gain
- Longer response time
- Greater dark-count rate
- Lower sensitivity for small photon flux
- Higher excess noise factor (arising from crosstalk among elements)
- Inferior dynamic range
- Smaller photosensitive areas (although monolithic SiPM arrays are available)
- Greater sensitivity to temperature variations
- Greater susceptibility to damage from ionizing radiation

Silicon photomultipliers (SiPMs) and photomultiplier tubes (PMTs) thus play complementary roles in the domain of single-photon and photon-number-resolving detection. The choice of which device to select depends on the intended use and the prevailing experimental conditions.

CMOS-integrated SiPMs. The performance of SiPMs can be significantly enhanced by using CMOS technology to incorporate the constituent SPADs together with their circuitry on the same silicon substrate. The compactness of such structures reduces response time and time jitter, and thus leads to faster devices. Digital SiPMs that employ direct pulse counting based on digital readout have also been developed. Much like their analog counterparts, these devices, which are sometimes called **digital photon-counting (DPC)** devices, deliver a collective output.

Superconducting Single-Photon Detectors (SSPDs)

Single-photon detection and photon-number-resolved detection can also be achieved by making use of **superconducting single-photon detectors** (**SSPDs**) such as the **transition-edge sensor** (**TES**). At its superconducting critical temperature, the TES behaves as a microbolometer in which the absorption of a photon results in a steep temperature-induced change in resistance. This enables a precise measurement of the energy associated with an arbitrary number of absorbed photons, allowing highly resolved photon-number detection. TES devices can be constructed from materials such as tungsten-on-silicon or NbN-on-sapphire, and are sensitive from the infrared to the X-ray. TES devices exhibit jitters of 100 ns, bandwidths of 100 kHz, negligible dark counts, and photon detection efficiencies in excess of 90% when embedded in suitable optical-cavity structures. Their use in practice is complicated by the need for cryogenic operation and by their small active areas.

In one configuration, photolithographically patterned 40-nm-thick tungsten thin films are deposited on a Si substrate to form a 25 μm $\times$ 25 μm TES device. The substrate is cooled to approximately 60 mK, about half the 100-mK superconducting-to-normal transition temperature. The transition width is about 1 mK. A bias voltage across the thin film maintains the temperature in the transition region via Joule heating. An incident photon absorbed by the tungsten film raises its electron temperature, thereby increasing its resistance. The time integral of the associated decrease in current multiplied by the bias voltage yields the total photoelectric energy absorbed by the thin film within its (slow) 15-μs thermal relaxation time. For light of a specified wavelength, the number of photons incident within the thermal relaxation time is determined by establishing the total energy transferred to the detector within this time. The signal is read out of the detector using an array of DC superconducting quantum-interference devices (SQUIDs), which operate as current-sensitive amplifiers.

19.5 ARRAY DETECTORS

An individual photodetector registers the photon flux incident upon it as a function of time. Similarly, an array containing a large number of photodetectors simultaneously registers the photon fluxes (as functions of time) at many spatial points. **Array detectors** thus allow electronic versions of optical images to be formed. One type of array detector, the microchannel plate [Fig. 19.1-2(c)], has already been discussed.

Modern microelectronics technology permits the fabrication of many types of array detectors. These contain large numbers of photodetector elements, known as **pixels**, that can operate as photoconductors, photodiodes, avalanche photodiodes, or thermal detectors such as microbolometers. A 2D array of photosensitive elements designed to record an electronic version of an image at the focal plane of an imaging system is known as a **focal-plane array** (**FPA**). Two principal forms of readout circuitry are used to transport the signals generated at the FPA: charge-coupled device (CCD) technology and complementary metal-oxide-semiconductor (CMOS) technology.

Photodetector Elements

The pixels in a focal-plane array take many forms, as indicated by the following examples:

- Microbolometer arrays are often used in thermal imaging cameras. Incident photons cause an increase in the temperature of the illuminated elements; the accompanying change in resistance is recorded by external circuitry. These devices operate at ambient temperature and have come to the fore in recent years as their resolution and sensitivity have improved dramatically. Vanadium oxide (VOx) microbolometer arrays offer hundreds of thousands of pixels, each $\approx$ 25 μm in

size, and are sensitive in the mid-infrared region. These devices find extensive use in military and commercial applications.

- Photoconductive arrays are typically used in the mid-infrared region. A photon whose energy is greater than the bandgap energy in a semiconductor such as InSb or HgCdTe creates an electron–hole pair that contributes to the conductivity of the material.
- Arrays of extrinsic semiconductors, such as Ge:Ga, are useful for making photoconductive FPAs that are sensitive in the far-infrared. A photon places a donor electron into the conduction band (or a receptor hole into the valence band) so that it contributes to the conductivity.
- Quantum-well infrared photodetectors (QWIPs) are used in megapixel focal-plane arrays. A photon provides sufficient energy to lift an electron out of a quantum well so that it contributes to the conductivity. Far-infrared and mid-infrared images are provided by GaAs/AlGaAs and GaAs/InGaAs/AlGaAs elements.
- Arrays fabricated from compound-semiconductor p–i–n photodiodes, such as InGaAs, GeSn, and HgCdTe, are used in the visible and infrared. A photon whose energy is greater than the bandgap energy creates an electron–hole pair that contributes to the diode current.
- Schottky-barrier photodiode elements fabricated from metal–semiconductor junctions are used in highly versatile FPA cameras. A photon whose energy is greater than the Schottky barrier creates an electron–hole pair that contributes to the diode current. PtSi can be used for imaging in many spectral regions since it is sensitive over a broad band of wavelengths stretching from the near ultraviolet to about 6 μm in the mid infrared. In spite of the fact that it has low quantum efficiency in the infrared, PtSi is widely used since it is easily manufactured and highly stable.
- Avalanche-photodiode detectors fabricated from p–n junctions with multiplication regions have been crafted into array detectors. A photon whose energy is greater than the bandgap energy creates an electron–hole pair that enters a high-field semiconductor region that provides gain. The resulting sub-nanosecond electrical pulse can have an amplitude sufficient to directly trigger a digital CMOS circuit.
- Single-photon avalanche detectors (SPADs) fabricated from reverse-biased p–n junctions make use of multiplication regions operated in Geiger mode. A photon whose energy is greater than the bandgap energy creates an electron–hole pair that enters the high-field semiconductor region, thereby causing avalanche breakdown and the concomitant generation of a large current pulse.
- SPAD arrays have been developed in which the constituent SPAD outputs can be individually read out and the entire array can be read out as a frame with ps resolution. Devices that make use of active quenching, in conjunction with in-pixel signal processing and analog-to-digital conversion, are available. Arrays that make use of time-to-digital conversion make 3D single-photon imaging via time-resolved detection feasible.
- Photosensitive arrays can also be operated as heterodyne detectors in which conversion gain is provided by a local oscillator (Sec. 25.4).

Readout Circuitry

Two principal forms of readout circuitry are used to transport the signals from the photodetector elements to the camera display or output: charge-coupled device (CCD) technology and complementary metal-oxide-semiconductor (CMOS) technology.

CCD technology. A **charge-coupled device (CCD)** operates by converting photons to photoelectrons at each detector element (pixel) and storing the photoelectrons in local potential wells. At a specified time, the charge is sequentially transferred, via a

buried CCD channel that serves as a shift register, from one detector position to another until it is transported to one corner of the chip, where it is read out. Numerous electrode structures and clocking schemes have been developed for periodically reading out the charge accumulated at each CCD element and generating the electronic data stream representing the image. In comparison with CMOS imaging systems, CCD systems are typically more complex, require more power, and provide slower readout, but they find widespread use in scientific and medical applications where high quality imaging is mandatory. A variation on the theme of CCDs is the **intensified charge-coupled device (ICCD)**, which makes use of a microchannel-plate image intensifier (Sec. 19.1A) placed before the CCD. Another variation is the **electron-multiplying charge-coupled device (EMCCD)**, which, just prior to readout, employs a supplementary high-voltage, serial electron-multiplication register that contains several hundred electrodes and provides a gain of several thousand via secondary emission. ICCDs and EMCCDs have comparable performance, although each has its own particular merits. Single-photon and single-electron sensitivity has recently been achieved by making use of a Si "skipper CCD" that reduces readout noise by sampling the charge associated with each pixel multiple times.[†]

CMOS technology. **Complementary metal-oxide-semiconductor (CMOS)** manufacturing technology is widely used for fabricating electronic devices and integrated circuits. Because it consumes little power, has good noise immunity, and is relatively inexpensive, this technology has spurred the mass production of FPAs; photosensitive group-IV detection elements can be directly integrated with the readout circuitry. Each element in the photodetector array is individually linked to several metal-oxide-semiconductor field-effect transistors (MOSFETs) that amplify and read out the detected signal. Unlike the sequential read-out required for CCDs, the detector elements in a CMOS array are read out in parallel, which provides a significant speed advantage. A variation on the theme of CMOS is **scientific complementary metal-oxide-semiconductor (sCMOS)**, which makes use of a more advanced readout technology that offers increased imaging area, higher frame rate, greater dynamic range, higher quantum efficiency, and lower readout noise. sCMOS and EMCCD devices are competitive; the choice depends principally on the application at hand. However, at extremely low light levels (< 100 photons/pixel), and in the absence of background, a common rule-of-thumb dictates that EMCCD sensors are superior. Photon-number-resolving megapixel image sensors, operating without the benefit of cooling or avalanche gain, have also recently been developed.[‡]

19.6 NOISE IN PHOTODETECTORS

Photodetectors are responsive to photon flux (or optical power). In accordance with (19.1-4), an incident photon flux Φ (optical power $P = h\nu\Phi$) gives rise to a proportional photocurrent $i_p = \eta e\Phi = \mathsf{R}P$. In actuality, however, the electric current generated in a photodetector is a *random quantity* i that takes on values both below and above its mean value $\bar{\imath} \equiv i_p = \eta e\Phi = \mathsf{R}P$. (We use the symbols $\overline{x}$ and $\langle x \rangle$ interchangeably to represent the mean value of x.) The fluctuations of i, which are generally regarded as noise, are characterized by the variance of the current, $\sigma_i^2 = \langle (i - \bar{\imath})^2 \rangle$, or by its standard deviation $\sigma_i = \sqrt{\langle (i - \bar{\imath})^2 \rangle}$. For a current of zero mean ($\bar{\imath} = 0$), the standard deviation becomes the root-mean-square (RMS) value $\sigma_i = \sqrt{\langle i^2 \rangle}$.

[†] See J. Tiffenberg, M. Sofo-Haro, A. Drlica-Wagner, R. Essig, Y. Guardincerri, S. Holland, T. Volansky, and T.-T. Yu, Single-Electron and Single-Photon Sensitivity with a Silicon Skipper CCD, *Physical Review Letters*, vol. 119, 131802, 2017.

[‡] See J. Ma, S. Masoodian, D. A. Starkey, and E. R. Fossum, Photon-Number-Resolving Megapixel Image Sensor at Room Temperature Without Avalanche Gain, *Optica*, vol. 4, pp. 1474–1481, 2017.

Sources of noise. A number of sources of noise are inherent in the process of photodetection:

- *Photon Noise.* The most fundamental source of noise is associated with the random arrivals of the photons themselves, which are usually described by Poisson statistics, as discussed in Sec. 13.2C.
- *Photoelectron Noise.* In a photodetector with quantum efficiency $\eta < 1$, a single arriving photon generates a photoelectron–hole pair with probability η and fails to do so with probability $1 - \eta$. Because the deletion process is random it serves as a source of noise, as shown in Sec. 13.2D.
- *Gain Noise.* The amplification process that provides internal gain in certain types of photodetectors, such as photoconductors and APDs, is random. Each detected photon (photoelectron) generates a random number of carriers G, with an average value $\overline{G}$. The gain fluctuations depend on the nature of the amplification process, as will be elucidated in Sec. 19.6B.
- *Receiver Circuit Noise.* Various components in the electrical circuitry of an optical receiver, such as resistors and transistors, contribute to receiver circuit noise, as will be considered in Sec. 19.6C.

These four sources of noise are illustrated schematically in Fig. 19.6-1. The mean signal entering the detector (input optical signal) has an associated intrinsic photon noise. The photoeffect converts the photons into photoelectrons. In the process, the mean signal decreases by the factor η (the quantum efficiency). The associated photoelectron noise also decreases, but by a lesser amount than the signal; thus the signal-to-noise ratio of the photoelectron signal is lower than that of the incident photon signal. If a photodetector gain mechanism is present, it amplifies both the photoelectron signal and noise. Moreover, it introduces its own gain noise. Finally, circuit noise enters at the point of current collection.

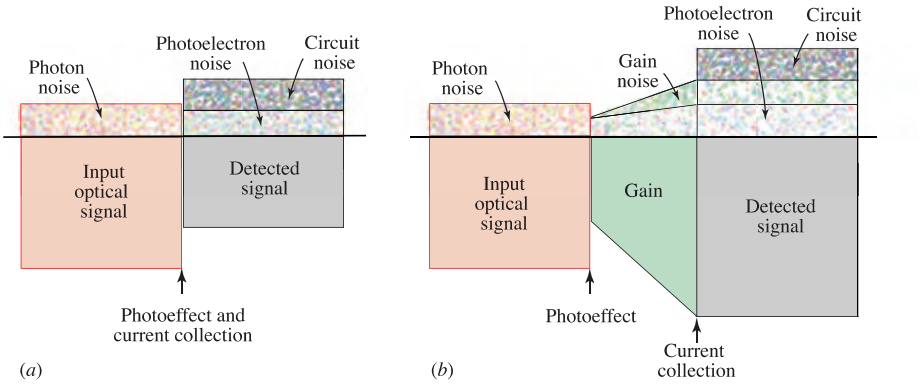


Figure 19.6-1 Input and detected signals along with various sources of noise for (a) a photodetector without gain, such as a p–i–n photodiode; and (b) a photodetector with gain, such as an avalanche photodiode.

Performance measures. As components of an information transmission system, photodetectors and optical receivers can be characterized by the following performance measures:

- The **signal-to-noise ratio** (SNR) of a random variable is defined as the ratio of the square of its mean to its variance. Thus, the SNR of the current i is $\text{SNR} = \overline{i}^2/\sigma_i^2$ while the SNR of the photon number $\overline{n}$ is $\text{SNR} = \overline{n}^2/\sigma_n^2$.

- The **minimum-detectable signal** is defined as the mean signal that yields unity SNR. In particular, the **noise-equivalent power (NEP)** is the signal power that yields unity SNR at a bandwidth of 1 Hz. The **specific detectivity** D^* is the reciprocal of the NEP after normalization by the square root of the bandwidth and the square root of the detector area.
- The **excess noise factor** F of a random variable is defined as the ratio of its mean-square to its square-mean. Thus, the excess noise factor of the photodetector gain G is $F = \langle G^2 \rangle / \langle G \rangle^2$.
- For an analog system, the **receiver sensitivity** is defined as the signal that corresponds to a prescribed value of the signal-to-noise ratio, $\mathrm{SNR} = \mathrm{SNR}_0$. While the minimum-detectable signal corresponds to a receiver sensitivity that provides $\mathrm{SNR}_0 = 1$, a higher value of SNR_0 is often specified to ensure a given level of accuracy (e.g., $\mathrm{SNR}_0 = 10^3$, corresponding to 30 dB). For a digital system, the receiver sensitivity is defined as the optical energy (or corresponding mean number of photons) per bit required to achieve a prescribed bit error rate, which is often set at $\mathrm{BER} = 10^{-9}$.
- The **bit error rate (BER)** is defined as the probability of error per bit in a digital optical receiver.

We proceed to derive expressions for the signal-to-noise ratio, as well as for some of the other performance measures discussed above, for photodetectors and optical receivers subject to the four key sources of noise highlighted earlier. Sources of noise that we do not explicitly consider include background noise and dark-current noise. **Background noise** is photon noise associated with light from extraneous optical sources (e.g., the sun, the stars) that manages to reach the photodetector. Background noise is particularly deleterious in detection systems that operate in the mid- and far-infrared spectral regions because of the copious thermal radiation emitted at these wavelengths by objects at room temperature (Fig. 14.4-4). Photodetectors also generate **dark-current noise**, which, as the name implies, is present even in the absence of light. Dark-current noise results from surface leakage current as well as from random electron–hole pairs generated by thermal and tunneling processes.

A. Photoelectron Noise

Photon Noise

As described in Sec. 13.2B, the photon flux associated with a fixed optical power P is inherently uncertain. The mean photon flux is $\Phi = P/h\nu$ (photons/s), but this quantity fluctuates randomly in accordance with a probability law that depends on the nature of the light source. The number of photons n counted in a time interval T is thus random with mean $\bar{n} = \Phi T$. For light from an ideal laser (Sec. 13.2C), or light from a multimode thermal source (Probs. 13.2-6–13.2-8), the photon number obeys the Poisson probability distribution, for which $\sigma_n^2 = \bar{n}$. If $\bar{n} = 100$, for example, the actual number of photons observed will lie approximately in the range 100 ± 10.

The photon-number signal-to-noise ratio $\mathrm{SNR} = \bar{n}^2/\sigma_n^2$ is therefore

$$\boxed{\mathrm{SNR} = \bar{n},} \qquad (19.6\text{-}1)$$

Poisson Photon-Number Signal-to-Noise Ratio

and the minimum-detectable photon number is $\bar{n} = 1$ photon. If the observation time $T = 1\ \mu$s and the wavelength $\lambda_o = 1.24\ \mu$m, this is equivalent to a minimum-detectable power of 0.16 pW. The receiver sensitivity for $\mathrm{SNR}_0 = 10^3$ (30 dB) is 1000 photons. If the time interval $T = 10$ ns, this is equivalent to a photon-flux sensitivity of 10^{11} photons/s or an optical power sensitivity of 16 nW at $\lambda_o = 1.24\ \mu$m.

Photoelectron Noise

A photon incident on a photodetector of quantum efficiency η generates a photoevent (i.e., creates a photoelectron–hole pair or liberates a photoelectron) with probability η, or fails to do so with probability $1 - \eta$. Photoevents are assumed to be selected at random from the photon stream. An incident mean photon flux Φ (photons/s) therefore results in a mean photoelectron flux $\eta\Phi$ (photoelectrons/s). The number of photoelectrons m detected in the time interval T is a random variable with mean

$$\overline{m} = \eta\overline{n}\,, \tag{19.6-2}$$

where $\overline{n} = \Phi T$ is the mean number of incident photons in the same time interval T. If the photon number is distributed in Poisson fashion, so too is the photoelectron number, as can be ascertained by using an argument parallel to that developed in Sec. 13.2D. It follows that the photoelectron-number variance is equal to its mean, i.e.,

$$\sigma_m^2 = \overline{m} = \eta\overline{n}\,. \tag{19.6-3}$$

The photoelectron noise is clearly not additive with the photon noise.

The underlying Poisson randomness inherent in the photon number, which constitutes a fundamental source of noise that must be contended with when using light to transmit a signal, thus results in a photoelectron-number signal-to-noise ratio

$$\boxed{\text{SNR} = \overline{m} = \eta\overline{n}\,,} \tag{19.6-4}$$

Photoelectron-Number Signal-to-Noise Ratio

in accord with (13.2-34). The minimum-detectable photoelectron number is $\overline{m} = \eta\overline{n} = 1$ photoelectron, corresponding to $1/\eta$ photons. The receiver sensitivity for $\text{SNR}_0 = 10^3$ is 1000 photoelectrons or $1000/\eta$ photons.

Photocurrent Noise

We now examine the properties of the electric current $i(t)$ induced in a circuit by a random photoelectron flux of mean $\eta\Phi$. The treatment we provide includes the effects of photon noise, photoelectron noise, and the characteristic time response of the detector and circuitry (filtering). Every photoelectron–hole pair generates a pulse of electric current of charge (area) e and time duration τ_p in the external circuit of the photodetector (Fig. 19.6-2). A photon stream incident on a photodetector therefore results in a stream of current pulses that add together to constitute the photocurrent $i(t)$. The randomness of the photon stream is transformed into a fluctuating electric current. If the incident photons are Poisson distributed, the current fluctuations are known as **shot noise**. More generally, for detectors with gain G the generated charge in each pulse is $q = Ge$.

Before providing an analytical derivation of the properties of the photocurrent $i(t)$, we examine the problem from a simplified perspective. Consider a photon flux Φ incident on a photoelectric detector of quantum efficiency η. Let the random number m of photoelectrons counted within a characteristic time interval $T = 1/2B$ (the resolution time of the circuit) generate a photocurrent $i(t)$, where t is the instant of time immediately following the interval T. For rectangular current pulses of duration T, the current and photoelectron-number random variables are related by $i = (e/T)m$. The photocurrent mean and variance are therefore given by

$$\overline{i} = \frac{e}{T}\,\overline{m} \tag{19.6-5}$$

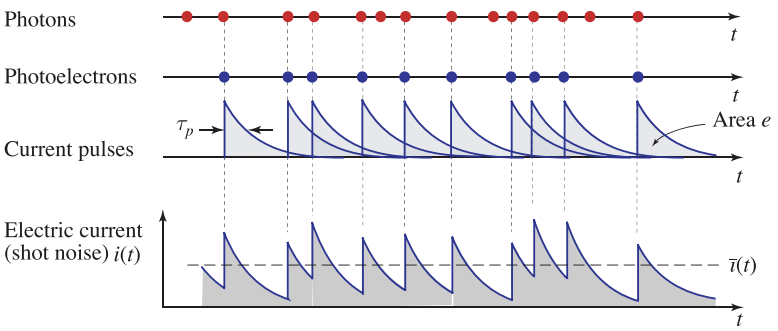


Figure 19.6-2 The photocurrent induced in a photodetector circuit comprises a superposition of current pulses, each associated with a detected photon. The individual pulses illustrated are exponentially decaying step functions but they can assume an arbitrary shape (see, e.g., Figs. 19.1-6(*b*) and 19.1-7). The superposition of the individual current pulses constitutes shot noise.

$$\sigma_i^2 = \left(\frac{e}{T}\right)^2 \sigma_m^2 , \tag{19.6-6}$$

respectively, where $\overline{m} = \eta\Phi T = \eta\Phi/2B$ is the mean number of photoelectrons collected in the time interval $T = 1/2B$. Substituting $\sigma_m^2 = \overline{m}$ for Poisson photoelectrons yields the shot-noise photocurrent mean and variance:

$$\boxed{\bar{\imath} = e\eta\Phi} \tag{19.6-7}$$
Photocurrent Mean

$$\boxed{\sigma_i^2 = 2e\bar{\imath}B.} \tag{19.6-8}$$
Photocurrent Variance

It follows that the signal-to-noise ratio of the shot-noise photocurrent, $\text{SNR} = \bar{\imath}^2/\sigma_i^2$, is

$$\boxed{\text{SNR} = \frac{\bar{\imath}}{2eB} = \frac{\eta\Phi}{2B} = \overline{m}.} \tag{19.6-9}$$
Photocurrent Signal-to-Noise Ratio

The current SNR is directly proportional to the photon flux Φ and inversely proportional to the electrical bandwidth of the circuit B. The result is identical to that for the photoelectron-number signal-to-noise ratio $\overline{m}$, as expected, since the circuit introduces no added randomness.

EXAMPLE 19.6-1. ***SNR and Receiver Sensitivity.*** For $\bar{\imath} = 10$ nA and $B = 100$ MHz, $\sigma_i \approx 0.57$ nA, corresponding to a signal-to-noise ratio $\text{SNR} = 310$ or 25 dB. An average of 310 photoelectrons are detected in every time interval $T = 1/2B = 5$ ns. The minimum-detectable photon flux is $\Phi = 2B/\eta$, and the receiver sensitivity for $\text{SNR}_0 = 10^3$ is $\Phi = 1000 \cdot (2B/\eta) = 2 \times 10^{11}/\eta$ photons/s.

□ **Derivation of the Photocurrent Mean and Variance.** We now proceed to prove (19.6-7) and (19.6-8) in the general case. Assume that a photoevent generated at $t = 0$ produces an electric pulse $h(t)$, of area e, in the external circuit. A photoevent generated at time t_1 then produces a displaced pulse, $h(t - t_1)$. Divide the time axis into incremental time intervals Δt so that the probability p that a photoevent occurs within an interval is $p = \eta\Phi\Delta t$. The electric current i at time t is written as

$$i(t) = \sum_l X_l\, h(t - l\Delta t), \tag{19.6-10}$$

where X_i assumes the value 1 with probability p, and 0 with probability $1 - p$. The variables $\{X_l\}$ are independent. The mean value of X_l is $0 \times (1 - p) + 1 \times p = p$. Its mean-square value is $\langle X_l^2 \rangle = 0^2 \times (1 - p) + 1^2 \times p = p$. The mean of the product $X_l X_k$ is p^2 if $l \neq k$, and p if $l = k$.

The mean and mean-square values of $i(t)$ are now determined via

$$\bar{\imath} = \langle i \rangle = \sum_l p\, h(t - l\Delta t) \tag{19.6-11}$$

$$\begin{aligned}\langle i^2 \rangle &= \sum_l \sum_k \langle X_l X_k \rangle\, h(t - l\Delta t)\, h(t - k\Delta t) \\ &= \sum \sum_{l \neq k} p^2\, h(t - l\Delta t)\, h(t - k\Delta t) + \sum_l p\, h^2(t - l\Delta t).\end{aligned} \tag{19.6-12}$$

Substituting $p = \eta\Phi\Delta t$, and taking the limit $\Delta t \to 0$ so that the summations become integrals, (19.6-11) and (19.6-12) yield, respectively,

$$\bar{\imath} = \eta\Phi \int_0^\infty h(t)\, dt = e\eta\Phi \tag{19.6-13}$$

$$\langle i^2 \rangle = (e\eta\Phi)^2 + \eta\Phi \int_0^\infty h^2(t)\, dt. \tag{19.6-14}$$

It follows that

$$\sigma_i^2 = \langle i^2 \rangle - \langle i \rangle^2 = \eta\Phi \int_0^\infty h^2(t)\, dt. \tag{19.6-15}$$

Defining

$$B = \frac{1}{2e^2} \int_0^\infty h^2(t)\, dt = \frac{\int_0^\infty h^2(t)\, dt}{2\left[\int_0^\infty h(t)\, dt\right]^2} \tag{19.6-16}$$

finally leads to (19.6-7) and (19.6-8). ■

The parameter B defined by (19.6-16) represents the device/circuit bandwidth. This is readily verified by noting that the Fourier transform of $h(t)$ is its transfer function $H(\nu)$. The area under $h(t)$ is simply $H(0) = e$. In accordance with Parseval's theorem (A.1-7), the area under $h^2(t)$ is equal to the area under the symmetric function $|H(\nu)|^2$, so that

$$B = \int_0^\infty \left| \frac{H(\nu)}{H(0)} \right|^2 d\nu. \tag{19.6-17}$$

In accordance with (A.2-10), the quantity B is therefore the power-equivalent spectral width of the function $|H(\nu)|$ (i.e., the bandwidth of the device/circuit combination). As an example, if $H(\nu) = 1$ for $-\nu_c < \nu < \nu_c$ and 0 elsewhere, (19.6-17) yields $B = \nu_c$.

These relations are applicable for all photoelectric detection devices without gain (e.g., phototubes and junction photodiodes). Use of the formulas requires knowledge of the bandwidth of the device, biasing circuit, and amplifier; B is determined by inserting the transfer function of the overall system into (19.6-17).

B. Gain Noise

Deterministic gain. The photocurrent mean and variance for a device with deterministic (fixed) gain G is obtained by replacing e with $q = Ge$ in (19.6-7) and (19.6-8), which leads to

$$\bar{\imath} = eG\eta\Phi = \frac{eG\eta P}{h\nu} \tag{19.6-18}$$

$$\sigma_i^2 = 2eG\bar{\imath}B = 2e^2G^2\eta B\Phi. \tag{19.6-19}$$

The signal-to-noise ratio, in accordance with (19.6-9), then becomes

$$\text{SNR} = \frac{\bar{\imath}}{2eGB} = \frac{\eta\Phi}{2B} = \bar{m}. \tag{19.6-20}$$

The SNR is independent of G because deterministic gain introduces no additional randomness. This is confirmed by observing that the mean current $\bar{\imath}$, along with its RMS value σ_i, are both multiplied by the same factor G.

Random gain. The simple results derived above do not apply when the gain itself is random, as is the case in photomultiplier tubes, photoconductors, and avalanche photodiodes. Appropriate expressions for the photocurrent mean and variance can be determined by modifying the derivation provided in the previous section. In particular, the electric current provided in (19.6-10) should be written as

$$i(t) = \sum_l X_l\, G_l\, h(t - l\Delta t), \tag{19.6-21}$$

where, as before, X_l takes the value 1 with probability $p = \eta\Phi\Delta t$, and 0 with probability $1 - p$. Now included in this expression is the random number G_l that represents the gain imparted to a photocarrier generated in the lth time slot, as shown in Fig. 19.6-3.

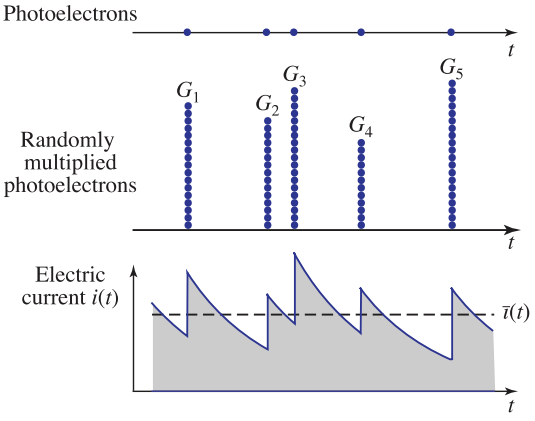


Figure 19.6-3 Each photoevent in a photodetector with gain generates a random number G_l of carriers that give rise to electrical current pulses of area eG_l. The total electric current in the detector circuit $i(t)$ is the superposition of these pulses.

If the random variable G_l has mean $\langle G\rangle \equiv \overline{G}$, and mean-square $\langle G^2\rangle$, an analysis analogous to that provided in (19.6-10)–(19.6-17) for Poisson photoelectrons yields

$$\boxed{\bar{\imath} = e\overline{G}\eta\Phi} \qquad \text{(19.6-22)}$$
Photocurrent Mean (Random Gain)

and

$$\boxed{\sigma_i^2 = 2e\overline{G}\bar{\imath}BF,} \qquad \text{(19.6-23)}$$
Photocurrent Variance (Random Gain)

where the **excess noise factor** F is defined as

$$\boxed{F = \frac{\langle G^2\rangle}{\langle G\rangle^2}.} \qquad \text{(19.6-24)}$$
Excess Noise Factor

The excess noise factor is related to the variance of the gain σ_G^2 by $F = 1+\sigma_G^2/\langle G\rangle^2$. In the special case of deterministic gain, $\sigma_G^2 = 0$ and $F = 1$, whereupon (19.6-23) reverts to (19.6-19). For random gain, we have $\sigma_G^2 > 0$ and $F > 1$; both of these quantities increase with the severity of the gain fluctuations. The resulting electric current i then exhibits fluctuations that are greater than those of shot noise.

In the presence of random gain, the current signal-to-noise ratio $\bar{\imath}^2/\sigma_i^2$ becomes

$$\boxed{\text{SNR} = \frac{\bar{\imath}}{2e\overline{G}BF} = \frac{\eta\Phi/2B}{F} = \frac{\overline{m}}{F},} \qquad \text{(19.6-25)}$$
Signal-to-Noise Ratio (Random Gain)

where $\overline{m}$ is the mean number of photoelectrons collected in the time $T = 1/2B$. The random-gain SNR is smaller than the deterministic-gain SNR by the factor F; the reduction is a consequence of gain randomness. It is clear that the excess noise factor F embodies the noise introduced by random gain in photodetectors.

EXAMPLE 19.6-2. ***Excess Noise Factor for a PMT.*** A photomultiplier tube achieves amplification by making use of secondary electron emission at its dynodes, as illustrated in Fig. 19.1-2(*b*). The excess noise factor is readily calculated by assuming that the secondary-emission gain random variable δ is identically distributed, with Poisson counting statistics and mean gain $\bar{\delta}$ for all dynodes except the first, which is endowed with gain $A\bar{\delta}$ (typically, $A \gg 1$). Under these conditions, the excess noise factor for the PMT is determined to be

$$F - 1 = \frac{1}{\overline{G}}\left[\frac{\overline{G}/A - 1}{(\overline{G}/A)^{1/N} - 1}\right], \qquad \text{(19.6-26)}$$

where $\overline{G} = A\bar{\delta}^N$ is the mean gain and N is the number of dynodes. The special case of all identical dynodes and large gain is considered in Prob. 19.6-3. Equation (19.6-26) is plotted vs. $\overline{G}$ as the dashed curves in Fig. 19.6-7 for $N = 1$, 4, and 10, assuming that $A = 10$ for the first (GaP) dynode. An estimate of the magnitude of the gain fluctuations may be obtained by considering a PMT in which the gain randomness yields an excess noise factor $F \approx 1.2$. Since $F = 1 + \sigma_G^2/\langle G\rangle^2$, the gain SNR $= \langle G\rangle^2/\sigma_G^2 = 1/(F-1) \approx 5$. If the PMT has a mean gain $\overline{G} = 10^6$, the standard deviation of the gain fluctuations is $\sigma_G = 10^6/\sqrt{5}$.

Excess Noise Factor for a Conventional APD

Conventional avalanche photodiodes (CAPDs) were examined in Sec. 19.4A. When photoelectrons are injected at the edge of a uniform multiplication region in a CAPD, the gain G of the device is given by (19.4-8). It depends on the electron ionization coefficient α_e and the ionization ratio $\mathrm{k} = \alpha_h/\alpha_e$, as well as on the width of the multiplication region w. The use of a similar (but more complex) analysis that incorporates the randomness associated with the gain process leads to an expression for the mean-square gain $\langle G^2 \rangle$ and the excess noise factor F. This more general derivation provides an expression for the mean gain $\overline{G}$ that is identical to that given in (19.4-8).

Calculations carried out by McIntyre in the mid-1960s reveal that the excess noise factor F for the CAPD is related to the mean gain and ionization ratio by

$$F = \mathrm{k}\overline{G} + (1 - \mathrm{k})\left(2 - \frac{1}{\overline{G}}\right). \tag{19.6-27}$$

Excess Noise Factor (Conventional APD)

This formula is plotted in Fig. 19.6-4 with the ionization ratio k as a parameter.

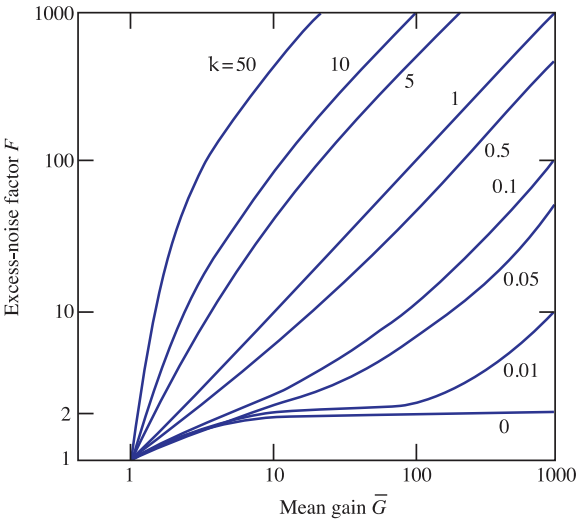


Figure 19.6-4 Excess noise factor F for a conventional APD (CAPD) with a uniform multiplication region, under electron injection, as a function of the mean gain $\overline{G}$, for different values of the ionization ratio k. For hole injection, $1/\mathrm{k}$ replaces k.

Equation (19.6-27) is valid when electrons are injected at the edge of the multiplication region, but both electrons and holes have the capacity to initiate impact ionizations. If only holes are injected, the same expression applies, provided that k is replaced by $1/\mathrm{k}$. Gain noise is minimized by injecting the carrier with the higher ionization coefficient, and by fabricating a structure with the lowest possible value of k if electrons are injected, or the highest possible value of k if holes are injected. In short, the ionization coefficients for the two carriers should be as different as possible. Equation (19.6-27) is said to be valid under conditions of **single-carrier-injection double-carrier multiplication (SCIDCM)** since both types of carrier have the capacity to impact ionize, even when only one type is injected. If electrons and holes are injected simultaneously, the overall result is the sum of the two partial results.

The gain noise associated with a CAPD arises from two sources: the randomness in the locations at which ionizations occur, and the feedback process associated with the fact that both kinds of carrier can produce impact ionizations. The first of these noise sources is present even when only one kind of carrier can multiply; it leads to $F - 1 = 1 - 1/\overline{G}$ (which is ≈ 1 for large values of the mean gain $\overline{G}$) as is apparent by setting $\mathrm{k} = 0$ in (19.6-27). This result is plotted vs. $\overline{G}$ as the dotted curve in Fig. 19.6-7 for purposes of comparison with the PMT and the staircase APD. The second source of noise, the feedback process, is potentially more detrimental since it can result in a far greater value of F.

EXAMPLE 19.6-3. ***Excess Noise Factor for a Si SAM APD.*** A separate absorption and multiplication (SAM) Si reach-through APD such as that depicted in Fig. 19.4-6 has peak sensitivity at a wavelength of 800 nm, quantum efficiency $\eta = 0.8$, and a gain–bandwidth product $GB = 350$ GHz. This SCIDCM device, which makes use of electron injection, has a mean gain $\overline{G} = 50$ and an ionization ratio $\mathrm{k} = 0.02$. Equation (19.6-27) yields $F \approx 3$ so that the gain mechanism reduces the signal-to-noise ratio by a factor of 3 while increasing the mean detected current by a factor of 50. Silicon APDs are sensitive over a wavelength range that stretches from 450 to 1 100 nm and can attain gains as high as 1000, depending on the device structure and reverse-bias voltage. In the presence of circuit noise the use of an APD can serve to increase the overall system SNR, as discussed in Sec. 19.6D.

EXAMPLE 19.6-4. ***Excess Noise Factor for an AlInAsSb SACM APD.*** The separate absorption, charge, and multiplication (SACM) $Al_{0.4}In_{0.6}As_{0.3}Sb_{0.7}/Al_{0.7}In_{0.3}As_{0.3}Sb_{0.7}$ APD considered in Example 19.4-4 makes use of absorption and multiplication layers with low and high Al content, corresponding to small and large energy bandgaps, respectively. This device is sensitive over telecommunications-band wavelengths ($1.3 \leq \lambda_o \leq 1.6$ μm). The ionization ratio $\mathrm{k} = 0.01$ for this SCIDCM APD is comparable to that of Si. For a mean gain $\overline{G} = 10$, (19.6-27) yields $F \approx 2$, indicating that the gain mechanism reduces the signal-to-noise ratio by a factor of 2 and increases the mean detected current by a factor of 10. As indicated above, the use of an APD can serve to increase the overall SNR in the presence of circuit noise (Sec. 19.6D).

Excess Noise Factor for APDs with History-Dependent Parameters

APDs with history-dependent parameters were introduced in Sec. 19.4B. A theory of APD noise that accommodates dead space, along with initial carrier-energy and impact-ionization threshold-energy effects, can be cast in the form of recurrence relations for the first and second moments, as well as the probability distribution, of the numbers of electrons and holes. These random variables are deterministically related to the random gain. Numerical solutions provide the mean gain and excess noise factor for arbitrary values of dead space and multiplication-region width. The theory properly predicts the performance of APDs in which history-dependent parameters play a role.

An example of the energy-band diagram for an APD with history-dependent parameters tailored to improve its performance is displayed in Fig. 19.6-5. Two thin multiplication layers, with relatively low threshold energy, surround a layer with higher threshold energy. Impact ionization is enhanced at the edges of the twin multiplication layers and is suppressed in the central region, which serves to impart energy to the carriers in transit. The materials are chosen so that hole-induced ionization is discouraged. The performance improvement is illustrated in Example 19.6-5.

EXAMPLE 19.6-5. ***Excess Noise Factor for a GaAs/AlGaAs APD Influenced by Dead Space.*** A very thin heterostructure APD similar to that displayed in Fig. 19.6-5 has a multiplication region comprising two 50-nm-thick layers of GaAs surrounding an 85-nm-thick layer of

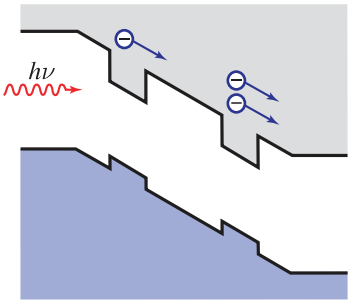


Figure 19.6-5 Energy-band diagram of a low-noise heterostructure APD with history-dependent parameters, under reverse-bias conditions.

$Al_{0.6}Ga_{0.4}As$. The measured excess noise factor is $F \approx 2.5$ at a mean gain $\overline{G} = 20$. In contrast, the excess noise factor predicted by (19.6-27) for a bulk GaAs homojunction APD ($k \approx 0.75$) is $F \approx 15.5$ at $\overline{G} = 20$ (Fig. 19.6-4). The noisiness of this heterostructure device is thus substantially lower than that predicted by the bulk theory, which ignores dead space as well as initial-energy and impact-ionization threshold-energy effects. Evidently, such effects materially reduce gain noise and must be accommodated when modeling thin-multiplication-region APDs. Other heterostructure configurations, such as a centered-well configuration, can exhibit even lower values of F at small values of $\overline{G}$.

Excess Noise Factor for APDs with Position-Dependent Parameters

Multilayer APDs with arbitrary structure and position-dependent parameters were introduced in Sec. 19.4B. A special case of this class of devices, the **staircase avalanche photodiode**, has the energy-band diagram displayed in Fig. 19.6-6. A three-stage device is illustrated under both unbiased and reverse-biased conditions. The bandgap is compositionally graded over a short distance, from a low value of E_{g1} to a high value of E_{g2}. Because of the material properties, hole-induced ionizations are discouraged. Other potential advantages of these devices include the discrete locations of the multiplications (at the jumps in the conduction band edges), the low operating voltage (which minimizes tunneling), and the fast response time (resulting from the reduced avalanche buildup time).

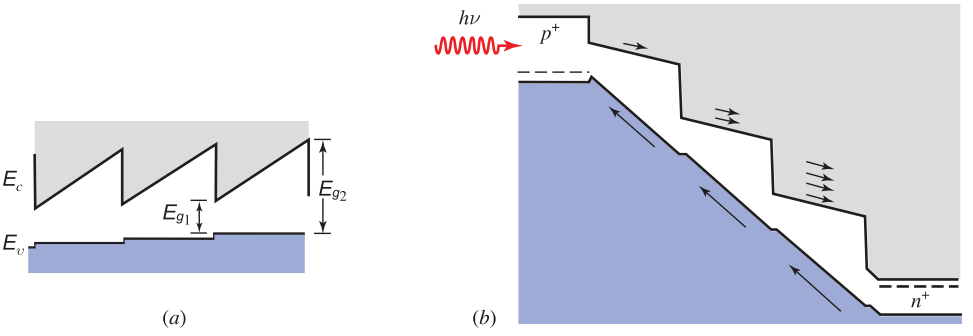


Figure 19.6-6 Energy-band diagram of a bandgap-engineered staircase avalanche photodiode under (*a*) unbiased and (*b*) reverse-biased conditions. The conduction-band steps encourage electron ionizations at discrete locations. (Adapted from F. Capasso, W.-T. Tsang, and G. F. Williams, Staircase Solid-State Photomultipliers and Avalanche Photodiodes with Enhanced Ionization Rates Ratio, *IEEE Transactions on Electron Devices*, vol. ED-30, pp. 381–390, Fig. 1 ©1983 IEEE.)

Under single-carrier-injection single-carrier-multiplication (SCISCM) conditions, the mean gain $\overline{G}$ of the staircase APD (Prob. 19.6-2) is

$$\overline{G} = (1+P)^N \tag{19.6-28}$$

and the excess noise factor is given by[†]

$$F = \frac{1}{\overline{G}} + \frac{2}{\overline{G}^{(1/N)}} - \frac{2}{\overline{G}^{(1+1/N)}}, \tag{19.6-29}$$

(19.6-29) Excess Noise Factor (Staircase APD)

where P is the probability of impact ionization at each stage and N is the number of stages. Equation (19.6-29) is plotted as the solid curves in Fig. 19.6-7 (as the **modified excess noise factor** $F-1$ vs. $\overline{G}$) for $N = 5$ and 10.

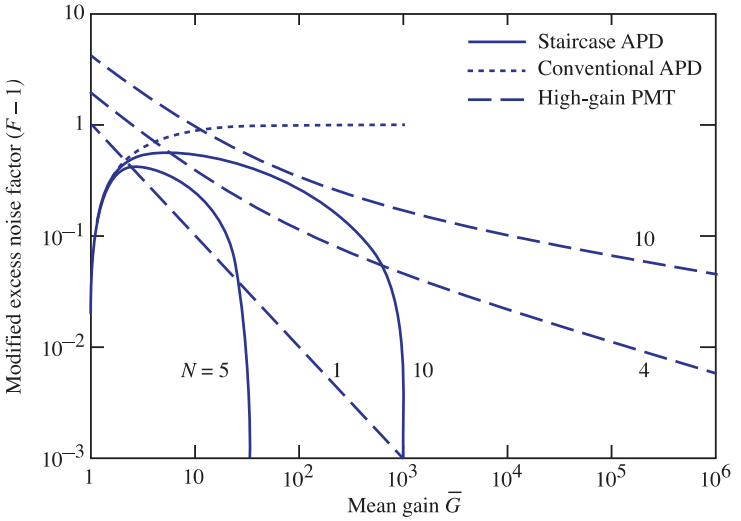


Figure 19.6-7 Modified excess noise factor $F-1$ versus mean gain $\overline{G}$ for three photodetectors: *The SCISCM Staircase APD*: The *solid curves* represent (19.6-29) with $N = 5$ and 10. *The SCISCM Conventional APD*: The *dotted curve* represents (19.6-27) with $k = 0$. *The SCISCM High-Gain-First-Dynode PMT*: The *dashed curves* represent (19.6-26) with $A = 10$ and $N = 1$, 4, and 10. The *modified excess noise factor* $F-1$ is displayed because it is conveniently plotted on double-logarithmic coordinates. The results presented represent optimal noise behavior for all three detectors. Though the excess noise factor of the ideal staircase APD always lies below that of the conventional APD, the difference is not large since $F < 2$ for both devices. The PMT offers high gain and excellent noise performance since the electrons travel in vacuum and it is a single-carrier device; however, its sensitivity is typically restricted to wavelengths shorter than about 1 μm and its quantum efficiency and responsivity are limited. (Adapted from M. C. Teich, K. Matsuo, and B. E. A. Saleh, Excess Noise Factors for Conventional and Superlattice Avalanche Photodiodes and Photomultiplier Tubes, *IEEE Journal of Quantum Electronics*, vol. QE-22, pp. 1184–1193, Fig. 3 ©1986 IEEE.)

Taking (19.6-29) to the limit as $N \to \infty$ leads to $F = 2 - 1/\overline{G}$, which is identical to the result obtained for the conventional APD under SCISCM conditions [(19.6-27)

[†] See K. Matsuo, M. C. Teich, and B. E. A. Saleh, Noise Properties and Time Response of the Staircase Avalanche Photodiode, *Journal of Lightwave Technology*, vol. LT-3, pp. 1223–1231, 1985 [simultaneously published in *IEEE Transactions on Electron Devices*, vol. ED-32, pp. 2615–2623, 1985].

with k = 0], as expected. Though we deal with the graded-gap staircase APD for purposes of illustration, (19.6-28) and (19.6-29) are applicable for any **superlattice avalanche photodiode (SAPD)** in which the carrier transport is perpendicular to the superlattice planes. In such structures, the carriers encounter a potential discontinuity at the heterointerfaces at each period of the multilayer structure.

Initial attempts to operate a staircase APD using GaAs/AlGaAs were impeded by insufficiently large conduction-band offsets and small energy separations between the direct and indirect band-structure valleys, so that the merits of the staircase structure could not be unequivocally demonstrated. However, recent work with the AlInAsSb/GaSb system has shown that the staircase concept is feasible (Example 19.6-6). The development of low-noise APDs based on narrow-bandgap semiconductor materials is a worthy enterprise since such structures can serve as "solid-state photomultipliers" at infrared wavelengths, where night-vision and thermal-imaging applications abound.

EXAMPLE 19.6-6. ***Excess Noise Factor for an AlInAsSb/GaSb Staircase APD.*** A one-stage staircase APD operates on the basis of photogenerated electrons in a wide-bandgap ($E_g = 1.16$ eV) AlInAsSb injector region entering a thin, narrow-bandgap ($E_g = 0.25$ eV) InAsSb multiplication region.[†] The conduction-band-edge energies at the interface of the materials differ by ≈ 0.6 eV, which is sufficient to permit impact ionization in the InAsSb with high probability, since its ionization threshold energy is ≈ 0.4 eV. The impact ionization is observed as a current gain of $\overline{G} \approx 1.8 \pm 0.2$ that persists over a broad range of excitation wavelengths, excitation intensities, reverse-bias voltages, and operating temperatures. Monte Carlo simulations confirm the presence of robust impact ionization by electrons within the thin InAsSb layer and essentially none by holes. The device thus behaves as a SCISCM staircase APD. In accordance with (19.6-28) and (19.6-29) for $N = 1$, we deduce that $P = \overline{G} - 1 \approx 0.8 \pm 0.2$ and $F = 3/\overline{G} - 2/\overline{G}^2 \approx 1$, respectively, over the range of observed values of $\overline{G}$.

Excess Noise Factor for Dark Current in the Multiplication Region

The output current of an APD fluctuates in the presence of light as well as in its absence. The dark-current noise arises from effects that are both external to, and internal to, the depletion and multiplication regions of the device. The surface-leakage dark current bypasses both regions and thus is not subject to gain noise. However, carrier pairs randomly generated by tunneling or thermal processes in the interior of many conventional and multilayer APDs are subject to multiplication, and thus to gain noise, much as for photogenerated carrier pairs. The dark carriers generated *within the multiplication region* are randomly distributed throughout it so that they experience a smaller mean gain, and a larger excess noise factor, than the carriers injected at its edge. Photogenerated carriers produced by light that impinges on the multiplication region are also subject to this increased excess noise. To minimize this effect, multiplication regions are generally fabricated using semiconductor materials of higher bandgap, which serves to limit tunneling and thermal processes, and the ensuing dark current. Expressions for the mean gain and excess noise factor for dark carriers generated within the multiplication region, as well as for dark and photogenerated carriers generated in the depletion region, and arbitrary superpositions thereof, have been set forth for multilayer APDs of arbitrary structure.[‡]

[†] See M. Ren, S. Maddox, Y. Chen, M. Woodson, J. C. Campbell, and S. Bank, AlInAsSb/GaSb Staircase Avalanche Photodiode, *Applied Physics Letters*, vol. 108, 081101, 2016.

[‡] See N. Z. Hakim, B. E. A. Saleh, and M. C. Teich, Generalized Excess Noise Factor for Avalanche Photodiodes of Arbitrary Structure, *IEEE Transactions on Electron Devices*, vol. 37, pp. 599–610, 1990.

C. Circuit Noise

Yet additional noise is introduced by the electronic circuitry associated with an optical receiver. Circuit noise results from the thermal motion of charged carriers in resistors and other dissipative elements (thermal noise) and from fluctuations of charge carriers in transistors used in the receiver amplifier, as well as from $1/f$-type effects.

Thermal Noise

Thermal noise (also called **Johnson noise** or **Nyquist noise**) arises from the random motion of mobile carriers in resistive electrical materials at finite temperatures; this gives rise to a random electric current $i(t)$ even in the absence of an external electrical power source. The thermal electric current in a resistance R is a random function $i(t)$ whose mean value $\langle i(t)\rangle = 0$. The variance of the current σ_i^2, which is the same as its mean-square value since the mean vanishes, increases with the temperature T.

Using the results of a derivation based on statistical mechanics, to be presented shortly, a resistance R at temperature T exhibits a random electric current $i(t)$ characterized by a power spectral density

$$S_i(f) = \frac{4}{R}\frac{hf}{\exp(hf/kT) - 1}, \tag{19.6-30}$$

where f is the frequency. In the region $f \ll kT/h$, which is the frequency range of principal interest since $kT/h = 6.24$ THz at room temperature, $\exp(hf/kT) \approx 1 + hf/kT$, so that

$$S_i(f) \approx 4kT/R. \tag{19.6-31}$$

The variance of the electric current is the integral of the power spectral density over all frequencies within the bandwidth B of the circuit, i.e.,

$$\sigma_i^2 = \int_0^B S_i(f)\, df. \tag{19.6-32}$$

Hence, for $B \ll kT/h$ this leads to

$$\boxed{\sigma_i^2 \approx 4kTB/R.} \tag{19.6-33}$$

Thermal Noise Current Variance (Resistance R)

Thus, as shown in Fig. 19.6-8, a resistor R at temperature T in a circuit of bandwidth B behaves as a noiseless resistor in parallel with a zero-mean noise current source with an RMS value σ_i specified by (19.6-33).

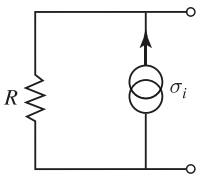


Figure 19.6-8 A resistance R at temperature T is equivalent to a noiseless resistor in parallel with a noise current source whose mean is zero and whose variance is given by $\sigma_i^2 = \langle i^2\rangle \approx 4kTB/R$, where B is the circuit bandwidth.

EXAMPLE 19.6-7. ***Thermal Noise in a Resistor.*** A 1-$k\Omega$ resistor at $T = 300°$ K, in a circuit of bandwidth $B = 100$ MHz, exhibits an RMS thermal noise current $\sigma_i \approx 41$ nA.

□ ⋆**Derivation of the Power Spectral Density of Thermal Noise.** We derive (19.6-30) by showing that the electrical power associated with the thermal noise in a resistance is identical to the electromagnetic power radiated by a one-dimensional blackbody. The factor $hf/[\exp(hf/kT) - 1]$ in (19.6-30) is recognized as the mean energy $\overline{E}$ of an electromagnetic mode of frequency f (the symbol ν is reserved for optical frequencies) in thermal equilibrium at temperature T [see (14.4-8)]. Equation (19.6-30) may therefore be written as $S_i(f)R = 4\overline{E}$. The electrical power dissipated by a noise current i passing through a resistance R is $\langle i^2 \rangle R = \sigma_i^2 R$, so that $S_i(f)R$ represents the electrical power density (per Hz) dissipated by the noise current $i(t)$ through R.

We now proceed to demonstrate that $4\overline{E}$ is the power density radiated by a one-dimensional blackbody. As discussed in Sec. 14.4B, an atomic system in thermal equilibrium with the electromagnetic modes in a cavity radiates a spectral energy density $\varrho(\nu) = M(\nu)\overline{E}$, where $M(\nu) = 8\pi\nu^2/c^3$ is the three-dimensional density of modes, and the spectral intensity density is $c\varrho(\nu)$. Though the charge carriers in a resistor move in all directions, only motion in the direction of the circuit current flow contributes. The density of modes in a single dimension is $M(f) = 4/c$ modes/m-Hz [see (11.1-10)] so that the corresponding energy density is $\varrho(f) = M(f)\overline{E} = 4\overline{E}/c$ and the radiated power density is $c\varrho(f) = 4\overline{E}$, as promised. ■

1/f Noise

Another form of noise associated with some components that comprise the electronic circuitry of an optical receiver exhibits a power spectral density with a power-law form: $S(f) \approx (f/f_0)^{-\alpha}$. The multiplicative constant f_0^{α} determines the absolute strength of the fluctuations at all frequencies while the exponent $-\alpha$ characterizes the relative strength of the fluctuations at different frequencies. Noise of this form is typically referred to as $1/f^{\alpha}$ noise or $1/f$-type noise. In the particular case when $\alpha = 1$, common appellations are $1/f$ noise, **excess noise**, **flicker noise**, and **pink noise**. The latter terminology arises from an analogy with visible light. For a spectrum of the form $S(f) \propto 1/f$, each octave is endowed with equal power so that lower frequencies ("red") are weighed more heavily than higher frequencies ("blue"), resulting in a spectrum with a pink tinge. Since no strict standard for this nomenclature exists, however, the foregoing descriptions are all used to describe $1/f^{\alpha}$ noise when α is roughly in the vicinity of unity.

Fluctuations of this form are ubiquitous in electronics and photonics. $1/f^{\alpha}$ noise was discovered, along with thermal noise, in early studies of low-frequency circuits. Such fluctuations are also widely observed in components, materials, and devices, including resistors, semiconductors, metal films, superconductors, thermionic-emission devices, and junction devices. In electronics, the range of frequencies over which such behavior is manifested can stretch over 12 orders of magnitude or more, and α typically lies between 0.8 and 1.4. From a practical perspective, devices and systems subject to $1/f^{\alpha}$ noise are often operated at frequencies that are sufficiently high so that this noise is negligible.

The origins of $1/f^{\alpha}$ noise remain obscure for many components, materials, and devices. The underlying mechanism is often associated with fluctuations of the number, or the mobility, of the charge carriers, but other causes have been postulated. $1/f$-type noise is thought by some to be a surface effect whereas others attribute it to bulk behavior. Moreover, behavior of this kind is not restricted to simple systems; complex systems also exhibit $1/f^{\alpha}$ noise. Although ubiquitous, this type of noise is not universal; it is not present, for example, in wire-wound resistors.

Circuit-Noise Parameter: Resistance-Limited and Amplifier-Limited Optical Receivers

It is convenient to lump the various sources of circuit noise in an optical receiver (thermal noise in resistors as well as noise in transistors and other circuit devices)

into a single random current source i_r at the receiver input that produces the same total noise at the receiver output (Fig. 19.6-9). The mean value of i_r is zero while its variance σ_r^2 depends on temperature, receiver bandwidth, circuit parameters, and device type.

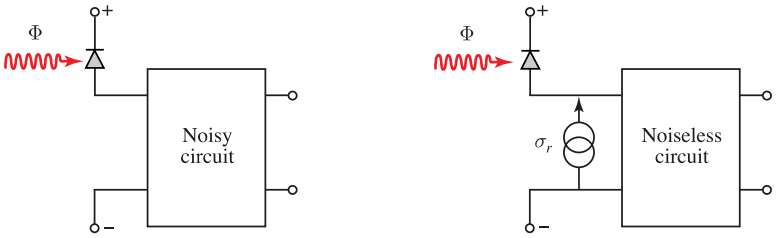


Figure 19.6-9 A noisy receiver circuit (left) can be replaced by a noiseless receiver circuit that has a single random current source with RMS value σ_r at its input (right).

It is also convenient to define a dimensionless circuit-noise parameter

$$\sigma_q = \frac{\sigma_r T}{e} = \frac{\sigma_r}{2eB}, \tag{19.6-34}$$

where B is the receiver bandwidth and $T = 1/2B$ is the receiver resolution time. Since σ_r is the RMS value of the noise current, σ_r/e is the RMS electron flux (electrons/s) arising from circuit noise, and $\sigma_q = (\sigma_r/e)T$ thus represents the RMS number of circuit-noise electrons collected in the time T. The circuit-noise parameter σ_q is a figure of merit that characterizes the quality of the optical receiver circuit, as will become apparent in Sec. 19.6D.

An optical receiver comprising a photodiode in series with a load resistor R_L, followed by an amplifier, is illustrated in Fig. 19.6-10. This simple receiver is said to be *resistance-limited* if the circuit-noise current arising from thermal noise in the load resistor substantially exceeds noise contributions from other sources. The amplifier may then be regarded as noiseless and the circuit-noise mean-square current is simply $\sigma_r^2 = 4kTB/R_L$. The circuit-noise parameter defined by (19.6-34) is therefore

$$\sigma_q = \sqrt{\frac{kT}{e^2 R_L B}}, \tag{19.6-35}$$

Circuit-Noise Parameter (Resistance-Limited Receiver)

which is inversely proportional to the square-root of the bandwidth B.

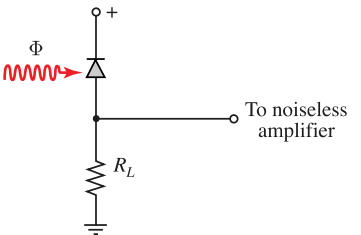


Figure 19.6-10 Resistance-limited optical receiver.

EXAMPLE 19.6-8. ***Circuit-Noise Parameter for a Resistance-Limited Receiver.*** At room temperature, a resistance $R_L = 50\ \Omega$ in a circuit of bandwidth $B = 100$ MHz generates a random current of RMS value $\sigma_r = 0.18\ \mu$A. This corresponds to a circuit-noise parameter $\sigma_q \approx 5700$.

A receiver using a well-designed low-noise amplifier can yield a smaller circuit-noise parameter than a resistance-limited receiver. Consider a receiver using an *FET amplifier*. If the noise arising from the high input resistance of the amplifier can be neglected, the receiver is limited by thermal noise in the channel between the FET source and drain. With the use of an equalizer to boost the high frequencies attenuated by the capacitive input impedance of the circuit, for typical circuit component values the circuit-noise parameter at room temperature turns out to be

$$\sigma_q \approx \frac{\sqrt{B}}{100} \quad (B \text{ in Hz}) . \tag{19.6-36}$$

(19.6-36) Circuit-Noise Parameter (FET Amplifier Receiver)

For example, if $B = 100$ MHz, then $\sigma_q = 100$, which is significantly smaller than the circuit-noise parameter associated with a 50-Ω resistance-limited amplifier of the same bandwidth (Example 19.6-8). The circuit-noise parameter σ_q increases with B because of the effect of the equalizer.[†]

A receiver that makes use of a *bipolar-transistor amplifier*, on the other hand, has a circuit-noise parameter σ_q that is independent of the bandwidth B over a wide range of frequencies. For bandwidths between 100 MHz and 2 GHz, σ_q is typically ≈ 500, provided that appropriate transistors are used and that they are optimally biased.

D. Signal-to-Noise Ratio and Analog Receiver Sensitivity

The simplest measure of the quality of reception in an analog communication system is the signal-to-noise ratio. The SNR of the current at the input to the noiseless circuit represented in Fig. 19.6-9 is the ratio of the square of the mean current to the sum of the variances of the constituent sources of noise:

$$\text{SNR} = \frac{\bar{\imath}^2}{2e\overline{G}\bar{\imath}BF + \sigma_r^2} = \frac{\left(e\overline{G}\eta\Phi\right)^2}{2e^2\overline{G}^2\eta B\Phi F + \sigma_r^2} . \tag{19.6-37}$$

(19.6-37) Optical Receiver Signal-to-Noise Ratio

The leftmost terms in the denominators represent photoelectron and gain noise [see (19.6-23)] while the rightmost terms represent circuit noise. For a detector devoid of gain, we have $\overline{G} = 1$ and $F = 1$. The noiseless circuit in Fig. 19.6-9 does not alter the signal-to-noise ratio even if it provides amplification.

EXERCISE 19.6-1

Signal-to-Noise Ratio of a Resistance-Limited Receiver. Assume that the optical receiver portrayed in Fig. 19.6-10 makes use of an ideal p–i–n photodiode ($\eta = 1$) and that the resistance $R_L = 50\ \Omega$ at $T = 300°$ K. The bandwidth is $B = 100$ MHz. At what value of the photon flux Φ is the photoelectron-noise current variance equal to the resistor thermal-noise current variance? What is the corresponding optical power at $\lambda_o = 1550$ nm?

[†] See S. D. Personick, *Optical Fiber Transmission Systems*, Plenum, 1981, Sec. 3.4; note that the parameter σ_q is equivalent to $Z/2$ in this reference.

It is useful to recast the SNR in (19.6-37) in terms of the mean number of detected photons $\overline{m}$ in the resolution time of the receiver $T = 1/2B$,

$$\overline{m} = \eta \Phi T = \frac{\eta \Phi}{2B}, \tag{19.6-38}$$

and the circuit noise-parameter $\sigma_q = \sigma_r/2Be$. The resulting expression is simply

$$\boxed{\text{SNR} = \frac{\overline{G}^2 \overline{m}^2}{\overline{G}^2 F \overline{m} + \sigma_q^2}.} \tag{19.6-39}$$

Signal-to-Noise Ratio for an Optical Receiver

Equation (19.6-39) has a straightforward interpretation. The numerator is the square of the mean number of multiplied photoelectrons detected in the receiver resolution time $T = 1/2B$. The denominator is the sum of the variances of the number of photoelectrons and the number of circuit-noise electrons collected in T.

For a photodiode without gain we have $\overline{G} = F = 1$, so that (19.6-39) reduces to

$$\boxed{\text{SNR} = \frac{\overline{m}^2}{\overline{m} + \sigma_q^2}.} \tag{19.6-40}$$

Signal-to-Noise Ratio (Unity-Gain Receiver)

The relative magnitudes of $\overline{m}$ and σ_q^2 establish the relative importance of photoelectron noise and circuit noise. The manner in which the parameter σ_q characterizes the circuit's performance as an optical receiver is now apparent. For example, if $\sigma_q = 100$, then circuit noise dominates photoelectron noise provided that the mean number of photoelectrons recorded per receiver resolution time is below 10 000.

We proceed now to examine the dependence of the SNR on photon flux Φ, receiver circuit-noise parameter σ_q, mean gain $\overline{G}$, and receiver bandwidth B. This will allow us to determine when the use of an avalanche photodiode is beneficial and will permit us to select an appropriate optical preamplifier for a given photon flux. In undertaking this parametric study, we rely on the expressions for the SNR provided in (19.6-37), (19.6-39), and (19.6-40).

Dependence of the SNR on Photon Flux

The dependence of the SNR on $\overline{m} = \eta\Phi/2B$ provides an indication of how the SNR varies with the photon flux Φ. Consider first a photodiode without gain, in which case (19.6-40) applies. Two limiting cases are of interest:

1. *Circuit-noise limit.* If Φ is sufficiently small, such that $\overline{m} \ll \sigma_q^2$ ($\Phi \ll 2B\sigma_q^2/\eta$), photon noise is negligible and circuit noise dominates, yielding

$$\text{SNR} \approx \frac{\overline{m}^2}{\sigma_q^2}. \tag{19.6-41}$$

2. *Photon-noise limit.* If the photon flux Φ is sufficiently large, such that $\overline{m} \gg \sigma_q^2$ ($\Phi \gg 2B\sigma_q^2/\eta$), the circuit-noise term can be neglected, whereupon

$$\text{SNR} \approx \overline{m}\,. \tag{19.6-42}$$

For small $\overline{m}$, therefore, the SNR is proportional to $\overline{m}^2$ and thereby to Φ^2, whereas for large $\overline{m}$ it is proportional to $\overline{m}$ and thereby to Φ, as illustrated in Fig. 19.6-11. For all levels of light, the SNR increases with increasing incident photon flux Φ; more light improves receiver performance.

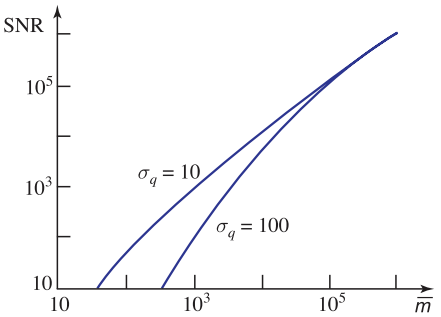


Figure 19.6-11 Signal-to-noise ratio (SNR) as a function of the mean number of photoelectrons per receiver resolution time, $\overline{m} = \eta\Phi/2B$, for a photodiode at two values of the circuit-noise parameter σ_q.

When is an APD Superior to a Photodiode?

We now compare two receivers that are identical in all respects except that one exhibits no gain, while the other exhibits mean gain $\overline{G}$ together with an excess noise factor F (e.g., an APD). For sufficiently small $\overline{m}$ (or photon flux Φ), circuit noise dominates. Amplifying the photocurrent above the level of circuit noise would then improve the SNR so that the APD receiver would be superior. For sufficiently large $\overline{m}$ (or photon flux), circuit noise is negligible. Amplifying the photocurrent then introduces gain noise, thereby reducing the SNR. The photodiode receiver would then be superior.

Comparing (19.6-39) and (19.6-40) reveals that the SNR of the APD receiver is greater than that of the photodiode receiver when $\overline{m} < \sigma_q^2(1 - 1/\overline{G}^2)/(F-1)$. For $\overline{G} \gg 1$, the APD provides an advantage when $\overline{m} < \sigma_q^2/(F-1)$. If this condition is not satisfied, the use of an APD compromises, rather than enhances, receiver performance. When σ_q is very small, for example, it is evident from (19.6-39) that the APD $\text{SNR} = \overline{m}/F$ is inferior to the photodiode $\text{SNR} = \overline{m}$. The SNR is plotted as a function of $\overline{m}$ for the two receivers in Fig. 19.6-12.

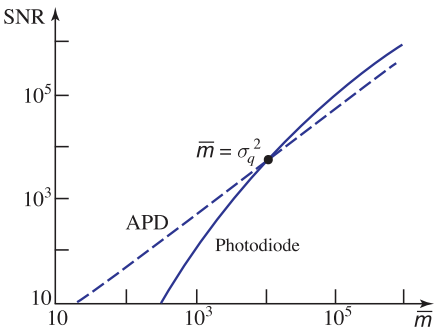


Figure 19.6-12 SNR versus $\overline{m} = \eta\Phi/2B$ for a photodiode receiver (solid curve) and for an APD receiver with mean gain $\overline{G} = 100$ and excess noise factor $F = 2$ (dashed curve) obtained from (19.6-39). The circuit-noise parameter $\sigma_q = 100$ in both cases. For small photon flux (circuit-noise-limited case), the APD yields a higher SNR than the photodiode. For large photon flux (photon-noise limited case), the photodiode receiver is superior to the APD receiver. The transition between the two regions occurs at $\overline{m} \approx \sigma_q^2/(F-1) = 10^4$.

Dependence of the SNR on APD Gain

As indicated above, the use of an APD with large gain is beneficial when the photon flux is sufficiently small, i.e., when $\overline{m} < \sigma_q^2/(F-1)$. The optimal gain of the APD is then determined by making use of (19.6-39):

$$\text{SNR} = \frac{\overline{G}^2\overline{m}}{\overline{G}^2 F + \sigma_q^2/\overline{m}}. \tag{19.6-43}$$

However, the excess noise factor F is itself a function of $\overline{G}$ for a conventional APD, as is clear from (19.6-27). Substituting (19.6-27) into (19.6-43) yields

$$\text{SNR} = \frac{\overline{G}^2\overline{m}}{\text{k}\overline{G}^3 + (1-\text{k})(2\overline{G}^2 - \overline{G}) + \sigma_q^2/\overline{m}}, \tag{19.6-44}$$

where k is the APD carrier ionization ratio specified in (19.4-1). Equation (19.6-44) is plotted in Fig. 19.6-13 for $\overline{m} = 1000$ and $\sigma_q = 500$, with k as a parameter. For the single-carrier-multiplication APD (k = 0), the SNR increases with gain and eventually saturates. For the double-carrier multiplication APD (k > 0), the SNR initially increases with increasing gain, but then reaches a maximum beyond which it decreases with increasing gain as a result of the sharp increase in gain noise. In general, therefore, there is an optimal value of the APD gain.

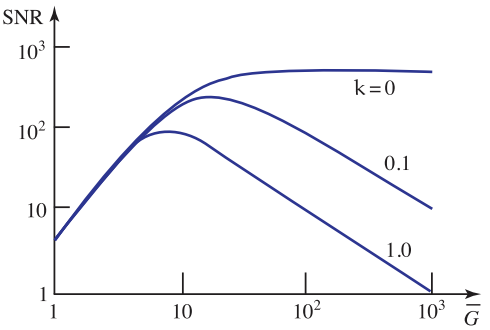


Figure 19.6-13 Dependence of the SNR on the APD mean gain $\overline{G}$ for different values of the ionization ratio k when $\overline{m} = 1000$ and $\sigma_q = 500$. Smaller values of k allow larger gain, higher receiver sensitivity, and larger values of the gain–bandwidth product.

Dependence of the SNR on Receiver Bandwidth

The relation between the SNR and the circuit bandwidth B is implicit in (19.6-37). It is governed by the dependence of the circuit-noise current variance σ_r^2 on B. Consider three receivers:

1. The *resistance-limited* receiver exhibits $\sigma_r^2 \propto B$ [see (19.6-33)] so that

$$\text{SNR} \propto 1/B. \tag{19.6-45}$$

2. The *FET amplifier* receiver obeys $\sigma_q \propto B^{1/2}$ [see (19.6-36)] so that $\sigma_r = 2eB\sigma_q \propto B^{3/2}$ [see (19.6-34)]. This indicates that the dependence of the SNR on B in (19.6-37) assumes the form

$$\text{SNR} \propto 1/(B + sB^3), \tag{19.6-46}$$

where s is a constant.

3. The *bipolar-transistor amplifier* receiver has a circuit-noise parameter σ_q that is approximately independent of B. Thus, $\sigma_r \propto B$ so that (19.6-37) takes the form

$$\text{SNR} \propto 1/(B + s'B^2), \tag{19.6-47}$$

where s' is a constant.

These relations are illustrated schematically in Fig. 19.6-14. The SNR always decreases with increasing B. For sufficiently small bandwidths, all three receivers exhibit an SNR that varies as $1/B$. For large bandwidths, the SNRs for the FET and bipolar-transistor receivers decline more sharply with bandwidth.

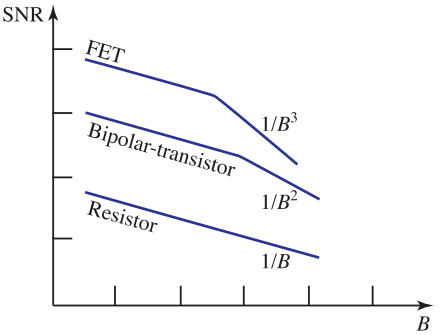


Figure 19.6-14 Double-logarithmic plot illustrating the dependence of the SNR on the circuit bandwidth B for the resistance-limited receiver, the bipolar-transistor receiver, and the FET receiver.

Analog Receiver Sensitivity

The receiver sensitivity is the minimum photon flux Φ_0, and its corresponding minimum optical power $P_0 = h\nu\Phi_0$ and minimum mean number of photoelectrons $\overline{m}_0 = \eta\Phi_0/2B$, required to achieve a prescribed value of signal-to-noise ratio SNR_0. The quantity $\overline{m}_0$ can be determined by solving (19.6-39) for $\text{SNR} = \text{SNR}_0$.

We consider only the unity-gain receiver, setting aside the more general solution as Exercise 19.6-2. Solving the quadratic equation (19.6-40) for $\overline{m}_0$, we obtain

$$\overline{m}_0 = \tfrac{1}{2}\left[\text{SNR}_0 + \sqrt{\text{SNR}_0^2 + 4\sigma_q^2\,\text{SNR}_0}\right]. \tag{19.6-48}$$

Two limiting cases emerge:

Photon-noise limit $\left(\sigma_q^2 \ll \tfrac{1}{4}\,\text{SNR}_0\right)$: $\overline{m}_0 = \text{SNR}_0$ (19.6-49)

Circuit-noise limit $\left(\sigma_q^2 \gg \tfrac{1}{4}\,\text{SNR}_0\right)$: $\overline{m}_0 = \sqrt{\text{SNR}_0}\,\sigma_q$. (19.6-50)

Receiver Sensitivity (Unity-Gain Receiver)

EXAMPLE 19.6-9. ***Sensitivity of an Analog Receiver.*** Assume that $\text{SNR}_0 = 10^4$, corresponding to an acceptable signal-to-noise ratio of 40 dB. If the receiver circuit-noise parameter $\sigma_q \ll 50$, the receiver is photon-noise limited and its sensitivity is $\overline{m}_0 = 10\,000$ photoelectrons per receiver resolution time. In the more likely situation for which $\sigma_q \gg 50$, the receiver sensitivity $\approx 100\,\sigma_q$. If $\sigma_q = 500$, for example, the sensitivity is $\overline{m}_0 = 50\,000$, which corresponds to $2B\overline{m}_0 = 10^5 B$ photoelectrons/s. The optical power sensitivity $P_0 = 2B\overline{m}_0 h\nu/\eta = 10^5 Bh\nu/\eta$ is directly proportional to the bandwidth. If $B = 100$ MHz and $\eta = 0.8$, then at $\lambda_o = 1550$ nm the receiver sensitivity is $P_0 \approx 1.6\ \mu$W.

When using (19.6-48) to determine the receiver sensitivity, it is important to keep in mind that the circuit-noise parameter σ_q is, in general, a function of the bandwidth B, in accordance with:

Resistance-limited receiver:	$\sigma_q \propto 1/\sqrt{B}$
FET amplifier:	$\sigma_q \propto \sqrt{B}$
Bipolar-transistor amplifier:	σ_q independent of B.

For these receivers, the receiver sensitivity $\overline{m}_0$ therefore depends on the receiver bandwidth B as illustrated in Fig. 19.6-15. The optimal choice of receiver therefore depends in part on the bandwidth B.

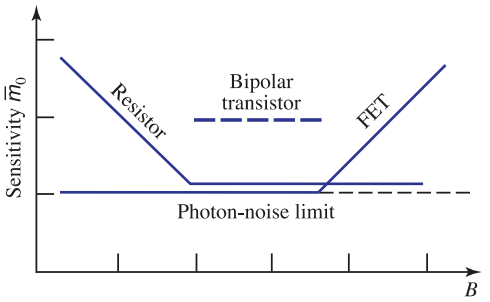


Figure 19.6-15 Double-logarithmic plot of receiver sensitivity $\overline{m}_0$ (minimum mean number of photoelectrons per receiver resolution time $T = 1/2B$ that guarantees a minimum signal-to-noise ratio SNR_0) as a function of the receiver bandwidth B, for three receivers. The curves approach the photon-noise limit at values of B for which $\sigma_q^2 \ll {}^1\!/_4\,\text{SNR}_0$. In the photon-noise limit (when circuit noise is negligible), the sensitivity $\overline{m}_0$ is equal to SNR_0 in all cases.

EXERCISE 19.6-2

Sensitivity of an Analog APD Receiver. Derive an expression analogous to (19.6-48) for the sensitivity of a receiver that incorporates an APD with mean gain $\overline{G}$ and excess noise factor F. Show that in the limit of negligible circuit noise, the receiver sensitivity reduces to

$$\overline{m}_0 = F \cdot \text{SNR}_0. \tag{19.6-51}$$

E. Bit Error Rate and Digital Receiver Sensitivity

The sensitivity of an analog receiver was defined in Sec. 19.6D as the minimum power of the received light (or the corresponding mean number of photons or photoelectrons) required to achieve a prescribed signal-to-noise ratio SNR_0. We now turn to the direct-detection digital communications receiver. In this case, the receiver sensitivity is defined as the minimum optical energy (or corresponding mean number of photons) per bit necessary to achieve a prescribed bit error rate (BER). The calculations are carried out in the context of an ON–OFF keying (OOK) system: the logic states "1" and "0" of a bit represent, respectively, the presence and absence of an optical pulse. We first determine the receiver sensitivity under ideal conditions, when the photodetector has unity quantum efficiency and only photon noise is present. We then consider the increase in sensitivity (decrease in performance) that results from incorporating photoelectron noise, background noise, photodetector gain noise, and circuit noise into the system. The performance of direct-detection and coherent-detection optical fiber communication systems are considered in detail in Secs. 25.2 and 25.4, respectively.

Sensitivity of the Ideal Digital Optical Receiver

Assume that the logic states "1" and "0" in an OOK system correspond to the presence and absence of optical energy, respectively; in state "1" an average of $\bar{n}$ photons is received while in state "0" no photons are received. If the two states are equally likely, the overall mean number of photons per bit is $\bar{n}_0 = \frac{1}{2}\bar{n}$. Since the actual number of detected photons is random, errors in logic-state identification occur, as portrayed in Fig. 19.6-16(a).

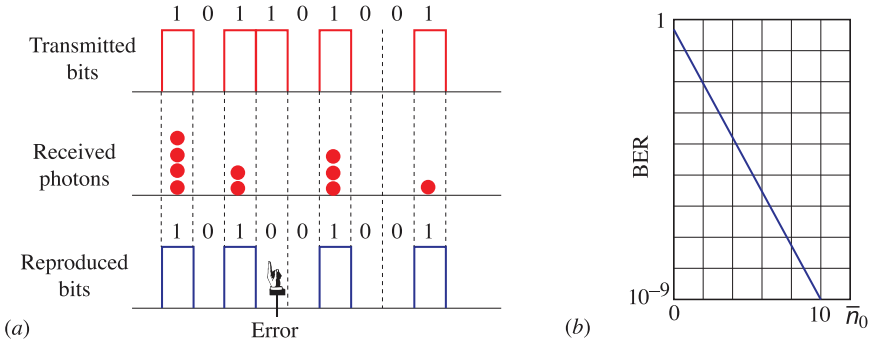


Figure 19.6-16 (a) Schematic illustrating errors that result from randomness in the photon number. (b) Bit error rate (BER) versus mean number of photons per bit $\bar{n}_0$ for an ideal, direct-detection OOK system in which the only source of noise is Poisson photon fluctuations.

For light generated by most optical sources, including laser diodes, lasers, and light-emitting diodes, the probability of finding n photons in a fixed time interval T obeys the Poisson distribution, $p(n) = \bar{n}^n \exp(-\bar{n})/n!$, where $\bar{n}$ is the mean number of photons (Sec. 13.2C). The ideal direct-detection OOK receiver is designed in such a way that it decides that "1" has been transmitted if it detects one or more photons. The probability p_1 of mistaking "1" for "0" is therefore given by the Poisson probability of detecting zero photons: $p_1 = p(0) = \exp(-\bar{n})$. When state "0" is transmitted, zero photons are detected; the receiver then correctly decides that state "0" has been transmitted and the error probability $p_0 = 0$. The bit error rate is the average of the two error probabilities,

BER $= \frac{1}{2}(p_1 + p_0)$, from which we find

$$\text{BER} = \tfrac{1}{2}\exp(-\overline{n}) = \tfrac{1}{2}\exp(-2\overline{n}_0). \tag{19.6-52}$$

Figure 19.6-16(b) portrays a semilogarithmic plot of this relation.

The receiver sensitivity is defined as the average number of photons per bit required to achieve a specified value of the BER. In particular, for a BER $= 10^{-9}$, an oft-chosen metric, (19.6-52) yields $\overline{n}_0 \approx 10$ photons per bit. We conclude that:

The receiver sensitivity for a direct-detection, binary OOK communication system that is ideal in every respect except for the Poisson statistics of the detected photon number is 10 *photons per bit at a BER of* 10^{-9}.

The ideal receiver sensitivity can be improved, in principle, by making use of photon-number-squeezed (sub-Poisson) light (Sec. 13.3C).

EXERCISE 19.6-3

Effect of Quantum Efficiency and Background Noise on Receiver Sensitivity.

(a) Show that for a receiver using a photodetector of quantum efficiency η, but that is otherwise ideal, BER $= \frac{1}{2}\exp(-2\eta\overline{n}_0)$, so that the receiver sensitivity is $\overline{n}_0 = 10/\eta$ photons per bit at a BER $= 10^{-9}$, corresponding to $\overline{m}_0 = \eta\overline{n}_0 = 10$ photoelectrons per bit.

(b) Assume that states "1" and "0" correspond, respectively, to mean photon numbers $\overline{n}+\overline{n}_B$ and $\overline{n}_B$, where $\overline{n}$ is the mean number of signal photons and $\overline{n}_B$ is the mean number of detected Poisson-distributed background photons that is independent of the signal. Determine an expression for the BER as a function of $\overline{n}$ and $\overline{n}_B$. Plot the BER versus $\overline{n}_0 = \frac{1}{2}\overline{n}$ for several values of $\overline{n}_B$. Determine the receiver sensitivity $\overline{n}_0$ as a function of $\overline{n}_B$ from this plot. [*Hint*: The sum of two Poisson-distributed random variables is also Poisson-distributed.]

Sensitivity of a Digital Receiver with Circuit Noise and Gain Noise

As elucidated in Sec. 19.6A, a photodetector transforms a fraction η of the incident photons into charge carriers, each of which then contributes a charge e to the electric current in the external circuit. The total charge accumulated in the bit time interval T is m (units of electrons). If the incident photons obey a Poisson distribution with mean $\overline{n}$, the photoelectrons also obey a Poisson distribution with mean $\overline{m} = \eta\overline{n}$ and variance $\overline{m}$.

Additional noise is introduced into the receiver circuit via a random electric current i_r with zero mean and a probability distribution that is approximately Gaussian with variance σ_r^2. Within the bit time interval T, the accumulated charge $q = i_r T/e$ (units of electrons) has an RMS value $\sigma_q = \sigma_r T/e$. The circuit-noise parameter σ_q depends on the receiver bandwidth B, as represented in (19.6-34).

The total accumulated charge per bit $s = m + q$ (units of electrons) is thus the sum of a Poisson random variable m and an independent Gaussian random variable q. Its mean μ is the sum of the means,

$$\mu = \overline{m} = \eta\overline{n}\,, \tag{19.6-53}$$

and its variance σ^2 is the sum of the variances,

$$\sigma^2 = \overline{m} + \sigma_q^2. \tag{19.6-54}$$

For $\overline{m}$ sufficiently large, the Poisson distribution may itself be approximated by a Gaussian distribution so that the overall distribution can be cast in the form of a Gaussian distribution of mean μ and variance σ^2. We adopt this approximation in the following analysis.

For an avalanche photodiode (APD) of mean gain $\overline{G}$, the mean number of photoelectrons is multiplied by the factor $\overline{G}$; gain noise is also introduced by the multiplication process. The mean of the total collected charge per bit s (units of electrons) is then

$$\mu = \overline{m}\,\overline{G} \tag{19.6-55}$$

while the variance is

$$\sigma^2 = \overline{m}\,\overline{G}^2 F + \sigma_q^2\,, \tag{19.6-56}$$

where $F = \langle G^2\rangle/\langle G\rangle^2$ is the APD excess noise factor [see (19.6-24)].

The direct-detection OOK receiver determines the charge s accumulated in each bit (by use of an integrator, for example) and compares it with a prescribed threshold ϑ. If $s > \vartheta$, state "1" is selected; otherwise, state "0" is selected. The probabilities of error, p_1 and p_0, are determined from two Gaussian probability distributions in s with the following parameters:

$$\begin{array}{lll} \text{mean } \mu_0 = 0, & \text{variance } \sigma_0^2 = \sigma_q^2 & \text{for state "0"} \\ \text{mean } \mu_1 = \overline{m}\,\overline{G}, & \text{variance } \sigma_1^2 = \overline{m}\,\overline{G}^2 F + \sigma_q^2 & \text{for state "1".} \end{array} \tag{19.6-57}$$

The probability p_0 of mistaking "0" for "1" is the integral of a Gaussian probability distribution $p(s)$ with mean μ_0 and variance σ_0^2, from $s = \vartheta$ to $s = \infty$. The probability p_1 of mistaking "1" for "0" is the integral of a Gaussian probability distribution with mean μ_1 and variance σ_1^2, from $s = -\infty$ to $s = \vartheta$. The threshold ϑ is selected such that the average probability of error, BER $= \frac{1}{2}(p_0 + p_1)$, is minimized.

An analysis along these lines is the basis of the conventional theory of binary detection in the presence of Gaussian noise, which is widely applicable (indeed we shall make use of it in our analysis of coherent communication systems in Sec. 25.4). If μ_0 and σ_0^2, and μ_1 and σ_1^2, are the means and variances associated with two Gaussian variables representing states "0" and "1", respectively, and if σ_0 and σ_1 are much smaller than $\mu_1 - \mu_0$, it can be shown that the bit error rate for an optimal-threshold receiver is given by

$$\text{BER} \approx \tfrac{1}{2}[1 - \text{erf}(Q/\sqrt{2})]. \tag{19.6-58}$$

Here

$$Q = \frac{\mu_1 - \mu_0}{\sigma_1 + \sigma_0} \tag{19.6-59}$$

and the error function $\text{erf}(z)$ is defined as

$$\text{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z \exp(-x^2)\,dx. \tag{19.6-60}$$

From (19.6-58) we know that a BER of 10^{-9} corresponds to $Q \approx 6$, whereupon (19.6-59) provides

$$\boxed{\mu_1 - \mu_0 \approx 6(\sigma_1 + \sigma_0).} \qquad \text{(19.6-61)}$$

Condition for BER $= 10^{-9}$ (Gaussian Approximation)

Substituting (19.6-57) into (19.6-61), defining $\overline{m}_0 = \frac{1}{2}\overline{m}$ as the mean number of photoelectrons detected per bit, and working through some algebra yields

$$\overline{m}_0 \approx 18F + 6\sigma_q/\overline{G}. \qquad \text{(19.6-62)}$$

Equation (19.6-62) relates the receiver sensitivity, specified by the mean number of photoelectrons per bit $\overline{m}_0$ required to render the BER $= 10^{-9}$, to the APD and circuit-noise parameters $\overline{G}$, F, and σ_q. The approximation serves well for the parameter values observed in actual systems.

When the APD gain is sufficiently large such that $3\overline{G}F \gg \sigma_q$, the second (circuit-noise-dependent) term on the right-hand side of (19.6-62) can be neglected, whereupon

$$\boxed{\overline{m}_0 \approx 18F.} \qquad \text{(19.6-63)}$$

APD Receiver Sensitivity (Absence of Circuit Noise)

According to (19.6-63), a receiver that has negligible circuit noise, and makes use of a photodiode with no gain ($\overline{G} = 1$ and $F = 1$), exhibits a receiver sensitivity $\overline{m}_0 \approx 18$ photoelectrons per bit at a BER $= 10^{-9}$. This result differs from the 10-photoelectrons-per-bit sensitivity established earlier for this ideal receiver. The discrepancy arises because of the replacement of the Poisson distribution by a Gaussian distribution, which is inaccurate for small photon numbers.

Typical sensitivities for several direct-detection, OOK receivers are provided in Table 19.6-1. It is of interest to compare these results with those presented in Table 25.4-1 for coherent-detection systems.

Table 19.6-1 Typical values of the sensitivity (number of photons per bit) for several direct-detection, OOK optical receivers with amplifier and circuit noise when operated at a BER $= 10^{-9}$, assuming that the photodetector quantum efficiency $\eta = 1$. The actual values depend on the amplifier parameters $\overline{G}$ and F as well as on the receiver circuit-noise parameter σ_q, which in turn depends on the bit rate $B_0 = 1/T$.

Receiver	Receiver Sensitivity (photons/bit)
Photon-limited ideal detector	10
Si APD	125
Er^{3+}-doped silica-fiber preamplifier/InGaAs p–i–n photodiode	215
InGaAs APD	500
p–i–n photodiode	6000

READING LIST

Photodetectors

See also the reading lists in Chapters 17 and 18.

B. Nabet, ed., *Photodetectors: Materials, Devices and Applications*, Elsevier-Woodhead, 2016.

S. N. Ahmed, *Physics and Engineering of Radiation Detection*, Elsevier, 2nd ed. 2015.

A. Beling and J. C. Campbell, Advances in Photodetectors and Optical Receivers, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

D. N. Bose, Photodetectors, in S. S. Jha, ed., *Perspectives in Optoelectronics*, World Scientific, 1995, Chapter 6, pp. 299–384.

Photoemission, Photomultiplier Tubes, and Microchannel Plates

A. G. Wright, *The Photomultiplier Handbook*, Oxford University Press, 2017.

C. Warner, Photomultiplier Tubes Detect Neutrinos Changing Identities, *SPIE Professional Magazine*, vol. 11, no. 2, pp. 20–23, 2016.

R. Locher, L. Castiglioni, M. Lucchini, M. Greif, L. Gallmann, J. Osterwalder, M. Hengsberger, and U. Keller, Energy-Dependent Photoemission Delays from Noble Metal Surfaces by Attosecond Interferometry, *Optica*, vol. 2, pp. 405–410, 2015.

K. Watase, Photomultiplier Tubes: μPMT is Key to High-Performance Portable Devices, *Laser Focus World*, vol. 49, no. 5, pp. 60–62, 2013.

T. Hakamaka *et al.*, eds., *Photomultiplier Tubes: Basics and Applications*, Hamamatsu Photonics, 3rd ed. 2007.

M. Lampton, The Microchannel Image Intensifier, *Scientific American*, vol. 245, no. 5, pp. 62–71, 1981.

R. W. Engstrom, *RCA Photomultiplier Handbook* (*PMT-62*), RCA Electro Optics and Devices (Lancaster, PA), 1980.

W. Shockley and J. R. Pierce, A Theory of Noise for Electron Multipliers, *Proceedings of the IRE*, vol. 26, pp. 321–332, 1938.

V. K. Zworykin, G. A. Morton, and L. Malter, The Secondary Emission Multiplier—A New Electronic Device, *Proceedings of the IRE*, vol. 24, pp. 351–375, 1936.

H. Iams and B. Salzberg, The Secondary Emission Phototube, *Proceedings of the IRE*, vol. 23, pp. 55–64, 1935.

A. Einstein, Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt, *Annalen der Physik*, vol. 322, pp. 132–148, 1905 [Translation: Concerning an Heuristic Point of View Toward the Emission and Transformation of Light, *American Journal of Physics*, vol. 33, pp. 367–374, 1965].

H. Hertz, Ueber einen Einfluss des ultravioletten Lichtes auf die elektrische Entladung [Concerning an Influence of Ultraviolet Light on the Electrical Discharge], *Annalen der Physik*, vol. 267, pp. 983–1000, 1887.

Photoconductivity

N. V. Joshi, ed., *Selected Papers on Photoconductivity*, SPIE Optical Engineering Press (Milestone Series Volume 56), 1992.

N. V. Joshi, *Photoconductivity: Art, Science, and Technology*, CRC Press, 1990.

A. Rose, *Concepts in Photoconductivity and Allied Problems*, Wiley, 1963; Krieger, reissued 1978.

Group-IV, Graphene, Plasmonic, and Organic Photodetectors

D. Thomson, A. Zilkie, J. E. Bowers, T. Komljenovic, G. T. Reed, L. Vivien, D. Marris-Morini, E. Cassan, L. Virot, J.-M. Fédéli, J.-M. Hartmann, J. H. Schmid, D.-X. Xu, F. Boeuf, P. O'Brien, G. Z. Mashanovich, and M. Nedeljkovic, Roadmap on Silicon Photonics, *Journal of Optics*, vol. 18, 073003, 2016.

T. Pham, W. Du, H. Tran, J. Margetis, J. Tolle, G. Sun, R. A. Soref, H. A. Naseem, B. Li, and S.-Q. Yu, Systematic Study of Si-Based GeSn Photodiodes with 2.6 μm Detector Cutoff for Short-Wave Infrared Detection, *Optics Express*, vol. 24, pp. 4519–4531, 2016.

A. Di Bartolomeo, Graphene Schottky Diodes: An Experimental Review of the Rectifying Graphene/Semiconductor Heterojunction, *Physics Reports*, vol. 606, pp. 1–58, 2016.

D. Natali and M. Caironi, Organic Photodetectors, in B. Nabet, ed., *Photodetectors: Materials, Devices and Applications*, Elsevier-Woodhead, 2016.

M. Kielar, O. Dhez, G. Pecastaings, A. Curutchet, and L. Hirsch, Long-Term Stable Organic Photodetectors with Ultra Low Dark Currents for High Detectivity Applications, *Scientific Reports*, vol. 6, 39201, 2016.

K.-J. Baeg, M. Binda, D. Natali, M. Caironi, and Y.-Y. Noh, Organic Light Detectors: Photodiodes and Phototransistors, *Advanced Materials*, vol. 25, pp. 4267–4295, 2013.

X. An, F. Liu, Y. J. Jung, and S. Kar, Tunable Graphene–Silicon Heterojunctions for Ultrasensitive Photodetection, *Nano Letters*, vol. 13, pp. 909–916, 2013.

A. V. Zayats and S. A. Maier, eds., *Active Plasmonics and Tuneable Plasmonic Metamaterials*, Wiley–Science Wise, 2013.

L. Vivien, A. Polzer, D. Marris-Morini, J. Osmond, J. M. Hartmann, P. Crozat, E. Cassan, C. Kopp, H. Zimmermann, and J. M. Fédéli, Zero-Bias 40 Gb/s Germanium Waveguide Photodetector on Silicon, *Optics Express*, vol. 20, pp. 1096–1101, 2012.

H. A. Atwater and A. Polman, Plasmonics for Improved Photovoltaic Devices, *Nature Materials*, vol. 9, pp. 205–213, 2010.

H. R. Stuart and D. G. Hall, Absorption Enhancement in Silicon-on-Insulator Waveguides Using Metal Island Films, *Applied Physics Letters*, vol. 69, pp. 2327–2329, 1996.

Infrared Photodetectors

Z. Jakšić, *Micro and Nanophotonics for Semiconductor Infrared Detectors: Towards an Ultimate Uncooled Device*, Springer-Verlag, 2014.

M. A. Kinch, *State-of-the-Art Infrared Detector Technology*, SPIE Optical Engineering Press, 2014.

A. Rogalski, *Infrared Detectors*, CRC Press/Taylor & Francis, 2nd ed. 2011.

C. Jagadish, ed., *Semiconductors and Semimetals*, S. Gunapala and D. Rhiger, eds., Volume 84, *Advances in Infrared Photodetectors*, Academic Press/Elsevier, 2011.

A. Daniels, *Field Guide to Infrared Systems, Detectors, and FPAs*, SPIE Optical Engineering Press, 2nd ed. 2010.

H. Schneider and H. C. Liu, *Quantum Well Infrared Photodetectors: Physics and Applications*, Springer-Verlag, 2007.

A. Rogalski, ed., *Selected Papers on Infrared Detectors: Developments*, SPIE Optical Engineering Press (Milestone Series Volume 179), 2004.

E. L. Dereniak and G. D. Boreman, *Infrared Detectors and Systems*, Wiley, 1996.

A. Rogalski, ed., *Selected Papers on Semiconductor Infrared Detectors*, SPIE Optical Engineering Press (Milestone Series Volume 66), 1992.

N. Sclar, Properties of Doped Silicon and Germanium Infrared Detectors, *Progress in Quantum Electronics*, vol. 9, pp. 149–257, 1984.

R. J. Keyes, ed., *Optical and Infrared Detectors*, Volume 19, *Topics in Applied Physics*, Springer-Verlag, 2nd ed. 1980.

R. K. Willardson and A. C. Beer, eds., *Semiconductors and Semimetals*, *Infrared Detectors II*, Academic Press, vol. 12, 1977.

R. D. Hudson, Jr. and J. W. Hudson, eds., *Benchmark Papers in Optics / 2: Infrared Detectors*, Dowden, Hutchinson & Ross, 1975.

R. K. Willardson and A. C. Beer, eds., *Semiconductors and Semimetals*, *Infrared Detectors*, Academic Press, vol. 5, 1970.

Avalanche Photodiodes and Gain Noise

S. R. Bank, J. C. Campbell, S. J. Maddox, M. Ren, A.-K. Rockwell, M. E. Woodson, and S. D. March, Avalanche Photodiodes Based on the AlInAsSb Materials System, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, 3800407, 2018.

J. C. Campbell, Recent Advances in Avalanche Photodiodes, *Journal of Lightwave Technology*, vol. 34, pp. 278–285, 2016.

D. Dai, M. Piels, and J. E. Bowers, Monolithic Germanium/Silicon Photodetectors with Decoupled Structures: Resonant APDs and UTC Photodiodes, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, 3802214, 2014.

A. Singh, V. Srivastav, and R. Pal, HgCdTe Avalanche Photodiodes: A Review, *Optics & Laser Technology*, vol. 43, pp. 1358–1370, 2011.

J. Beck, C. Wan, M. Kinch, J. Robinson, P. Mitra, R. Scritchfield, F. Ma, and J. Campbell, The HgCdTe Electron Avalanche Photodiode, *Journal of Electronic Materials*, vol. 35, pp. 1166–1173, 2006.

M. M. Hayat, O.-H. Kwon, S. Wang, J. C. Campbell, B. E. A. Saleh, and M. C. Teich, Boundary Effects on Multiplication Noise in Thin Heterostructure Avalanche Photodiodes: Theory and Experiment, *IEEE Transactions on Electron Devices*, vol. 49, pp. 2114–2123, 2002.

P. Yuan, S. Wang, X. Sun, X. G. Zheng, A. L. Holmes, Jr., and J. C. Campbell, Avalanche Photodiodes with an Impact-Ionization-Engineered Multiplication Region, *IEEE Photonics Technology Letters*, vol. 12, pp. 1370–1372, 2000.

M. M. Hayat, B. E. A. Saleh, and M. C. Teich, Effect of Dead Space on Gain and Noise of Double-Carrier-Multiplication Avalanche Photodiodes, *IEEE Transactions on Electron Devices*, vol. 39, pp. 546–552, 1992.

N. Z. Hakim, B. E. A. Saleh, and M. C. Teich, Generalized Excess Noise Factor for Avalanche Photodiodes of Arbitrary Structure, *IEEE Transactions on Electron Devices*, vol. 37, pp. 599–610, 1990.

M. C. Teich, K. Matsuo, and B. E. A. Saleh, Excess Noise Factors for Conventional and Superlattice Avalanche Photodiodes and Photomultiplier Tubes, *IEEE Journal of Quantum Electronics*, vol. QE-22, pp. 1184–1193, 1986.

R. Chin, N. Holonyak, Jr., G. E. Stillman, J. Y. Tang, and K. Hess, Impact Ionization in Multilayered Heterojunction Structures, *Electronics Letters*, vol. 16, pp. 467–469, 1980.

K. Nishida, K. Taguchi, and Y. Matsumoto, InGaAsP Heterostructure Avalanche Photodiodes with High Avalanche Gain, *Applied Physics Letters*, vol. 35, pp. 251–253, 1979.

H. W. Ruegg, An Optimized Avalanche Photodiode, *IEEE Transactions on Electron Devices*, vol. ED-14, pp. 239–251, 1967.

R. J. McIntyre, Multiplication Noise in Uniform Avalanche Diodes, *IEEE Transactions on Electron Devices*, vol. ED-13, pp. 164–168, 1966.

K. M. Johnson, High-Speed Photodiode Signal Enhancement at Avalanche Breakdown Voltage, *IEEE Transactions on Electron Devices*, vol. ED-12, pp. 55–63, 1965.

SPADs, SiPMs, and SSPDs

S. Piatek and E. Hergert, Detector Options for Low-Light Applications, *Photonics Spectra*, vol. 51, no. 10, pp. 54–58, 2017.

D. Durini, U. Paschen, A. Schwinger, and A. Spickermann, Silicon Based Single-Photon Avalanche Diode (SPAD) Technology for Low-Light and High-Speed Applications, in B. Nabet, ed., *Photodetectors: Materials, Devices and Applications*, Elsevier-Woodhead, 2016.

N. Dinu, Silicon Photomultipliers (SiPM), in B. Nabet, ed., *Photodetectors: Materials, Devices and Applications*, Elsevier-Woodhead, 2016.

T. Gerrits, A. Lita, B. Calkins, and S. W. Nam, Superconducting Transition Edge Sensors for Quantum Optics, in R. H. Hadfield and G. Johansson, eds., *Superconducting Devices in Quantum Optics*, Springer, 2016.

M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov, Invited Review Article: Single-Photon Sources and Detectors, *Review of Scientific Instruments*, vol. 82, 071101, 2011.

R. H. Hadfield, Single-Photon Detectors for Optical Quantum Information Applications, *Nature Photonics*, vol. 3, pp. 696–705, 2009.

A. E. Lita, A. J. Miller, and S. W. Nam, Counting Near-Infrared Single-Photons with 95% Efficiency, *Optics Express*, vol. 16, pp. 3032–3040, 2008.

D. A. Ramirez, M. M. Hayat, G. Karve, J. C. Campbell, S. N. Torres, B. E. A. Saleh, and M. C. Teich, Detection Efficiencies and Generalized Breakdown Probabilities for Nanosecond-Gated Near Infrared Single-Photon Avalanche Photodiodes, *IEEE Journal of Quantum Electronics*, vol. 42, pp. 137–145, 2006.

G. N. Gol'tsman, O. Okunev, G. Chulkova, A. Lipatov, A. Semenov, K. Smirnov, B. Voronov, A. Dzardanov, C. Williams, and R. Sobolewski, Picosecond Superconducting Single-Photon Optical Detector, *Applied Physics Letters*, vol. 79, pp. 705–707, 2001.

A. Smith, ed., *Selected Papers on Photon-Counting Detectors*, SPIE Optical Engineering Press (Milestone Series Volume 143), 1998.

D. H. Andrews, W. F. Brucksch, Jr., W. T. Ziegler, and E. R. Blanchard, Attenuated Superconductors I. For Measuring Infra-Red Radiation, *Review of Scientific Instruments*, vol. 13, pp. 281–292, 1942.

Array and Imaging Detectors

X. Sun, J. B. Abshire, J. D. Beck, P. Mitra, K. Reiff, and G. Yang, HgCdTe Avalanche Photodiode Detectors for Airborne and Spaceborne Lidar at Infrared Wavelengths, *Optics Express*, vol. 25, pp. 16589–16602, 2017.

B. Aull, Geiger-Mode Avalanche Photodiode Arrays Integrated to All-Digital CMOS Circuits, *Sensors*, vol. 16, 495, 2016.

T. Kuroda, *Essential Principles of Image Sensors*, CRC Press/Taylor & Francis, 2014.

D. Durini, ed., *High Performance Silicon Imaging: Fundamentals and Applications of CMOS and CCD Sensors*, Elsevier-Woodhead, 2014.

G. C. Holst and T. S. Lomheim, *CMOS/CCD Sensors and Camera Systems*, SPIE Optical Engineering Press, 2nd ed. 2011.

W. S. Boyle, Nobel Lecture: CCD — An Extension of Man's View, *Reviews of Modern Physics*, vol. 82, pp. 2305–2306, 2010.

G. E. Smith, Nobel Lecture: The Invention and Early History of the CCD, *Reviews of Modern Physics*, vol. 82, pp. 2307–2312, 2010.

A. Rogalski, J. Antoszewski, and L. Faraone, Third-Generation Infrared Photodetector Arrays, *Journal of Applied Physics*, vol. 105, 091101, 2009.

M. G. Kang, ed., *Selected Papers on CCD and CMOS Imagers*, SPIE Optical Engineering Press (Milestone Series Volume 177), 2003.

Photon, Photoelectron, and Circuit Noise in Photodetectors

E. Säckinger, *Analysis and Design of Transimpedance Amplifiers for Optical Receivers*, Wiley, 2018.

S. B. Lowen and M. C. Teich, *Fractal-Based Point Processes*, Wiley, 2005.

H. A. Haus, *Electromagnetic Noise and Quantum Optical Measurements*, Springer-Verlag, 2000.

M. C. Teich and B. E. A. Saleh, Branching Processes in Quantum Electronics, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 6, pp. 1450–1457, 2000.

A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw–Hill, 3rd ed. 1991.

A. van der Ziel, *Noise in Solid State Devices and Circuits*, Wiley, 1986.

M. J. Buckingham, *Noise in Electron Devices and Systems*, Wiley, 1983.

D. R. Cox and V. Isham, *Point Processes*, Chapman & Hall, 1980.

B. E. A. Saleh, *Photoelectron Statistics*, Springer-Verlag, 1978.

J. B. Johnson, The Schottky Effect in Low Frequency Circuits, *Physical Review*, vol. 26, pp. 71–85, 1925.

W. Schottky, Über spontane Stromschwankungen in verschiedenen Elektrizitätsleitern [Concerning Spontaneous Current Fluctuations in Various Electrical Conductors], *Annalen der Physik*, vol. 57, pp. 541–567, 1918.

S. D. Poisson, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités* [*Research on the Probability of Judgments in Criminal and Civil Matters, Preceded by the General Rules for the Calculation of Probabilities*], Bachelier, 1837.

PROBLEMS

19.1-1 **Effect of Reflectance on Quantum Efficiency.** Calculate the factor $\mathcal{T} = 1 - \mathcal{R}$ in the expression for the photoelectric-detector quantum efficiency provided in (19.1-3), under both normal and 45° incidence, for an unpolarized light beam incident from air onto Si, GaAs, and InSb (refer to Sec. 6.2 and Table 17.2-1).

19.1-2 **Maximum Responsivity.** Determine the maximum responsivity of an ideal (unity quantum efficiency and unity gain) semiconductor photodetector fabricated from (a) Si; (b) GaAs; and (c) InSb.

19.1-3 **Transit Time Current and Duration for a Single Carrier Pair.** Referring to Fig. 19.1-6, assume that a photon generates an electron–hole pair at the position $x = w/3$, that $v_e = 3v_h$, and that the carriers recombine at the contacts. For each carrier in the semiconductor, find the magnitudes of the currents, i_h and i_e, and the durations of the currents, τ_h and τ_e. Express your results in terms of e, w, and v_e. Verify that the total charge induced in the circuit is e. For $v_e = 6 \times 10^7$ cm/s and $w = 10$ μm, sketch the time course of $i_h(t)$, $i_e(t)$, and $i(t)$.

*19.1-4 **Transit-Time Spread for a Uniformly Illuminated Semiconductor Photodetector.** Consider a semiconductor material (as in Fig. 19.1-6) exposed to a broad-area impulse of light at $t = 0$ that uniformly generates N electron–hole pairs between 0 and w. Let the electron and hole velocities in the material be v_e and v_h, respectively. Show that the hole current can be written as

$$i_h(t) = \begin{cases} -\dfrac{Nev_h^2}{w^2}\, t + \dfrac{Nev_h}{w}, & 0 \le t \le \dfrac{w}{v_h} \\ 0, & \text{elsewhere,} \end{cases}$$

and that the electron current can be written as

$$i_e(t) = \begin{cases} -\dfrac{Nev_e^2}{w^2}\, t + \dfrac{Nev_e}{w}, & 0 \le t \le \dfrac{w}{v_e} \\ 0, & \text{elsewhere,} \end{cases}$$

so that the total current is

$$i(t) = \begin{cases} \dfrac{Ne}{w}\left[(v_h + v_e) - \dfrac{1}{w}(v_h^2 + v_e^2)t\right], & 0 \le t \le \dfrac{w}{v_e} \\ \dfrac{Nev_h}{w}\left[1 - \dfrac{v_h}{w}t\right], & \dfrac{w}{v_e} \le t \le \dfrac{w}{v_h}. \end{cases}$$

These three currents are illustrated in Fig. 19.1-7. Verify that the electrons and holes each contribute charge $Ne/2$ to the external circuit so that the total charge delivered to the circuit is Ne.

*19.1-5 **Two-Photon Photodetectors.** Consider a beam of photons of energy $h\nu$ and photon-flux density ϕ (photons/cm^2-s) incident on a semiconductor detector with bandgap energy $h\nu < E_g < 2h\nu$, so that one photon has insufficient energy to raise an electron from the valence to the conduction band. However, two photons can occasionally conspire to jointly surrender their energy to the electron. Assume that the *current density* induced in such a detector is $J_p = \zeta\phi^2$, where ζ is a constant. Show that the responsivity (A/W) of the two-photon photodetector is given by $\mathcal{R} = [\zeta/(hc_o)^2]\lambda_o^2 P/A$, where P is the optical power and A is the illuminated detector area. Provide a rationale for the proportionality of $\mathcal{R}$ to λ_o^2 and P/A. Two-photon photoemission behaves similarly.[†]

19.2-1 **Photoconductive Detector Circuit.** A photoconductive detector is often connected in series with a load resistor R and a DC voltage source V, and the voltage V_p across the load resistor is measured. If the conductance of the detector is proportional to the optical power P, sketch the dependence of V_p on P. Under what conditions is this dependence linear?

19.2-2 **Photoconductivity in Intrinsic Si.** The concentration of charge carriers in a sample of intrinsic Si is $\mathrm{n_i} = 1.5 \times 10^{10}$ cm^{-3} and the recombination lifetime $\tau = 10$ μs. If the material is illuminated with light, and an optical power density of 1 mW/cm^3 at $\lambda_o = 1$ μm is absorbed by the material, determine the percentage increase in its conductivity. Assume that the quantum efficiency $\eta = {}^1\!/_2$.

[†] See M. C. Teich, J. M. Schröer, and G. J. Wolga, Double-Quantum Photoelectric Emission from Sodium Metal, *Physical Review Letters*, vol. 13, pp. 611–614, 1964.

19.3-1 **Quantum Efficiency and Responsivity of a Photodiode Detector.** A particular p–i–n photodiode illuminated by a pulse of light containing 6×10^{12} incident photons at a wavelength of $\lambda_o = 1550$ nm gives rise, on average, to 2×10^{12} electrons collected at the terminals of the device. Determine the quantum efficiency η and the responsivity R of the photodiode at this wavelength.

19.4-1 **Quantum Efficiency of an APD.** An APD with gain $\overline{G} = 20$ operates at a wavelength $\lambda_o = 1550$ nm. If its responsivity at this wavelength is $\mathsf{R} = 12$ A/W, calculate its quantum efficiency η. What is the photocurrent i_p at the output of the device if a photon flux $\Phi = 10^{10}$ photons/s, at this same wavelength, is incident on it?

19.4-2 **Gain of a Ge APD.** Show that a conventional APD with ionization ratio $\mathsf{k} \approx 1$, such as a device fabricated from germanium, has a gain given by $\overline{G} = 1/(1 - \alpha_e w)$, where α_e is the electron ionization coefficient and w is the width of the multiplication layer. Note that (19.4-8) cannot be used to provide a proper result for the gain when $\mathsf{k} = 1$.

19.6-1 **Comparison of Excess Noise Factors for SCISCM and SCIDCM Conventional APDs.** Show that a single-carrier-injection single-carrier-multiplication (SCISCM) conventional APD with pure electron injection and no hole multiplication ($\mathsf{k} = 0$) has an excess noise factor $F \approx 2$ for all appreciable values of the gain. Use (19.4-8) to show that the mean gain is then $G = \exp(\alpha_e w)$. Calculate the responsivity of a Si APD illuminated by photons of energy equal to its bandgap energy, $E_g = 1.12$ eV, assuming that $\eta = 0.8$ and $\overline{G} = 70$. Determine the excess noise factor for a single-carrier-injection double-carrier-multiplication (SCIDCM) Si APD when $\mathsf{k} = 0.01$. Compare the SCISCM and SCIDCM results for F.

*19.6-2 **Mean Gain of a Staircase APD.** Use the Bernoulli probability law to demonstrate that the mean gain of a single-carrier-injection single-carrier-multiplication (SCISCM) multilayer APD, such as the staircase device displayed in Fig. 19.6-6, is $\overline{G} = (1 + P)^N$, where P is the probability of impact ionization at each stage and N is the number of stages. Show that $\overline{G}$ reduces to the result for the conventional SCISCM APD when $P \to 0$ and $N \to \infty$. In this limit there are an infinite number of stages and the probability is vanishingly small that a carrier is produced by impact ionization in any one given stage of the device.

19.6-3 **Excess Noise Factor for a Photomultiplier Tube.**

(a) Derive an expression for the excess noise factor F of a one-stage photomultiplier tube (PMT) assuming that the number of secondary-emission electrons per incident primary electron is Poisson distributed with mean $\bar{\delta}$. Show that the results are a special case of those provided in (19.6-26) for the N-dynode PMT.

(b) When the N dynodes are all identical ($A = 1$), in the limit of high mean gain ($\overline{G} \gg 1$), show that (19.6-26) can be written as $F \approx \overline{G}^{1/N}/(\overline{G}^{1/N} - 1) = \bar{\delta}/(\bar{\delta} - 1)$. This signifies that the gain provided by a PMT is nearly noise free even without the benefit of a high-gain first dynode.

*19.6-4 **Excess Noise Factor for a Photoconductive Detector.** The gain of a photoconductive detector was specified in (19.2-3) to be $G = \tau/\tau_e$, where τ is the excess-carrier electron–hole recombination lifetime and τ_e is the electron transit time across the sample. In a more realistic representation, the gain G is taken to be a random quantity since the process of electron–hole recombination is random. Show that an exponentially distributed probability density function for the random recombination lifetime, $P(\tau) = (1/\overline{\tau}) \exp(-\tau/\overline{\tau})$, results in an excess noise factor $F = 2$. In accordance with (19.6-25), this reveals that photoconductor **generation-recombination (GR) noise** degrades the photoconductor current SNR by a factor of 2.

19.6-5 **Bandwidth of an RC Circuit.** Using the definition of bandwidth provided in (19.6-16), show that a circuit of impulse response function $h(t) = (e/\tau) \exp(-t/\tau)$ has a bandwidth $B = 1/4\tau$. What is the bandwidth of an RC circuit? Determine the thermal noise current associated with a resistance $R = 1$ kΩ at $T = 300°$ K connected to a capacitance $C = 5$ pF.

19.6-6 **Signal-to-Noise Ratio for an Analog APD Receiver.** Assuming that circuit noise is negligible, by what factor does the signal-to-noise ratio of a receiver that uses a conventional APD of mean gain $\overline{G} = 100$ change if the ionization ratio k is increased from $\mathsf{k} = 0.1$ to $\mathsf{k} = 0.2$? Show that if $\overline{G} \gg 1$ and $\overline{G} \gg 2(1 - \mathsf{k})/\mathsf{k}$, the SNR is approximately inversely proportional to $\overline{G}$.

19.6-7 **Noise in an Analog APD Receiver.** An optical receiver using a conventional APD has the following parameters: quantum efficiency $\eta = 0.8$; mean gain $\overline{G} = 100$; ionization ratio

$\mathrm{k} = 0.5$; load resistance $R_L = 1$ kΩ; temperature $T = 300°$ K; bandwidth $B = 100$ kHz; and dark/leakage current $i_d = 1$ nA. An optical signal of power 10 nW at $\lambda_o = 0.87$ μm is received. Determine the RMS values of the different noise currents, and the SNR. Assume that the dark/leakage current has a noise variance that obeys the same law as photocurrent noise and that the receiver is resistance limited.

19.6-8 **Optimal Gain for an APD in an Analog Receiver.** An analog receiver using a p–i–n photodiode has a ratio of circuit-noise variance to photoelectron-noise variance of 100. If a conventional APD with ionization ratio $\mathrm{k} = 0.2$ is used instead, determine the optimal mean gain for maximizing the SNR and the corresponding improvement in SNR.

19.6-9 **Analog Receiver Sensitivity.** Determine the receiver sensitivity, in terms of the optical power required to achieve a $\mathrm{SNR} = 10^3$, for a photodetector of quantum efficiency $\eta = 0.8$ at $\lambda_o = 1300$ nm in a circuit of bandwidth $B = 100$ MHz when there is no circuit noise. The receiver measures the electric current i.

19.6-10 **Noise Comparison for Three Photodetectors.** Consider a photodetector connected in series with a 50-Ω load resistor maintained at 77° K (liquid-nitrogen temperature) that is to be used in a 1-μm-wavelength analog optical system with a bandwidth of 1 GHz. Compare the performance of three photodetectors: 1) a p–i–n photodiode with quantum efficiency $\eta = 0.9$; 2) an APD with quantum efficiency $\eta = 0.6$, gain $\overline{G} = 100$, and ionization ratio $\mathrm{k} = 0$; and 3) a 10-stage photomultiplier tube (PMT) with quantum efficiency $\eta = 0.3$, overall mean gain $\overline{G} = 4^{10}$, and overall gain variance $\sigma_G^2 = \overline{G}^2/4$.

(a) For each photodetector, determine the photocurrent SNR when it is illuminated by a photon flux of 10^{10} s^{-1}.

(b) Which devices render the signal detectable?

*19.6-11 **Sensitivity of an AM Receiver.** A receiver with negligible circuit noise, bandwidth B, and a photodetector with responsivity R (A/W) measures a modulated optical power $P(t) = P_0 + P_s \cos(2\pi f t)$, where $f < B$. If $P_0 \gg P_s$, derive an expression for the minimum modulation power P_s that is measurable with signal-to-noise ratio $\mathrm{SNR}_0 = 30$ dB. What is the effect of the background power P_0 on the minimum observable signal P_s?

19.6-12 **Dependence of Digital Receiver Sensitivity on Wavelength.** The receiver sensitivity of an ideal digital receiver (with unity quantum efficiency and no circuit noise) operating at a wavelength 870 nm is −76 dBm. What is the sensitivity at 1300 nm if the receiver is operated at the same data rate?

19.6-13 **Bit Error Rate for a Digital Receiver.** An ideal digital receiver that makes use of a p–i–n photodiode with $\eta = 1$ is devoid of noise except for Poisson photon noise. The receiver mistakes a 870-nm optical signal of power P (logic state "1") that is present for one that is absent (logic state "0") with probability 10^{-10}. What is the probability of error under each of the following altered conditions?

(a) The wavelength is $\lambda_o = 1300$ nm.

(b) Original conditions, but now the power is doubled.

(c) Original conditions, but the photodetector quantum efficiency is now $\eta = 0.5$.

(d) Original conditions, but an ideal APD with $\eta = 1$, gain $G = 100$, and $F = 1$ (no gain noise) is used.

(e) As in (d), but the APD has an excess noise factor $F = 2$ instead.

19.6-14 **Sensitivity of a Photon-Counting Receiver.** A photodetector with quantum efficiency $\eta = 0.5$ records the number of photoelectrons received in successive time intervals of duration $T = 1$ μs. Determine the receiver sensitivity (mean number of photons required to achieve a $\mathrm{SNR} = 10^3$) assuming that the photon-number distribution is Poisson. Assuming that the wavelength of the light is $\lambda_o = 870$ nm, what is the corresponding optical power? If this optical power is detected, what is the probability that the detector registers zero counts?

*19.6-15 **Single-Dynode PMT Illuminated by Photon-Number-Squeezed Light.** Consider a photomultiplier tube (PMT) with quantum efficiency η and a single dynode. Incident on the photocathode is light from a specially designed photon source for which the probability of observing n photons in the counting time T is

$$p(n) = \begin{cases} 1/2, & n = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

When an electron strikes the dynode, either two or three secondary electrons are emitted and these proceed to the anode to be registered. The gain distribution $P(G)$ is given by

$$P(G) = \begin{cases} 1/3, & G = 2 \\ 2/3, & G = 3 \\ 0, & \text{otherwise,} \end{cases}$$

so that it is twice as likely that three secondary electrons are produced as two.

(a) Calculate the SNR and the photon-number variance-to-mean ratio for the input photon number and compare these quantities to those for a Poisson photon number of the same mean.

(b) Find the probability distribution for the photoelectron number $p(m)$, along with its SNR and variance-to-mean ratio [see Sec. 13.2D].

(c) Demonstrate that as the quantum efficiency η decreases and approaches zero, the photoelectron-number variance-to-mean ratio $\sigma_m^2/\overline{m}$ approaches unity from below, but never becomes equal to or exceeds unity.[†] The photoelectron statistics thus retain their sub-Poisson character no matter how small η becomes.

(d) Determine the mean gain $\langle G \rangle$ and the mean-square gain $\langle G^2 \rangle$ associated with the secondary-emission process.

(e) Find the excess noise factor F.

(f) Assuming that the quantum efficiency of the PMT is $\eta = 1/4$, and the counting time $T = 1.3$ ns, determine the mean anode current $\bar{\imath}$ in a circuit of bandwidth $B = 1/2T$, and calculate the current SNR.

(g) Calculate the PMT responsivity if the incident light is of wavelength $\lambda_o = 1550$ nm.

(h) Explain why (19.6-23) for σ_i^2 is not applicable.

[†] See M. C. Teich and B. E. A. Saleh, Effects of Random Deletion and Additive Noise on Bunched and Antibunched Photon-Counting Statistics, *Optics Letters*, vol. 7, pp. 365–367, 1982.

CHAPTER

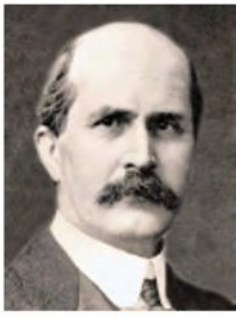
20

ACOUSTO-OPTICS

20.1 INTERACTION OF LIGHT AND SOUND 945
- A. Bragg Diffraction
- *B. Coupled-Wave Theory
- C. Bragg Diffraction of Beams

20.2 ACOUSTO-OPTIC DEVICES 958
- A. Modulators
- B. Scanners
- C. Space Switches
- D. Filters, Frequency Shifters, and Isolators

*20.3 ACOUSTO-OPTICS OF ANISOTROPIC MEDIA 967



Sir William Henry Bragg (1862–1942, left) and **Sir William Lawrence Bragg (1890–1971**, right), a father-and-son team, were awarded the Nobel Prize in 1915 for their studies of the diffraction of light from periodic structures such as those created by sound.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

The refractive index of an optical medium can be altered by the presence of sound. **Acousto-optics** is the study of how sound modifies the effect such a medium has on light, as schematically illustrated in Fig. 20.0-1. Many useful photonic devices make use of the ability of sound to control light; these include optical modulators, switches, deflectors, filters, isolators, frequency shifters, and spectrum analyzers.

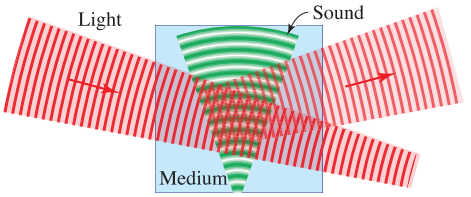


Figure 20.0-1 Sound can modify the effect of an optical medium on light.

Sound is a dynamic strain involving molecular vibrations that takes the form of a wave traveling at a velocity characteristic of the medium (the velocity of sound). As an example, a harmonic plane wave of compressions and rarefactions in a gas is depicted in Fig. 20.0-2. In those regions where the medium is compressed, the density of gas is higher and its refractive index is larger; where the medium is rarefied, its density and refractive index are smaller. In solids, sound involves vibrations of the molecules about their equilibrium positions; this alters the optical polarizability of the material and thus its refractive index.

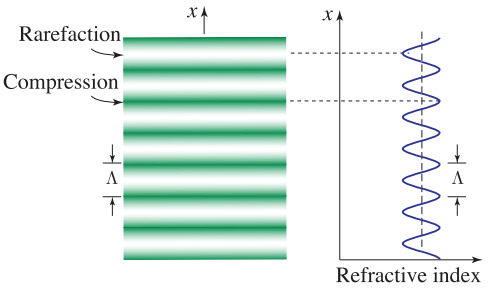


Figure 20.0-2 Variation of the refractive index of a material accompanying a harmonic sound wave. The pattern has a period Λ (the wavelength of sound) and travels with the velocity of sound in the medium.

Hence, an acoustic wave creates a perturbation of the refractive index in the form of a wave. The medium becomes a *dynamic graded-index medium* — an inhomogeneous medium with a time-varying, stratified refractive index. The theory of acousto-optics deals with the perturbation of the refractive index caused by sound, and with the propagation of light through this perturbed, time-varying, inhomogeneous medium.

The propagation of light in static (as opposed to time-varying) inhomogeneous (graded-index) media has been examined in Secs. 1.3, 2.4C, and 5.2B. Since optical frequencies are far higher than acoustic frequencies, the variations of the refractive index in a medium perturbed by sound are invariably very slow in comparison with the optical period. As a consequence, an adiabatic approach is suitable wherein the optical propagation problem is solved separately at every instant of time during the relatively slow course of the acoustic cycle, always treating the material as if it were a static (frozen) inhomogeneous medium. In this quasi-stationary approximation, acousto-optics reduces to the optics of inhomogeneous media, usually periodic, controlled by sound.

The simplest form of the interaction of light and sound is the partial reflection of an optical plane wave from the stratified parallel planes representing the refractive-index variations created by an acoustic plane wave (Fig. 20.0-3). A set of parallel reflectors separated by the wavelength of sound Λ will reflect light if the angle of incidence θ satisfies the **Bragg condition** for constructive interference (2.5-13),

$$\sin\theta_B = \frac{\lambda}{2\Lambda}, \tag{20.0-1}$$

Bragg Condition

where λ is the wavelength of light in the medium [Exercise 2.5-3 and (7.1-42)]. This form of light–sound interaction is known as **Bragg diffraction**, **Bragg reflection**, or **Bragg scattering**. Devices that make use of it are known as **Bragg reflectors**, **Bragg deflectors**, **Bragg cells**, **acousto-optic cells**, or **acousto-optic modulators**.

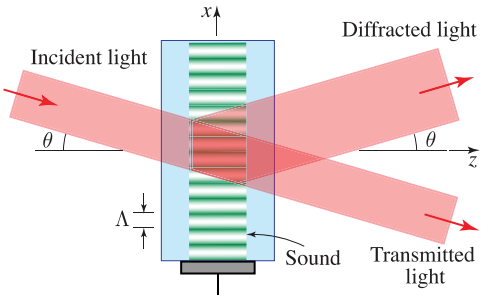


Figure 20.0-3 Bragg diffraction: an acoustic plane wave acts as a partial reflector of light (a beamsplitter) when the angle of incidence θ satisfies the Bragg condition specified in (20.0-1).

This Chapter

In Sec. 20.1, a simplified theory of the optics of Bragg diffraction is presented for linear, nondispersive, and isotropic media. Though the theory is based on wave optics, so that the polarization of light and sound are ignored, a simple quantum interpretation of the interaction emerges. Bragg cells using electrically controlled acoustic transducers have found widespread application in photonics and their use for the modulation and scanning of light is considered in Sec. 20.2. Section 20.3 provides a brief introduction to anisotropic and polarization effects in acousto-optics.

20.1 INTERACTION OF LIGHT AND SOUND

The effect of a scalar acoustic wave on a scalar optical wave is described in this section. We first consider optical and acoustic plane waves, and then examine the interaction of optical and acoustic beams.

A. Bragg Diffraction

Consider an acoustic plane wave of frequency f (angular frequency $\Omega = 2\pi f$) and wavelength $\Lambda = v_s/f$ (wavenumber $q = 2\pi/\Lambda$) traveling in the x direction (vertically)

in a medium with sound velocity v_s. The strain (relative displacement) at time t and position x in the medium is

$$s(x,t) = S_0 \cos(\Omega t - qx), \tag{20.1-1}$$

where S_0 is the strain amplitude. The acoustic intensity I_s (units of $\mathrm{W/m^2}$) is

$$I_s = \tfrac{1}{2}\varrho\, v_s^3 S_0^2, \tag{20.1-2}$$

where ϱ is the mass density of the medium.

Refractive-index perturbation. The medium is assumed to be optically transparent and its refractive index in the absence of sound is n. The sound-induced strain $s(x,t)$ creates a proportional perturbation of the refractive index that is obtained via a Taylor-series expansion analogous to that used for the Pockels effect in (21.1-4),

$$\Delta n(x,t) \approx -\tfrac{1}{2}\mathrm{p}n^3 s(x,t), \tag{20.1-3}$$

where p is a dimensionless phenomenological coefficient known as the **elasto-optic coefficient** or **strain-optic coefficient**. The minus sign indicates that positive strain (dilation) leads to a reduction of the refractive index. As a consequence, the medium has a time-varying inhomogeneous refractive index that takes the form of a wave

$$n(x,t) = n - \Delta n_0 \cos(\Omega t - qx), \tag{20.1-4}$$

with amplitude

$$\Delta n_0 = \tfrac{1}{2}\mathrm{p}n^3 S_0. \tag{20.1-5}$$

Substituting (20.1-2) into (20.1-5) reveals that the refractive-index change is proportional to the square root of the acoustic intensity,

$$\boxed{\Delta n_0 = \sqrt{\tfrac{1}{2}\mathcal{M} I_s}\,.} \tag{20.1-6}$$

The quantity

$$\boxed{\mathcal{M} = \frac{\mathrm{p}^2 n^6}{\varrho\, v_s^3}} \tag{20.1-7}$$

Acousto-Optic Figure of Merit

is a material parameter that represents the effectiveness of sound in altering the refractive index. The quantity $\mathcal{M}$ is thus a figure of merit that indicates the strength of the acousto-optic effect in the material.

EXAMPLE 20.1-1. ***Acousto-Optic Figure of Merit for Flint Glass.*** Extra-dense flint glass is characterized by the parameters $\varrho = 6.3 \times 10^3\ \mathrm{kg/m^3}$, $v_s = 3.1\ \mathrm{km/s}$, $n = 1.92$, and $\mathrm{p} = 0.25$, so that $\mathcal{M} = 1.67 \times 10^{-14}\ \mathrm{m^2/W}$. An acoustic wave of intensity $10\ \mathrm{W/cm^2}$ thus creates a refractive-index wave of amplitude $\Delta n_0 = 2.89 \times 10^{-5}$.

Amplitude reflectance. Consider now an optical plane wave traveling in this medium with frequency ν, angular frequency $\omega = 2\pi\nu$, free-space wavelength $\lambda_o = c_o/\nu$, wavelength in the unperturbed medium $\lambda = \lambda_o/n$ corresponding to wavenumber $k = n\omega/c_o$, and wavevector $\mathbf{k}$ lying in the x–z plane and making an angle θ with the z axis, as illustrated in Fig. 20.1-1.

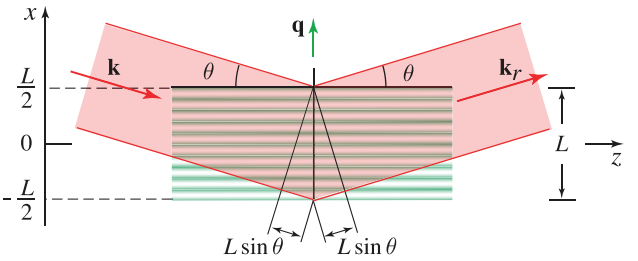


Figure 20.1-1 Reflections from layers of an inhomogeneous medium.

Because the acoustic frequency f is typically smaller than the optical frequency ν by at least five orders of magnitude, we consider the light–sound interaction in terms of an adiabatic approach, as mentioned earlier. Hence, the refractive index is taken to be a static "frozen" sinusoidal function

$$n(x) = n - \Delta n_0 \cos(qx - \varphi), \tag{20.1-8}$$

where φ is a fixed phase. We determine the reflected light from this inhomogeneous (graded-index) medium and track its slow variation with time by taking $\varphi = \Omega t$.

To determine the amplitude of the reflected wave (wavevector $\mathbf{k}_r$) we divide the medium into incremental planar layers orthogonal to the x axis. The incident optical plane wave is partially reflected at each layer because of the refractive-index change. To begin, we assume that the reflectance is sufficiently small so that the transmitted light from each layer approximately maintains its original magnitude (i.e., is not depleted) as it penetrates through the subsequent layers of the medium.

If $\Delta\mathrm{r} = (d\mathrm{r}/dx)\,\Delta x$ is the incremental complex amplitude reflectance of a layer of incremental width Δx at position x, the total complex amplitude reflectance for an overall length L (Fig. 20.1-1) is the sum of all incremental reflectances,

$$\mathrm{r} = \int_{-L/2}^{L/2} e^{j2kx\sin\theta}\,\frac{d\mathrm{r}}{dx}\,dx. \tag{20.1-9}$$

The phase factor $e^{j2kx\sin\theta}$ accommodates the fact that the reflected wave at a position x is advanced by a distance $2x\sin\theta$, corresponding to a phase shift $2kx\sin\theta$, relative to the reflected wave at $x = 0$. The wavenumbers of the incident and reflected waves are taken to be the same.

Using (20.1-8), we write

$$\frac{d\mathrm{r}}{dx} = \frac{d\mathrm{r}}{dn}\,\frac{dn}{dx} = \frac{d\mathrm{r}}{dn}\,q\,\Delta n_0 \sin(qx - \varphi), \tag{20.1-10}$$

where the derivative $d\mathrm{r}/dn$, which may be obtained from the Fresnel equations of reflection as will be shown subsequently, is not dependent on x. Substituting (20.1-10)

into (20.1-9), and using complex notation for $\sin(qx-\varphi) = [e^{j(qx-\varphi)} - e^{-j(qx-\varphi)}]/2j$, we have

$$\mathrm{r} = j\mathrm{r}_0 e^{j\varphi}\frac{1}{L}\int_{-\frac{1}{2}L}^{\frac{1}{2}L} e^{j(2k\sin\theta - q)x}\,dx - j\mathrm{r}_0 e^{-j\varphi}\frac{1}{L}\int_{-\frac{1}{2}L}^{\frac{1}{2}L} e^{j(2k\sin\theta + q)x}\,dx, \tag{20.1-11}$$

where

$$\mathrm{r}_0 = \tfrac{1}{2}\Delta n_0\, qL\,\frac{d\mathrm{r}}{dn}. \tag{20.1-12}$$

Carrying out the integrals in (20.1-11) and substituting $\varphi = \Omega t$, we finally obtain

$$\mathrm{r} = \mathrm{r}_+ + \mathrm{r}_-\,, \tag{20.1-13}$$

where

$$\boxed{\mathrm{r}_\pm = \pm j\mathrm{r}_0\ \mathrm{sinc}\left[(2k\sin\theta \mp q)\frac{L}{2\pi}\right] e^{\pm j\Omega t}} \tag{20.1-14}$$

Amplitude Reflectance

and $\mathrm{sinc}(x) \equiv \sin(\pi x)/(\pi x)$.

For reasons that will become clear shortly, the terms r_+ and r_- are called the upshifted and downshifted Bragg amplitude reflectances, respectively. The upshifted reflectance r_+ has its maximum value when $2k\sin\theta = q$, whereas the downshifted reflection is maximum when $2k\sin\theta = -q$. If L is sufficiently large, these maxima are sharp, so that any slight deviation from the angles $\theta = \pm\sin^{-1}(q/2k)$ renders the corresponding term negligible. Thus, only one of these two terms can be significant at a time, depending on the angle θ. We first consider the upshifted condition, $2k\sin\theta \approx q$, in which case the downshifted reflection is negligible; we comment on the downshifted case subsequently.

Bragg Condition

The sinc function in (20.1-14) attains its maximum value of 1.0 when its argument is zero, i.e., when $q = 2k\sin\theta$ for upshifted reflection. This occurs when $\theta = \theta_B$, where $\theta_B = \sin^{-1}(q/2k)$ is the **Bragg angle**. Since $q = 2\pi/\Lambda$ and $k = 2\pi/\lambda$, we have

$$\boxed{\sin\theta_B = \frac{\lambda}{2\Lambda}.} \tag{20.1-15}$$

Bragg Angle

The Bragg angle is the angle at which the incremental reflections from planes separated by an acoustic wavelength Λ have a phase shift of 2π so that they interfere constructively [see Exercise 2.5-3 and (7.1-42)].

EXAMPLE 20.1-2. ***Bragg Angle for Flint Glass.*** An acousto-optic cell is made of extra-dense flint glass of refractive index $n = 1.92$, in which the velocity of sound is $v_s = 3.1$ km/s. The Bragg angle for reflection of an optical wave of free-space wavelength $\lambda_o = 633$ nm ($\lambda = \lambda_o/n \approx 330$ nm) from a sound wave of frequency $f = 100$ MHz ($\Lambda = v_s/f = 31$ μm) is $\theta_B = 5.3$ mrad $\approx 0.30°$. This angle is internal (i.e., inside the medium). If the cell is placed in air, θ_B corresponds to an external angle $\theta'_B \approx n\theta_B = 0.59°$. A sound wave of 10 times greater frequency ($f = 1$ GHz) corresponds to a Bragg angle $\theta_B = 3.0°$.

The Bragg condition can also be stated as a simple relation between the wavevectors of the sound wave and the optical waves. If $\mathbf{q} = (q, 0, 0)$, $\mathbf{k} = (-k \sin\theta, 0, k \cos\theta)$, and $\mathbf{k}_r = (k \sin\theta, 0, k \cos\theta)$ represent the components of the wavevectors of the sound wave, the incident light wave, and the reflected light wave, respectively, the condition $q = 2k \sin\theta_B$ is equivalent to the vector relation

$$\mathbf{k}_r = \mathbf{k} + \mathbf{q}, \tag{20.1-16}$$

Bragg Condition

as illustrated by the vector diagram in Fig. 20.1-2.

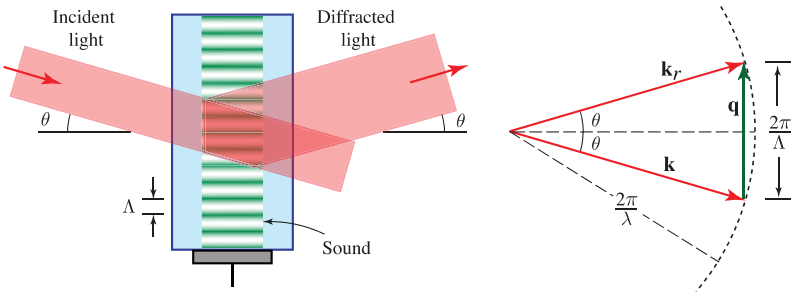


Figure 20.1-2 The Bragg condition $\sin\theta_B = q/2k$ is equivalent to the vector relation $\mathbf{k}_r = \mathbf{k} + \mathbf{q}$. For a sound wave traveling in the upward direction, the directions of the incident optical wave and the sound wave form an acute angle and the frequency of the diffracted wave is upshifted.

Tolerance in the Bragg Condition

The dependence of the complex amplitude reflectance on the angle θ is governed by the symmetric function $\text{sinc}[(q - 2k \sin\theta)L/2\pi] = \text{sinc}[(\sin\theta - \sin\theta_B)2L/\lambda]$ in (20.1-14). This function reaches its peak value when $\theta = \theta_B$ and drops sharply when θ differs slightly from θ_B. When $\sin\theta - \sin\theta_B = \lambda/2L$ the sinc function reaches its first zero and the intensity reflectance $|r|^2$ vanishes (Fig. 20.1-3). Because θ_B is usually very small, $\sin\theta \approx \theta$, and the reflectance vanishes at an angular deviation from the Bragg angle of approximately $\theta - \theta_B \approx \lambda/2L$. Since L is typically much greater than λ, this is an extremely small angular width. This sharp reduction of the reflectance for slight deviations from the Bragg angle occurs as a result of the destructive interference between the incremental reflections from the sound wave.

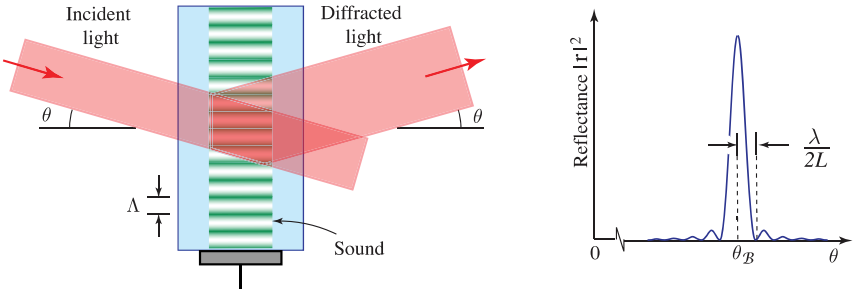


Figure 20.1-3 Dependence of the intensity reflectance $|r|^2$ on the angle θ. Maximum reflection occurs at the Bragg angle $\theta_B = \sin^{-1}(\lambda/2\Lambda)$.

Doppler Shift

In accordance with (20.1-14), the complex amplitude reflectance r_+ is proportional to $\exp(j\Omega t)$. Since the angular frequency of the incident light is ω [i.e., $E \propto \exp(j\omega t)$], the reflected wave $E_r = \mathrm{r}_+ E \propto \exp[j(\omega + \Omega)t]$ has angular frequency

$$\boxed{\omega_r = \omega + \Omega.} \qquad \text{(20.1-17)}$$
Doppler Shift

The process of reflection is therefore accompanied by a frequency shift equal to the frequency of the sound. This can be viewed as a Doppler shift (Exercise 2.6-1 and Sec. 14.3D). The incident light is reflected from surfaces that move with velocity v_s. Its Doppler-shifted angular frequency is therefore $\omega_r = \omega(1 + 2v_s \sin\theta/c)$, where $v_s \sin\theta$ is the component of velocity of these surfaces along the direction of the incident and reflected waves. Using the relations $\sin\theta = \lambda/2\Lambda$, $v_s = \Lambda\Omega/2\pi$, and $c = \lambda\omega/2\pi$ reproduces (20.1-17). The magnitude of the Doppler shift equals the sound frequency.

Because $\Omega \ll \omega$, the frequencies of the incident and reflected waves are approximately equal (the difference is typically smaller than 1 part in 10^5) so that the wavelengths of the two waves are therefore also approximately equal. In writing (20.1-9) we have implicitly made use of this assumption by using the same wavenumber k for the two waves. Also, in drawing the vector diagram in Fig. 20.1-2 it was assumed that the vectors $\mathbf{k}_r$ and $\mathbf{k}$ have approximately the same length, $n\omega/c_o$.

Intensity Reflectance

The reflectance $\mathcal{R} = |\mathrm{r}_+|^2$ is the ratio of the intensity of the reflected optical wave to that of the incident optical wave. At the Bragg angle $\theta = \theta_B$, (20.1-14) gives $\mathcal{R} = |\mathrm{r}_0|^2$ so that substituting from (20.1-12) yields

$$\mathcal{R} = \tfrac{1}{4}\,\Delta n_0^2\, q^2 L^2 \left|\frac{d\mathrm{r}}{dn}\right|^2. \qquad \text{(20.1-18)}$$

An expression for the derivative $d\mathrm{r}/dn$ may be obtained by use of the Fresnel equations (Sec. 6.2) to determine the incremental complex-amplitude reflectance $\Delta\mathrm{r}$ in terms of the incremental refractive-index change Δn between two adjacent layers. For TE (orthogonal) polarization, (6.2-8) is used with $n_1 = n + \Delta n$, $n_2 = n$, and $\theta_1 = 90^\circ - \theta$; Snell's law $n_1 \sin\theta_1 = n_2 \sin\theta_2$ provides θ_2. When terms of second order in Δn are neglected, the result is $\Delta\mathrm{r} \approx -\Delta n/2n \sin^2\theta$ so that

$$\frac{d\mathrm{r}}{dn} \approx \frac{-1}{2n\sin^2\theta}. \qquad \text{(20.1-19)}$$

Equation (6.2-9) is similarly used for the TM (parallel) polarization, yielding

$$\frac{d\mathrm{r}}{dn} \approx \frac{-\cos 2\theta}{2n\sin^2\theta}. \qquad \text{(20.1-20)}$$

In most acousto-optic devices θ is very small, so that $\cos 2\theta \approx 1$, rendering (20.1-19) approximately applicable for both polarizations.

Substituting (20.1-19) into (20.1-18), and using the Bragg condition $q = 2k\sin\theta = (4\pi n \sin\theta/\lambda_o)$, then leads to

$$\mathcal{R} = \frac{\pi^2}{\lambda_o^2}\left(\frac{L}{\sin\theta}\right)^2 \Delta n_0^2. \qquad \text{(20.1-21)}$$

Using (20.1-6), we conclude that the intensity reflectance

$$\mathcal{R} = \frac{\pi^2}{2\lambda_o^2}\left(\frac{L}{\sin\theta}\right)^2 \mathcal{M} I_s \tag{20.1-22}$$

Intensity Reflectance

is proportional to the intensity of the acoustic wave I_s, to the material parameter $\mathcal{M}$ defined in (20.1-7), and to the square of the oblique distance $L/\sin\theta$ of penetration of light through the acoustic wave. Finally, substituting $\sin\theta = \lambda/2\Lambda$ into (20.1-22) gives rise to

$$\mathcal{R} = 2\pi^2 n^2 \frac{L^2\Lambda^2}{\lambda_o^4}\mathcal{M} I_s\,. \tag{20.1-23}$$

Hence, the intensity reflectance is inversely proportional to λ_o^4 (or directly proportional to ω^4). The dependence of the efficiency of scattering on the fourth power of the optical frequency is typical for light-scattering phenomena (see Secs. 5.6B, 10.3A, and 14.5C).

The purported proportionality between the reflectance and the sound intensity is problematical, however. As the sound intensity increases, $\mathcal{R}$ would eventually exceed unity, and the reflected light would be more intense than the incident light. This result is a consequence of a violation of the assumptions of this approximate theory. It was assumed at the outset that the incremental reflection from each layer was too small to deplete the transmitted wave that reflects from subsequent layers. Clearly, this assumption does not hold when the sound wave is intense. Saturation then occurs, ensuring that $\mathcal{R}$ does not exceed unity (see also Sec. 7.1C). As will be shown in Sec. 20.1B, a more careful analysis that accommodates depletion of the incident optical wave leads to an expression for the exact reflectance $\mathcal{R}_e$ given by

$$\mathcal{R}_e = \sin^2\sqrt{\mathcal{R}}\,, \tag{20.1-24}$$

where $\mathcal{R}$ is the approximate reflectance provided in (20.1-22). Evidently, when $\mathcal{R} \ll 1$, $\sin\sqrt{\mathcal{R}} \approx \sqrt{\mathcal{R}}$, so that $\mathcal{R}_e \approx \mathcal{R}$, as illustrated in Fig. 20.1-4.

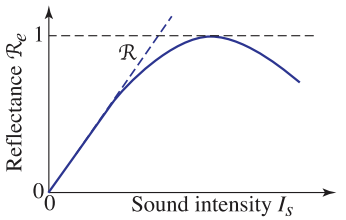


Figure 20.1-4 Dependence of the exact reflectance $\mathcal{R}_e$ of a Bragg reflector on the intensity of sound I_s. When I_s is small $\mathcal{R}_e \approx \mathcal{R}$, which is a linear function of I_s.

EXAMPLE 20.1-3. ***Reflectance of a Flint-Glass Bragg Cell.*** Consider a Bragg cell made of extra-dense flint glass, which has a figure of merit $\mathcal{M} = 1.67 \times 10^{-14}$ m^2/W (Example 20.1-1). For $\lambda_o = 633$ nm (a wavelength of the He–Ne laser), a sound intensity $I_s = 10$ W/cm^2, and a penetration length of the light through the sound $L/\sin\theta = 1$ mm, we find that $\mathcal{R} = 0.0206$ and $\mathcal{R}_e = 0.0205$, so that approximately 2% of the light is reflected from the cell. If the sound intensity is increased to 100 W/cm^2, however, then $\mathcal{R} = 0.206$ and $\mathcal{R}_e = 0.192$, revealing that the depletion of the incident optical wave must be accommodated and that the reflectance increases to $\approx 19\%$.

Downshifted Bragg Diffraction

According to (20.1-14), another possible geometry for Bragg diffraction is that for which $2k \sin\theta = -q$. This is satisfied when the angle θ is negative; i.e., when the directions of the incident optical wave and the sound wave make an obtuse angle, as illustrated in Fig. 20.1-5. In this case, the downshifted reflectance r_- in (20.1-14) attains its maximum value, while the upshifted reflectance r_+ is negligible. The complex amplitude reflectance is then

$$\mathrm{r}_- = -j\mathrm{r}_0 e^{-j\Omega t}. \tag{20.1-25}$$

In this geometry, the frequency of the reflected optical wave, denoted ω_s, is downshifted, so that

$$\omega_s = \omega - \Omega \tag{20.1-26}$$

and the wavevectors of the light and sound waves satisfy the relation

$$\mathbf{k}_s = \mathbf{k} - \mathbf{q}, \tag{20.1-27}$$

as illustrated in Fig. 20.1-5. Equation (20.1-27) is a phase-matching condition, ensuring that the light reflections add in phase. The frequency downshift in (20.1-26) is consistent with the Doppler shift since the light and its parallel sound-wave component travel in the same direction in this configuration.

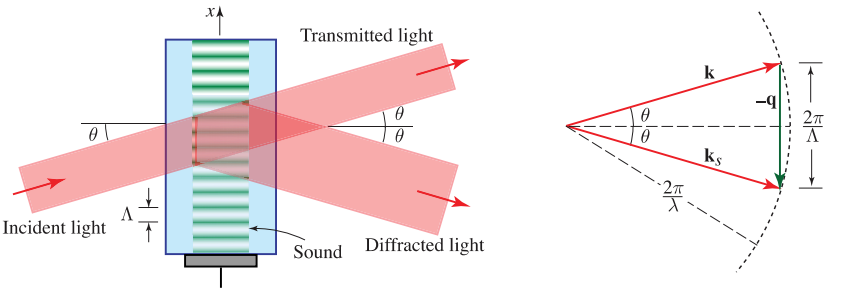


Figure 20.1-5 Geometry of downshifted diffraction of light from a sound wave traveling in the upward direction. The directions of the incident optical wave and the sound wave make an obtuse angle with respect to each other and the frequency of the diffracted wave is downshifted.

Quantum Interpretation

In accordance with the quantum theory of light (Chapter 13), an optical wave of angular frequency ω and wavevector $\mathbf{k}$ is viewed as a stream of photons, each of energy $\hbar\omega$ and momentum $\hbar\mathbf{k}$. An acoustic wave of angular frequency Ω and wavevector $\mathbf{q}$ may be similarly regarded as a stream of acoustic quanta, called **phonons**, each of energy $\hbar\Omega$ and momentum $\hbar\mathbf{q}$.

From a quantum perspective, the interaction of light and sound involves a photon and a phonon combining to generate a new photon with the sum energy and momentum. An incident photon of frequency ω and wavevector $\mathbf{k}$ thus interacts with a phonon of frequency Ω and wavevector $\mathbf{q}$ to generate a new photon of frequency ω_r and wavevector $\mathbf{k}_r$, as illustrated in Fig. 20.1-6 (see also Fig. 14.5-5). Conservation of energy and momentum require that $\hbar\omega_r = \hbar\omega + \hbar\Omega$ and $\hbar\mathbf{k}_r = \hbar\mathbf{k} + \hbar\mathbf{q}$, from which the Doppler shift formula (20.1-17), $\omega_r = \omega + \Omega$, and the Bragg condition

(20.1-16), $\mathbf{k}_r = \mathbf{k} + \mathbf{q}$, are recovered. The preceding argument is applicable for upshifted Bragg diffraction; a parallel argument exists for the downshifted case. The quantum interpretation provided here is similar to that offered for Brillouin and Raman scattering in Sec. 14.5C.

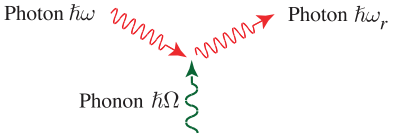


Figure 20.1-6 Bragg diffraction from a quantum perspective: a photon combines with a phonon to generate a new photon with modified frequency and momentum.

*B. Coupled-Wave Theory

Bragg Diffraction as a Scattering Process

As described in Sec. 5.2B, light propagating through a homogeneous medium with a slowly varying inhomogeneous refractive-index perturbation Δn is described by the wave equation (5.2-20),

$$\nabla^2 \mathcal{E} - \frac{1}{c^2}\frac{\partial^2 \mathcal{E}}{\partial t^2} \approx -\mathcal{S}, \tag{20.1-28}$$

where the radiation source

$$\mathcal{S} = -\mu_o \frac{\partial^2 \Delta \mathcal{P}}{\partial t^2} = -2\mu_o \epsilon_o n \frac{\partial^2}{\partial t^2}(\Delta n\, \mathcal{E}) \tag{20.1-29}$$

is proportional to the second derivative of the product $\Delta n\, \mathcal{E}$. For Bragg diffraction the perturbation Δn is created by a sound wave, so the scattering source is dependent on both the acoustic field and the optical field $\mathcal{E}$, which includes both the incident and scattered fields. An approximate method for solving this scattering problem makes use of the **first Born approximation**, which relies on the assumption that the scattering source $\mathcal{S}$ is created by the incident field, rather than by the actual field. Once the scattering source is known, the wave equation can be solved for the scattered field.

We assume that the incident optical field is a plane wave,

$$\mathcal{E} = \text{Re}\{A \exp[j(\omega t - \mathbf{k}\cdot\mathbf{r})]\} \tag{20.1-30}$$

and that the perturbation engendered by the acoustic wave is also a plane wave,

$$\Delta n = -\Delta n_0 \cos(\Omega t - \mathbf{q}\cdot\mathbf{r}). \tag{20.1-31}$$

Substituting these two equations into (20.1-29), and reordering the terms of the product $\Delta n\, \mathcal{E}$, leads to

$$\mathcal{S} = -\left(\frac{\Delta n_0}{n}\right)\left(k_r^2\, \text{Re}\left\{A \exp\left[j(\omega_r t - \mathbf{k}_r\cdot\mathbf{r})\right]\right\} + k_s^2\, \text{Re}\left\{A \exp\left[j(\omega_s t - \mathbf{k}_s\cdot\mathbf{r})\right]\right\}\right), \tag{20.1-32}$$

where $\omega_r = \omega + \Omega$, $\mathbf{k}_r = \mathbf{k} + \mathbf{q}$, $k_r = \omega_r/c$; and $\omega_s = \omega - \Omega$, $\mathbf{k}_s = \mathbf{k} - \mathbf{q}$, $k_s = \omega_s/c$. The two sources of radiation, with frequencies $\omega \pm \Omega$ and wavevectors $\mathbf{k} \pm \mathbf{q}$, may emit upshifted or downshifted Bragg-reflected plane waves. Upshifted reflection occurs if the geometry of the experimental configuration is such that the magnitude of the vector

$\mathbf{k}+\mathbf{q}$ equals $\omega_r/c \approx \omega/c$, as can easily be seen from the vector diagram in Fig. 20.1-2. Downshifted reflection occurs if the vector $\mathbf{k}-\mathbf{q}$ has magnitude $\omega_s/c \approx \omega/c$, as illustrated in Fig. 20.1-5. Clearly, these two conditions cannot be simultaneously satisfied.

The foregoing analysis provides an independent proof for the Bragg condition and Doppler-shift formula based on a scattering approach. Equation (20.1-32) confirms that the intensity of the emitted light is proportional to ω^4, so that the efficiency of scattering is inversely proportional to the fourth power of the wavelength. This analysis can be pursued further to obtain an expression for the intensity reflectance by determining the intensity of the wave emitted by the scattering source (Prob. 20.1-2).

Coupled-Wave Equations

To reach beyond the first Born approximation, we must include the contribution made by the scattered field to the source $\mathcal{S}$. Assuming that the geometry is that of upshifted Bragg diffraction, the field $\mathcal{E}$ then comprises the incident and Bragg-reflected waves: $\mathcal{E} = \text{Re}\{E\exp(j\omega t)\} + \text{Re}\{E_r\exp(j\omega_r t)\}$. With the help of the relation $\Delta n = -\Delta n_0\cos(\Omega t - \mathbf{q}\cdot\mathbf{r})$, (20.1-29) provides

$$\mathcal{S} = \text{Re}\big\{S\exp(j\omega t) + S_r\exp(j\omega_r t)\big\} + \text{terms of other frequencies}, \tag{20.1-33}$$

where

$$S = -k^2\frac{\Delta n_0}{n}E_r, \quad S_r = -k_r^2\frac{\Delta n_0}{n}E. \tag{20.1-34}$$

Comparing terms of equal frequencies on both sides of the wave equation (20.1-28) leads to a pair of coupled Helmholtz equations for the incident wave and the Bragg-reflected wave:

$$(\nabla^2 + k^2)E = -S, \qquad (\nabla^2 + k_r^2)E_r = -S_r. \tag{20.1-35}$$

These equations, with the help of (20.1-34), may be solved to determine E and E_r.

Consider, as an example, the case of small-angle Bragg diffraction ($\theta \ll 1$), so that the two waves travel approximately in the z direction as portrayed in Fig. 20.1-7. Assuming that $k \approx k_r$, the fields E and E_r are described by $E = A\exp(-jkz)$ and $E_r = A_r\exp(-jkz)$, respectively, where the envelopes A and A_r are slowly varying functions of z. Using the slowly varying envelope approximation set forth in Sec. 2.2C, we can set $(\nabla^2 + k^2)A\exp(-jkz) \approx -j2k(dA/dz)\exp(-jkz)$, whereupon (20.1-34) and (20.1-35) yield coupled first-order differential equations for the envelopes

$$\boxed{\begin{aligned}\frac{dA}{dz} &= j\tfrac{1}{2}\gamma A_r \qquad (20.1\text{-}36a)\\ \frac{dA_r}{dz} &= j\tfrac{1}{2}\gamma A, \qquad (20.1\text{-}36b)\end{aligned}}$$

with

$$\gamma = k\frac{\Delta n_0}{n}. \tag{20.1-37}$$

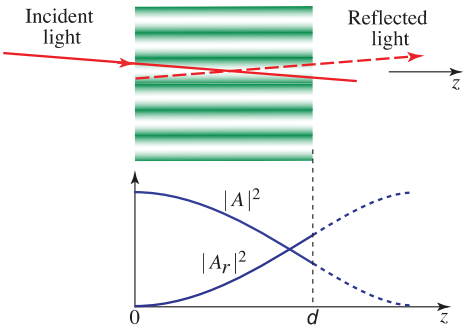


Figure 20.1-7 *Top*: Incident and reflected waves for small-angle Bragg diffraction. *Bottom*: The intensity of the incident optical wave decays while that of the Bragg-reflected wave grows as the distance of travel through the acoustic wave in the Bragg cell increases.

If the Bragg cell extends from $z = 0$ to $z = d$ (Fig. 20.1-7), and we use the boundary condition $A_r(0) = 0$, (20.1-36) have the harmonic solutions

$$A(z) = A(0)\cos\left(\frac{\gamma z}{2}\right) \tag{20.1-38a}$$

$$A_r(z) = jA(0)\sin\left(\frac{\gamma z}{2}\right). \tag{20.1-38b}$$

These equations describe the growth of the reflected wave and the decay of the incident wave, as illustrated in Fig. 20.1-7. The reflectance $\mathcal{R}_e = |A_r(d)|^2/|A(0)|^2$ is therefore $\mathcal{R}_e = \sin^2(\gamma d/2)$, so that $\mathcal{R}_e = \sin^2\sqrt{\mathcal{R}}$, where $\mathcal{R} = (\gamma d/2)^2$. Using (20.1-37), we then have $\mathcal{R} = (\pi^2/\lambda_o^2)\, d^2\, \Delta n_0^2$, which exactly matches the expression for the weak-sound reflectance provided in (20.1-21) with $d = L/\sin\theta$.

C. Bragg Diffraction of Beams

It has been shown thus far that an *optical plane wave* of wavevector $\mathbf{k}$ interacts with an *acoustic plane wave* of wavevector $\mathbf{q}$ to produce an optical plane wave of wavevector $\mathbf{k}_r = \mathbf{k}+\mathbf{q}$, provided that the Bragg condition is satisfied (i.e., the angle between $\mathbf{k}$ and $\mathbf{q}$ is such that the magnitude $k_r = |\mathbf{k}+\mathbf{q}| \approx k = 2\pi/\lambda$, as illustrated in Fig. 20.1-2).

The interaction between a *beam of light* and a *beam of sound* can be understood if the beams are regarded as superpositions of plane waves traveling in different directions, each with its own wavevector (see the introduction to Chapter 4).

Diffraction of an Optical Beam from an Acoustic Plane Wave

Consider an *optical beam* of width D interacting with an *acoustic plane wave*. In accordance with Fourier optics (Sec. 4.3A), the optical beam can be decomposed into a collection of plane waves whose directions occupy a cone of half-angle

$$\delta\theta = \frac{\lambda}{D}. \tag{20.1-39}$$

The coefficient of proportionality in (20.1-39) is taken to be unity for simplicity, but actually depends on the beam profile: 1) For a rectangular profile of width D, the angular width from the peak to the first zero of the Fraunhofer diffraction pattern is $\delta\theta = \lambda/D$ in accordance with (4.3-7); 2) For a circular profile of diameter D, $\delta\theta = 1.22\lambda/D$ in accordance with (4.3-9); and 3) For a Gaussian beam of waist diameter $D = 2W_0$, $\delta\theta = \lambda/\pi W_0 = (2/\pi)\lambda/D \approx 0.64\lambda/D$ in accordance with (3.1-20).

Though there is only a single wavevector $\mathbf{q}$ for the acoustic plane wave, there are many wavevectors $\mathbf{k}$ (all of the same length $2\pi/\lambda$) within the cone of angle $\delta\theta$ for the

optical beam. As illustrated in Fig. 20.1-8, however, there is only one direction of $\mathbf{k}$ for which the Bragg condition is satisfied. The diffracted optical wave thus has only a single wavevector $\mathbf{k}_r$ and is therefore a plane wave.

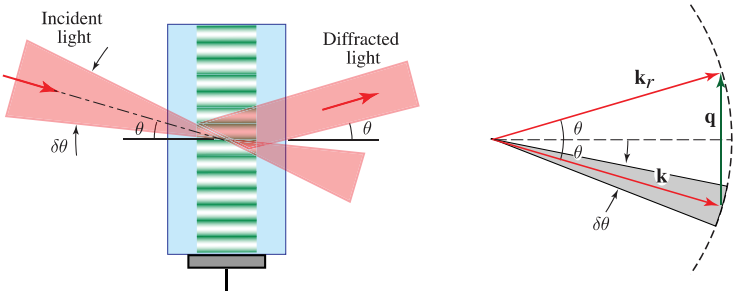


Figure 20.1-8 Diffraction of an *optical beam* from an *acoustic plane wave*. There is only a single plane-wave component of the incident light beam that satisfies the Bragg condition. The diffracted light is therefore a plane wave.

Diffraction of an Optical Beam from an Acoustic Beam

Suppose now that the acoustic wave itself is a beam of width D_s. If the sound frequency is sufficiently high so that the wavelength is much smaller than the width of the medium, sound propagates as an unguided (free-space) wave and has properties analogous to those of light, and exhibits an angular divergence

$$\delta\theta_s = \frac{\Lambda}{D_s}. \tag{20.1-40}$$

The sound beam thus comprises a collection of acoustic plane waves whose directions lie within the divergence angle $\delta\theta_s$. The Bragg diffraction of an *optical beam* from an *acoustic beam* can be determined by identifying matching pairs of optical and acoustic plane waves that satisfy the Bragg condition. The sum of the reflected waves constitutes the reflected optical beam. There are many vectors $\mathbf{k}$ (all of the same length $2\pi/\lambda$) and many vectors $\mathbf{q}$ (all of the same length $2\pi/\Lambda$), but only those pairs of vectors that result in isosceles triangles obeying the Bragg condition contribute, as illustrated in Fig. 20.1-9.

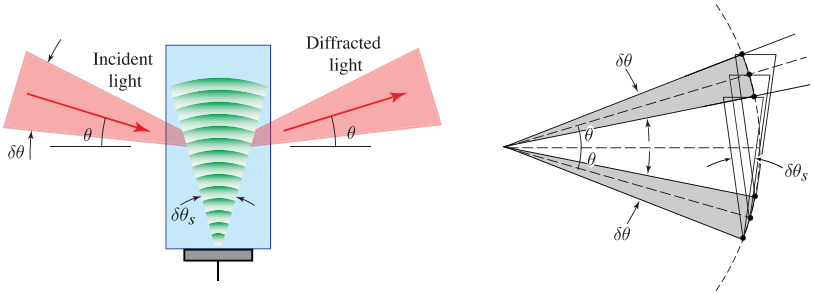


Figure 20.1-9 Diffraction of an *optical beam* from an *acoustic beam*. There are many plane-wave components of the incident light beam that satisfy the Bragg condition so the diffracted light is an optical beam.

If the acoustic-beam divergence is greater than that of the optical-beam ($\delta\theta_s \gg \delta\theta$), and if the central directions of the two beams satisfy the Bragg condition, every incident optical plane wave finds an acoustic match and the diffracted light beam has the same

angular divergence as the incident optical beam $\delta\theta$. The distribution of acoustic energy in the sound beam can then be monitored as a function of direction by using a probe light beam of much narrower divergence and measuring the diffracted light as the angle of incidence is varied.

Diffraction of an Optical Plane Wave from a Thin Acoustic Beam: Raman–Nath Diffraction

Since a thin acoustic beam comprises plane waves traveling in many directions, it can diffract light at angles that are significantly different from the Bragg angle corresponding to the beam's principal direction. Consider, for example, the geometry of Fig. 20.1-10 in which an incident optical plane wave impinges perpendicularly to the principal direction of a thin acoustic beam. This configuration for the acousto-optic diffraction of light by sound is known as **Raman–Nath diffraction** or **Debye–Sears diffraction**.

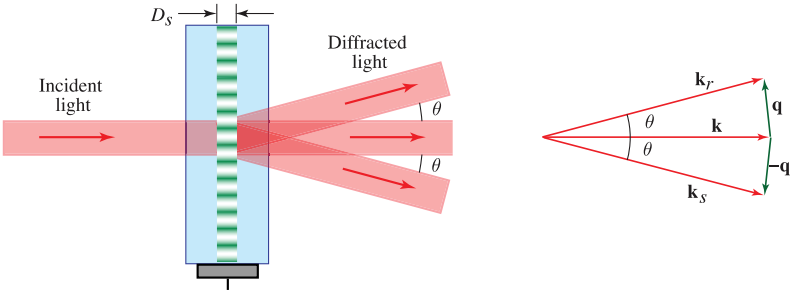


Figure 20.1-10 Raman–Nath diffraction of light by sound. An optical plane wave normally incident on a thin-beam acoustic traveling or standing wave is partially deflected into two directions, at angles $\theta \approx \pm\lambda/\Lambda$ with respect to the direction of the incident wave.

As can be understood from the vector diagram provided in Fig. 20.1-10, the Bragg condition is satisfied if the diffracted wavevector makes angles $\pm\theta$ with respect to the direction of the incident optical wave, where the angle θ is given by

$$\sin\frac{\theta}{2} = \frac{\lambda}{2\Lambda}. \tag{20.1-41}$$

If θ is small, we have $\sin(\theta/2) \approx \theta/2$, which leads to

$$\boxed{\theta \approx \frac{\lambda}{\Lambda}.} \tag{20.1-42}$$

The incident light is diffracted at the angles $\pm\theta$, whether the thin acoustic beam is traveling upward or downward. For an acoustic standing-wave beam, the optical wave is diffracted at both angles from both traveling-wave components that comprise the standing wave.

The angle $\theta \approx \lambda/\Lambda$ is in fact the angle at which a diffraction grating of period Λ diffracts an incident plane wave (Exercise 2.4-5). Indeed, the thin acoustic beam serves to modulate the refractive index of the acousto-optic material, creating a pattern of period Λ confined to a thin planar layer. The acousto-optic medium thus acts as a thin diffraction grating.

Such an acousto-optic phase grating also allows light to be diffracted into higher orders, at angles $\pm 2\theta, \pm 3\theta, \ldots$, as described by (2.4-12) and illustrated in Fig. 20.1-11(a). Such higher-order diffracted waves can be interpreted via a generalized quantum picture of the light–sound interaction portrayed in Fig. 20.1-6: an incident photon combines with *two* phonons to form a photon of the second-order diffracted wave. Conservation of momentum dictates that the diffracted beams then have wavevectors $\mathbf{k} \pm 2\mathbf{q}$ [one of these beams, $\mathbf{k}_r = \mathbf{k} + 2\mathbf{q}$, is illustrated in Fig. 20.1-11(b)]; the second-order diffracted light is concomitantly frequency shifted to $\omega \pm 2\Omega$. The same approach can be used to describe higher orders of diffraction.

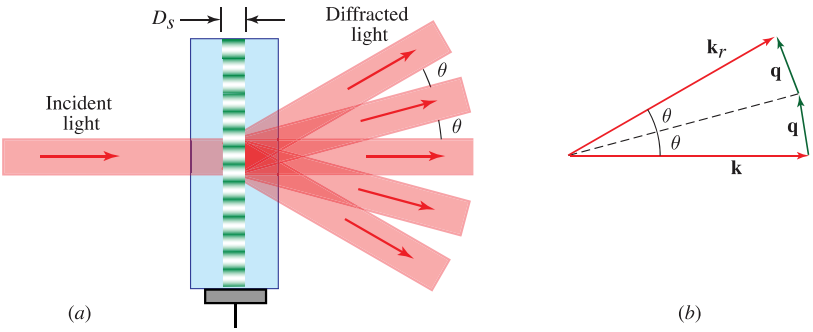


Figure 20.1-11 (a) A thin acoustic beam acts as a diffraction grating. (b) Conservation-of-momentum diagram for second-order acousto-optic diffraction.

20.2 ACOUSTO-OPTIC DEVICES

The materials used for the fabrication of acousto-optic devices are generally chosen on the basis of the figure of merit $\mathcal{M} = \mathrm{p}^2 n^6 / \varrho\, v_s^3$ that characterizes the strength of the acousto-optic effect, as provided in (20.1-7). In addition to the region of wavelength transparency, key material parameters are the refractive index, mass density, sound velocity, and elasto-optic coefficient. Commonly used materials in the visible and near-IR include extra-dense flint glass (Examples 20.1-1, 20.1-2, and 20.1-3), fused silica, crystalline quartz, and TeO_2; in the mid-IR, Ge and chalcogenide glasses are often used. Longitudinal acoustic waves are most commonly employed in such devices because of their high diffraction efficiency, but transverse acoustic waves have the merit that they offer polarization-independent operation in certain configurations. Integrated-optic devices often rely on $LiNbO_3$, which is piezoelectric so that it allows the generation of on-chip surface acoustic waves via surface-mounted metallic electrodes.

As discussed in this section, acousto-optic modulators are useful in a wide variety of applications in photonics, including:

- Analog modulation of optical intensity
- Digital switching of optical intensity
- Scanning an optical beam
- Routing an optical beam to selected directions
- Spectral analysis of an acoustic beam
- Spectral filtering of an optical beam
- Shifting the frequency of an optical beam
- Serving as an optical isolator

A. Modulators

The intensity of the light diffracted from a Bragg cell is proportional to the intensity of the applied acoustic wave, provided that the latter is sufficiently weak [see (20.1-22)]. Using an electrically controlled acoustic transducer [Fig. 20.2-1(a)], the reflected light intensity can thus be varied proportionally. Such a device, known as an **acousto-optic modulator (AOM)**, serves as a linear analog modulator of light.

As the acoustic intensity increases and saturation sets in, it is possible to attain essentially unity reflectance (Fig. 20.1-4). The modulator can then serve as an **acousto-optic switch**. As illustrated in Fig. 20.2-1(b), switching the sound ON and OFF results in the reflected light being switched ON and OFF, while the transmitted light is concomitantly switched OFF and ON. Acousto-optic switches find use in Q switching and cavity dumping for solid-state lasers, as well as in active mode-locking and laser-pulse selection (Sec. 16.4), among other uses.

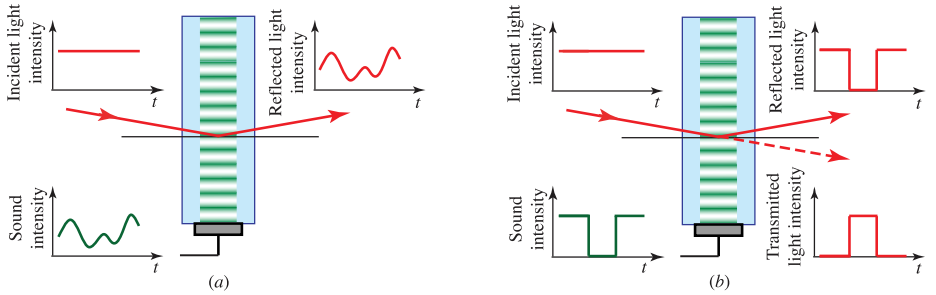


Figure 20.2-1 (a) An acousto-optic modulator that uses an electrically controlled piezoelectric transducer. The intensity of the diffracted light is proportional to the intensity of the sound. (b) An acousto-optic switch. The presence of sound causes the light to be reflected while its absence allows the light to be transmitted.

Modulation Bandwidth

The bandwidth of the modulator is the maximum frequency at which it can efficiently impart modulation to the diffracted light. When the amplitude of an acoustic wave of frequency f_0 is varied as a function of time via amplitude modulation with a signal of bandwidth B, the acoustic wave is no longer a single-frequency harmonic function; it then has frequency components that lie within a band $f_0 \pm B$ centered about the frequency f_0 (Fig. 20.2-2). How does monochromatic light interact with this multifrequency acoustic wave and what is the maximum value of B that can be accommodated by the acousto-optic modulator?

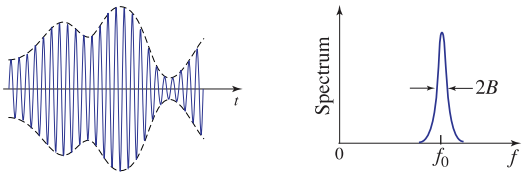


Figure 20.2-2 Sample waveform of an amplitude-modulated acoustic signal and its spectrum.

When both the incident optical wave and the acoustic wave are plane waves, the component of sound of frequency f corresponds to a Bragg angle

$$\theta = \sin^{-1}\frac{\lambda}{2\Lambda} = \sin^{-1}\frac{f\lambda}{2v_s} \approx \frac{\lambda}{2v_s}f, \tag{20.2-1}$$

assuming that θ is small. For a fixed angle of incidence θ, an incident monochromatic optical plane wave of wavelength λ thus interacts with one and only one harmonic component of the acoustic wave, i.e., the component with frequency f that satisfies (20.2-1). This frequency corresponds to a wavenumber $q = (2\pi/v_s)f$, and the vector diagram illustrated in Fig. 20.2-3 shows that the diffracted wave is then *monochromatic* with frequency $\nu + f$. Though the acoustic wave is modulated, the diffracted optical wave is not. Evidently, under this idealized condition the bandwidth of the modulator is zero!

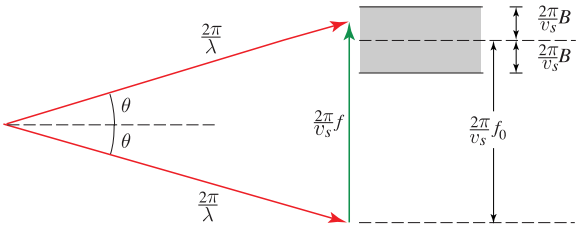


Figure 20.2-3 Interaction of an optical plane wave with a modulated (multifrequency) acoustic plane wave. Only one frequency component of the sound diffracts the light wave. The diffracted wave is monochromatic and not modulated.

To achieve modulation with a bandwidth B, each of the acoustic frequency components within the band $f_0 \pm B$ must interact with the incident light wave. A situation with more tolerance is evidently required. Suppose that the incident light is a beam of width D and angular divergence $\delta\theta = \lambda/D$ and that the modulated sound wave is planar. Each frequency component of the sound can then interact with the optical plane wave that has the matching Bragg angle (Fig. 20.2-4).

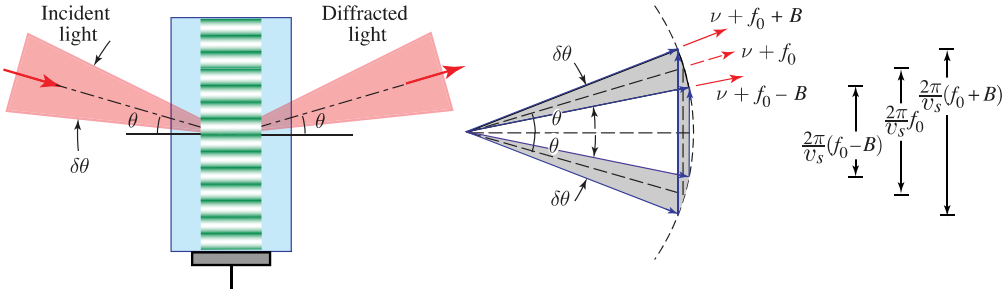


Figure 20.2-4 Interaction of an optical beam of angular divergence $\delta\theta$ with acoustic plane waves whose frequencies are confined to the band $f_0 \pm B$. There are many parallel $\mathbf{q}$ vectors of different lengths; each matches a direction in the incident light beam.

As is clear from the figure, the frequency band $f_0 \pm B$ is accommodated by an optical beam of angular divergence

$$\delta\theta \approx \frac{(2\pi/v_s)B}{2\pi/\lambda} = \frac{\lambda}{v_s}B. \tag{20.2-2}$$

The bandwidth of the modulator is therefore determined by

$$B = v_s\frac{\delta\theta}{\lambda} = \frac{v_s}{D}, \tag{20.2-3}$$

which we write as

$$B = \frac{1}{T} \quad \text{where} \quad T = \frac{D}{v_s}. \tag{20.2-4}$$

Modulator Bandwidth

The quantity T represents the transit time of the sound across the waist of the light beam, a result that is readily understood since a time T is required to change the amplitude of the sound wave at all points in the light–sound interaction region. The bandwidth can therefore be increased by focusing the light beam to a smaller diameter.

EXERCISE 20.2-1

Parameters of Acousto-Optic Modulators. Determine the Bragg angle and maximum bandwidth for the following acousto-optic modulators:

Modulator 1

Material: Fused silica ($n = 1.46$, $v_s = 6$ km/s)
Sound: Frequency $f = 50$ MHz
Light: He–Ne laser, wavelength $\lambda_o = 633$ nm; angular divergence $\delta\theta = 1$ mrad

Modulator 2

Material: Tellurium ($n = 4.8$, $v_s = 2.2$ km/s)
Sound: Frequency $f = 100$ MHz
Light: CO_2 laser, wavelength $\lambda_o = 10.6$ μm; beam width $D = 1$ mm

B. Scanners

An acousto-optic cell can be used to scan light. The fundamental concept is based on the linear relation between the Bragg angle θ and the sound frequency f expressed in (20.2-1). The angle of deflection 2θ is therefore given by

$$2\theta \approx \frac{\lambda}{v_s} f, \tag{20.2-5}$$

where θ is taken to be sufficiently small so that $\sin\theta \approx \theta$. By changing the sound frequency f, the deflection angle 2θ can be varied.

Scanning with an acoustic plane wave. One difficulty is that θ represents both the angle of incidence and the angle of reflection. Effecting a change in just the angle of reflection requires a simultaneous change in both the angle of incidence and the sound frequency. Changing the angle of incidence may be accomplished by tilting the sound beam, as illustrated in Fig. 20.2-5. However, creating such a tilt requires a complex system that uses, for example, a phased array of acoustic transducers (several acoustic transducers driven at relative phases selected to impart a tilt to the overall generated sound wave). Changing the sound frequency requires a frequency modulator (FM), which must be synchronized with phased array governing the angle of tilt.

Scanning with an acoustic beam. The requirement of tilting the sound wave may be mitigated by making use a sound beam with an angular divergence equal to or greater than the entire range of directions to be scanned. As the sound frequency is changed, the Bragg angle is altered and the incoming light wave selects only that acoustic plane-wave component with the matching direction. Though the efficiency of such a system is expected to be low, we proceed to examine some of its properties.

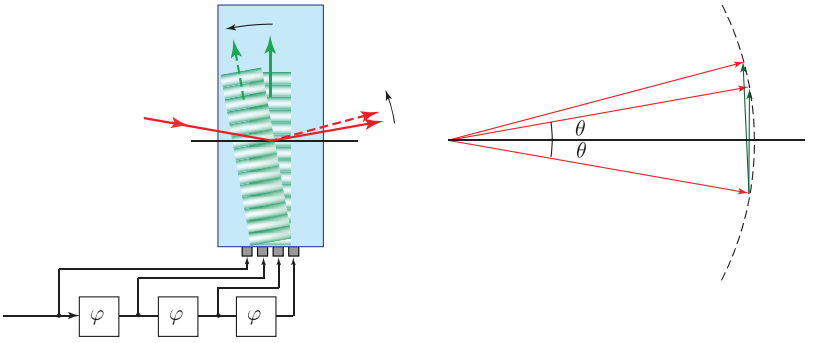


Figure 20.2-5 Scanning by changing the sound frequency *and* acoustic-wave direction. The planar sound wave is tilted by using an array of transducers driven by signals that differ by a phase φ.

Scan Angle

For a sound frequency f, the incident light wave interacts with the sound component at an angle $\theta = (\lambda/2v_s)f$ and is deflected by an angle $2\theta = (\lambda/v_s)f$, as illustrated in Fig. 20.2-6. By varying the sound frequency from f_0 to $f_0 + B$, the deflection angle 2θ is swept over a scan angle

$$\boxed{\Delta\theta = \frac{\lambda}{v_s}B.} \qquad \begin{matrix}(20.2\text{-}6)\\ \text{Scan Angle}\end{matrix}$$

This assumes, of course, that the sound beam has an angular width $\delta\theta_s = \Lambda/D_s \geq \Delta\theta$. Since the scan angle is inversely proportional to the speed of sound, larger scan angles are obtained by using materials with a small sound velocity v_s.

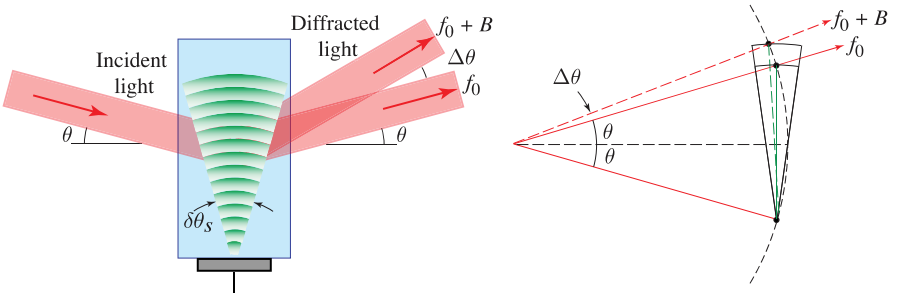


Figure 20.2-6 Scanning an optical wave by varying the frequency of a sound beam over the frequency range $f_0 \leq f \leq f_0 + B$. The angular divergence of the sound beam is $\delta\theta_s$.

Number of Resolvable Spots

If the optical wave itself has a residual angular width $\delta\theta = \lambda/D$, and assuming that $\delta\theta \ll \delta\theta_s$, then the deflected beam also has a width $\delta\theta$. The number of resolvable spots of the scanner (the number of nonoverlapping angular widths within the scanning range) is then given by

$$N = \frac{\Delta\theta}{\delta\theta} = \frac{(\lambda/v_s)B}{\lambda/D} = \frac{D}{v_s}B, \qquad (20.2\text{-}7)$$

which we write as

$$N = TB \quad \text{with} \quad T = \frac{D}{v_s}, \tag{20.2-8}$$

Number of Resolvable Spots

where B is the bandwidth of the FM modulator used to generate the sound and $T = D/v_s$ is the transit time of sound across the light beam (Fig. 20.2-7). The number of resolvable spots, which is equal to the time–bandwidth product TB, represents the degrees of freedom of the scanner and is a significant indicator of its capabilities. Increasing N requires a large transit time T, which is opposite to the design requirement for an acousto-optic modulator, where the modulation bandwidth $B = 1/T$ is enhanced by selecting a small value of T.

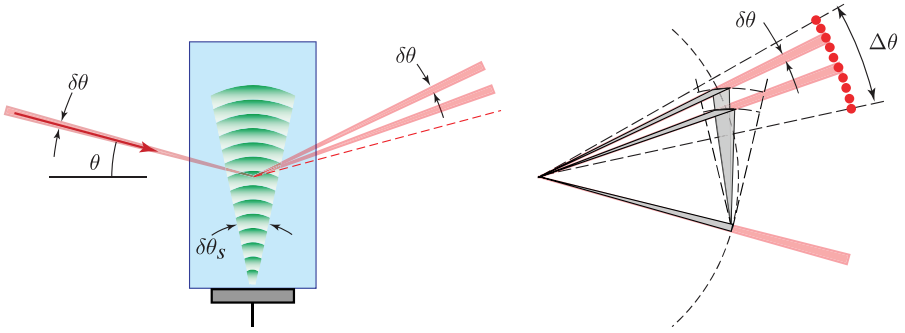


Figure 20.2-7 Number of resolvable spots of an acousto-optic scanner.

EXERCISE 20.2-2

Parameters of an Acousto-Optic Scanner. A fused-silica acousto-optic scanner ($v_s = 6$ km/s, $n = 1.46$) is used to scan a He–Ne laser beam ($\lambda_o = 633$ nm). The sound frequency is scanned over the range 40 to 60 MHz. To what width should the laser beam be focused so that the number of resolvable spots is $N = 100$? What is the scan angle $\Delta\theta$? What would be the effect of using a material with a smaller velocity of sound, such as extra-dense flint glass for which $v_s = 3.1$ km/s?

The Acousto-Optic Scanner as an Acoustic Spectrum Analyzer

The proportionality between the Bragg angle and the sound frequency stated in (20.2-1) can be used to make an acoustic spectrum analyzer. A sound wave containing a collection of different frequencies disperses the incident light in corresponding different directions. The intensity of the deflected light in a given direction is proportional to the intensity of the sound component at the corresponding frequency (Fig. 20.2-8).

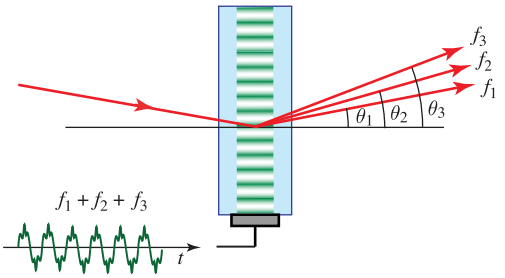


Figure 20.2-8 Each frequency component of the sound wave deflects light in a different direction. The acousto-optic cell thus serves as an acoustic spectrum analyzer.

C. Space Switches

An acousto-optic cell can be used as a space switch (Sec. 24.3) that routes information carried by one or more optical beams to one or more selected directions. Several interconnection schemes are possible:

- An acousto-optic cell in which the frequency of the acoustic wave is one of N possible values, $f_1, f_2, \ldots,$ or f_N, reflects an incident optical beam in one of N corresponding directions, $\theta_1, \theta_2, \ldots,$ or θ_N, as illustrated in Fig. 20.2-9. The device routes one beam to any of N directions.

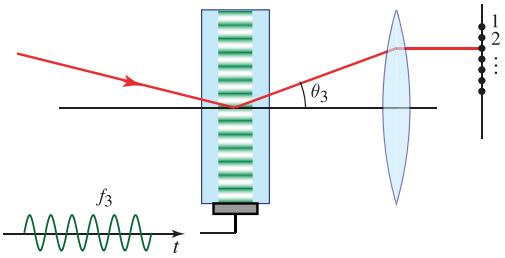


Figure 20.2-9 Routing an optical beam to one of N directions. By applying an acoustic wave of frequency f_3, for example, the optical beam is reflected at an angle θ_3 and routed to point 3.

- Using an acoustic wave comprising two frequencies that are simultaneously present, f_1 and f_2, allows the incident optical beam to be simultaneously reflected in the two corresponding directions, θ_1 and θ_2. A single beam is thereby connected to any pair of many possible directions, as illustrated in Fig. 20.2-10. More generally, an acoustic wave comprising M simultaneously present frequencies allows the incident beam to be simultaneously routed in M directions. An example of this configuration is the acoustic spectrum analyzer schematized in Fig. 20.2-8. The light beam is simultaneously routed to M points, and the intensity at each point is proportional to the intensity of the corresponding sound-frequency component.

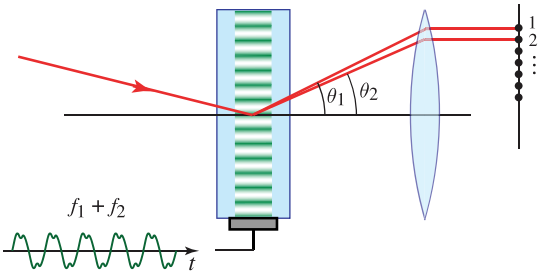


Figure 20.2-10 Routing a light beam simultaneously in a number of directions.

- The length of the acousto-optic cell may be divided into two segments. At a certain time, an acoustic wave of frequency f_1 is present in one segment and an acoustic wave of frequency f_2 is present in the other. This can be accomplished

by generating the acoustic wave from a frequency shift keyed electric signal in the form of two pulses: a pulse of frequency f_1 followed by another of frequency f_2, each with a duration $T/2$, where $T = w/v_s$ is the transit time of sound through the length w of the cell (Fig. 20.2-11). When the leading edge of the acoustic wave reaches the end of the cell, the cell processes two incoming optical beams by reflecting the top beam in the direction θ_1 corresponding to f_1, and the bottom beam in the direction θ_2 corresponding to f_2. This is a switch that connects each of two beams to any of many possible directions. By placing more than one frequency component in each segment, each of the two beams can itself be simultaneously routed in several directions.

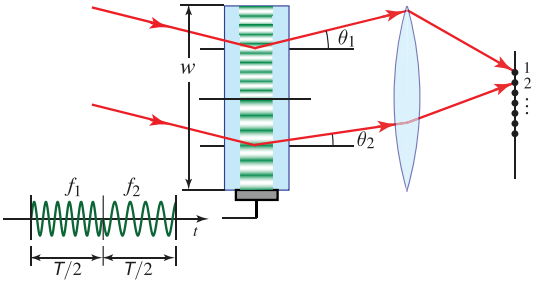


Figure 20.2-11 Routing each of two light beams in a set of specified directions. The acoustic wave is generated by a frequency shift keyed electric signal.

- The cell may also be divided into N segments, each carrying a harmonic acoustic wave of the same frequency f but of different amplitude. The result is a **spatial light modulator** (**SLM**) that modulates the intensities of N input beams (Fig. 20.2-12). Spatial light modulators are useful in optical signal processing (Sec. 21.1E).

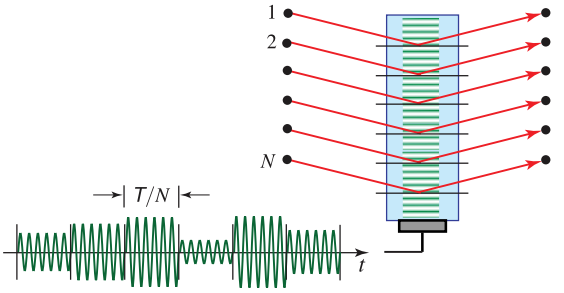


Figure 20.2-12 The spatial light modulator modulates N optical beams. The acoustic wave is driven by an amplitude-modulated electric signal.

- The most general interconnection architecture is one for which the cell is divided into L segments, each carrying an acoustic wave with M frequencies. The device acts as a random access switch that simultaneously routes each of L incoming beams in M directions (Fig. 20.2-13).

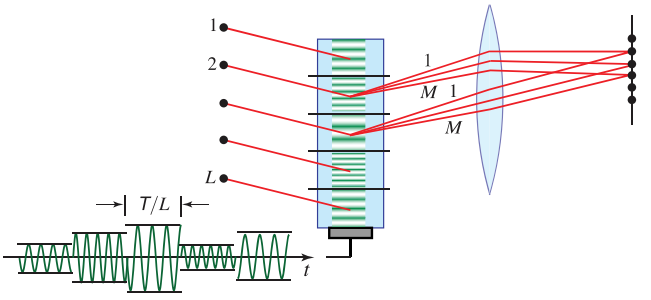


Figure 20.2-13 An arbitrary-interconnection switch routing each of L incoming light beams for random access to M points.

Interconnection Capacity

There is an upper limit to the number of interconnections that may be established by an acousto-optic device, as will be shown below. If an acousto-optic cell is used to route each of L incoming optical beams to a maximum of M directions simultaneously, the product ML cannot exceed the time–bandwidth product $N = TB$, where T is the transit time through the cell and B is the bandwidth of the acoustic wave:

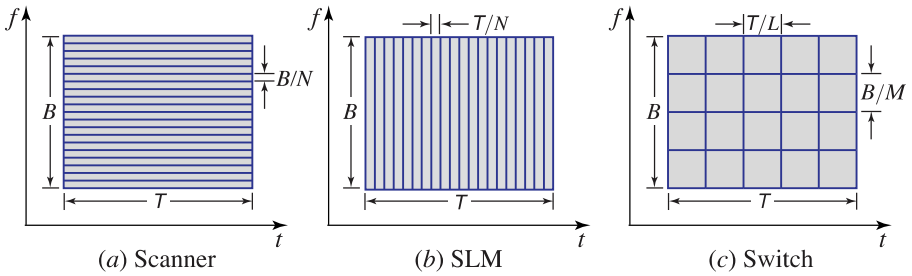
$$ML \leq N. \tag{20.2-9}$$

Interconnection Capacity

This upper bound on the number of interconnections is called the **interconnection capacity** of the device.

An acousto-optic cell with L segments relies on an acoustic wave composed of L segments, each of time duration T/L. For each segment to address M independent points, the acoustic wave must carry M independent frequency components per segment. For a signal of duration T/L there is an inherent frequency uncertainty of L/T Hz. The M frequency components must therefore be separated by at least that uncertainty. For the M components to be placed within the available bandwidth B, we must have $M(L/T) \leq B$, from which $ML \leq TB$ so that (20.2-9) follows.

A single optical beam ($L = 1$), for example, can be connected to any of $N = TB$ points, but each of two beams can be connected to at most $N/2$ points, and so on. It is a question of dividing an available time–bandwidth product $N = TB$ in the form of L time segments, each containing M independent frequencies. Examples of the possible choices are illustrated in the time–frequency diagram presented in Fig. 20.2-14.



(a) Scanner (b) SLM (c) Switch

Figure 20.2-14 Several examples of dividing the time–bandwidth region TB in the time–frequency diagram into $N = TB$ subdivisions (in this diagram $N = 20$). (*a*) *Scanner*: a single time segment containing N frequency segments. (*b*) *Spatial light modulator*: N time segments, each containing one frequency component. (*c*) *Interconnection switch*: L time segments, each containing $M = N/L$ frequency segments (in this diagram, $N = 20$, $M = 4$, and $L = 5$).

D. Filters, Frequency Shifters, and Isolators

The acousto-optic cell is useful for a number of other devices, including filters, frequency shifters, and optical isolators.

Tunable Acousto-Optic Filters

The Bragg condition $\sin\theta = \lambda/2\Lambda$ relates the angle θ, the acoustic wavelength Λ, and the optical wavelength λ. If θ and Λ are specified, Bragg reflection can occur only for a single optical wavelength, $\lambda = 2\Lambda \sin\theta$. This wavelength-selective property can be used to filter an optical wave containing a broad spectrum of wavelengths. The filter is tuned by changing the angle θ or the sound frequency f.

EXERCISE 20.2-3

Resolving Power of an Acousto-Optic Filter. Show that the spectral resolving power $\lambda/\Delta\lambda$ of an acousto-optic filter is equal to fT, where f is the sound frequency, T is the transit time, and $\Delta\lambda$ is the minimum resolvable wavelength difference.

Frequency Shifters

Optical frequency shifters are useful in many photonic applications, including optical heterodyne systems, optical FM modulators, and laser Doppler velocimeters. An acousto-optic cell may be used as a tunable frequency shifter since the Bragg reflected light is frequency shifted (up or down) by the frequency of the sound (Sec. 20.1A). In a heterodyne optical receiver, a received amplitude-modulated or phase-modulated optical signal is mixed with a coherent optical wave from a local light source, acting as a local oscillator with a different frequency. The two optical waves beat with each other (Sec. 2.6B) and the detected signal varies at the difference frequency. Information relating to the amplitude and phase of the received signal can be extracted from the detected signal (Sec. 25.4). The acousto-optic cell offers a convenient means for imparting the frequency shift required for the heterodyning process.

Optical Isolators

An optical isolator is a one-way optical valve that is often used to prevent reflected light from retracing its path back into the original light source (Secs. 6.6D and 24.1C). Optical isolators are sometimes used with laser diodes since the reflected light can adversely affect the lasing process. The acousto-optic cell can serve as an isolator. If part of the frequency-upshifted Bragg-diffracted light is reflected onto itself by a mirror and traces its path back into the cell, as illustrated in Fig. 20.2-15, it undergoes a second Bragg diffraction accompanied by a second frequency upshift. Since the frequency of the returning light differs from that of the original light by twice the sound frequency, a filter may be used to block it. Even without a filter, the laser process may be insensitive to the frequency-shifted light.

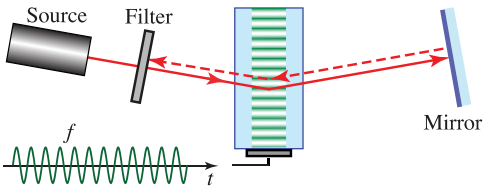


Figure 20.2-15 An acousto-optic isolator.

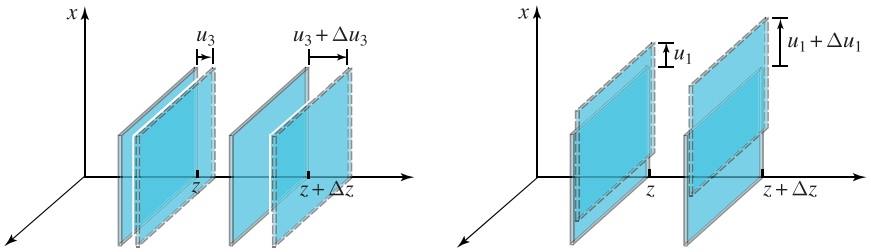
*20.3 ACOUSTO-OPTICS OF ANISOTROPIC MEDIA

The scalar theory of the interaction of light and sound is generalized in this section to include the anisotropic properties of the medium and the effects of polarization of light and sound.

Acoustic Waves in Anisotropic Materials

An acoustic wave is a wave of material strain. Strain is defined in terms of the displacements of molecules relative to their equilibrium positions. If $\mathbf{u} = (u_1, u_2, u_3)$ is the

vector of displacement of the molecules located at position $\mathbf{x} = (x_1, x_2, x_3)$, the **strain tensor**, which is symmetrical, has components $s_{ij} = \frac{1}{2}(\partial u_i/\partial x_j + \partial u_j/\partial x_i)$, where the indices $i, j = 1, 2, 3$ denote the coordinates (x, y, z). The element $s_{33} = \partial u_3/\partial x_3$, for example, represents tensile strain (stretching) in the z direction [Fig. 20.3-1(a)], whereas s_{13} represents shear strain since $\partial u_1/\partial x_3$ is the relative movement in the x direction of two incrementally separated parallel planes normal to the z direction, as illustrated in Fig. 20.3-1(b).



(a) Tensile strain (b) Shear

Figure 20.3-1 Displacements associated with tensile strain and shear.

An acoustic wave can be longitudinal or transverse, as illustrated by the following examples.

EXAMPLE 20.3-1. ***Longitudinal Acoustic Wave.*** A wave with the displacement $u_1 = 0$, $u_2 = 0$, $u_3 = A_0 \sin(\Omega t - qz)$, where A_0 is a constant, corresponds to a strain tensor in which all components vanish except

$$s_{33} = S_0 \cos(\Omega t - qz), \tag{20.3-1}$$

where $S_0 = -qA_0$. This is a wave that stretches in the z direction and also travels in the z direction. Since the vibrations lie along the same direction as the wave propagation, the wave is longitudinal.

EXAMPLE 20.3-2. ***Transverse Acoustic Wave.*** A wave with the displacement $u_1 = A_0 \sin(\Omega t - qz)$, $u_2 = 0$, $u_3 = 0$, corresponds to a strain tensor in which all components vanish except

$$s_{13} = s_{31} = S_0 \cos(\Omega t - qz), \tag{20.3-2}$$

where $S_0 = -\frac{1}{2}qA_0$. This wave travels in the z direction but vibrates in the x direction. It is therefore a transverse (shear) wave.

The velocities of longitudinal and transverse acoustic waves are characteristics of the medium and generally depend on the direction of propagation.

The Photoelastic Effect

The optical properties of an anisotropic medium are completely characterized by the electric impermeability tensor $\boldsymbol{\eta} = \epsilon_o \boldsymbol{\epsilon}^{-1}$ (Sec. 6.3). Given $\boldsymbol{\eta}$, we can determine the index ellipsoid and hence the refractive indices for an optical wave of arbitrary polarization traveling in an arbitrary direction.

In the presence of strain, the electric impermeability tensor is modified so that η_{ij} becomes a function of the elements of the strain tensor, $\eta_{ij} = \eta_{ij}(s_{kl})$. This dependence is called the **photoelastic effect**. Each of the nine functions $\eta_{ij}(s_{kl})$ may

be expanded in a Taylor series in terms of the nine variables s_{kl}. Maintaining only the linear terms provides

$$\eta_{ij}(s_{kl}) \approx \eta_{ij}(0) + \sum_{kl} \mathrm{p}_{ijkl} s_{kl}, \qquad i,j,l,k = 1,2,3, \tag{20.3-3}$$

where the quantities $\mathrm{p}_{ijkl} = \partial\eta_{ij}/\partial s_{kl}$ are the components of a fourth-rank tensor known as the **photoelasticity tensor** (sometimes called the **elasto-optic tensor** or the **strain-optic tensor**).

Since both $\{\eta_{ij}\}$ and $\{s_{kl}\}$ are symmetrical tensors, the coefficients $\{\mathrm{p}_{ijkl}\}$ are invariant to permutations of i and j, and to permutations of k and l. There are therefore only six (rather than nine) independent values for the set (i,j) and six independent values for (k,l). The pair of indices (i,j) is usually contracted to a single index $I = 1,2,\ldots,6$ (see Table 21.2-1). The indices (k,l) are similarly contracted and denoted by the index $K = 1,2,\ldots,6$. The fourth-rank tensor p_{ijkl} is then described by a 6×6 matrix p_{IK}.

Moreover, the symmetry of the crystal requires that some of the coefficients p_{IK} vanish and that certain coefficients are related. The matrix p_{IK} for a cubic crystal, for example, has the structure

$$\mathrm{p}_{IK} = \begin{bmatrix} \mathrm{p}_{11} & \mathrm{p}_{12} & \mathrm{p}_{12} & 0 & 0 & 0 \\ \mathrm{p}_{12} & \mathrm{p}_{11} & \mathrm{p}_{12} & 0 & 0 & 0 \\ \mathrm{p}_{12} & \mathrm{p}_{12} & \mathrm{p}_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{p}_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{p}_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{p}_{44} \end{bmatrix}. \tag{20.3-4}$$

Photoelasticity Matrix (Cubic Crystal)

This matrix is also applicable for isotropic media, with the additional constraint $\mathrm{p}_{44} = \frac{1}{2}(\mathrm{p}_{11} - \mathrm{p}_{12})$, which leads to only two independent coefficients. It is analogous to the Kerr electro-optic matrix s_{IK} for isotropic media displayed in Table 21.2-3.

EXAMPLE 20.3-3. ***Longitudinal Acoustic Wave in a Cubic Crystal.*** The longitudinal acoustic wave described in Example 20.3-1 travels along one of the axes of a cubic crystal of refractive index n. Substituting (20.3-1) and (20.3-4) into (20.3-3), we find that the associated strain results in an impermeability tensor with elements

$$\eta_{11} = \eta_{22} = \frac{1}{n^2} + \mathrm{p}_{12} S_0 \cos(\Omega t - qz) \tag{20.3-5}$$

$$\eta_{33} = \frac{1}{n^2} + \mathrm{p}_{11} S_0 \cos(\Omega t - qz) \tag{20.3-6}$$

$$\eta_{ij} = 0, \qquad i \neq j. \tag{20.3-7}$$

Thus, the initially optically isotropic cubic crystal becomes a uniaxial crystal with the optic axis along the direction of the acoustic wave (z direction) and with ordinary and extraordinary refractive indices, n_o and n_e, respectively, given by

$$\frac{1}{n_o^2} = \frac{1}{n^2} + \mathrm{p}_{12} S_0 \cos(\Omega t - qz) \tag{20.3-8}$$

$$\frac{1}{n_e^2} = \frac{1}{n^2} + \mathrm{p}_{11} S_0 \cos(\Omega t - qz). \tag{20.3-9}$$

The shape of the index ellipsoid is altered periodically in time and space in the form of a wave, but the principal axes remain unchanged (Fig. 20.3-2). Since the change in the refractive indices

is typically small, the second terms in (20.3-8) and (20.3-9) are small, so that the approximation $(1+\Delta)^{-1/2} \approx 1 - \Delta/2$ for $|\Delta| \ll 1$ may be applied, whereupon (20.3-8) and (20.3-9) become

$$n_o \approx n - \tfrac{1}{2} n^3 \mathrm{p}_{12} S_0 \cos(\Omega t - qz) \tag{20.3-10}$$

$$n_e \approx n - \tfrac{1}{2} n^3 \mathrm{p}_{11} S_0 \cos(\Omega t - qz). \tag{20.3-11}$$

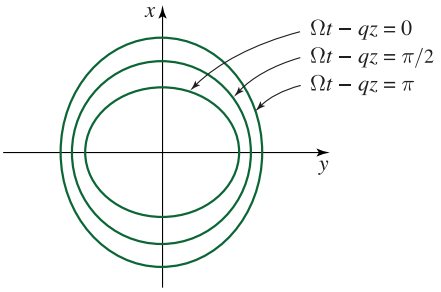


Figure 20.3-2 A longitudinal acoustic wave traveling in the z direction in a cubic crystal alters the shape of the index ellipsoid from a sphere to an ellipsoid of revolution whose dimensions vary sinusoidally with time; the optic axis is along the z direction.

EXERCISE 20.3-1

Transverse Acoustic Wave in a Cubic Crystal. The transverse acoustic wave described in Example 20.3-2 travels along one of the axes of a cubic crystal. Show that the crystal becomes biaxial with principal refractive indices

$$n_1 \approx n - \tfrac{1}{2} n^3 \mathrm{p}_{44} S_0 \cos(\Omega t - qz) \tag{20.3-12}$$

$$n_2 \approx n \tag{20.3-13}$$

$$n_3 \approx n + \tfrac{1}{2} n^3 \mathrm{p}_{44} S_0 \cos(\Omega t - qz). \tag{20.3-14}$$

In Example 20.3-3 and Exercise 20.3-1, the acoustic wave alters the principal values of the index ellipsoid but not its principal directions, so that the ellipsoid maintains its orientation. This is not always the case, however. Acoustic waves with other directions and polarizations relative to the principal axes of the crystal result in alterations of the principal refractive indices as well as the principal axes of the crystal.

Bragg Diffraction

The interaction of a linearly polarized optical wave with a longitudinal or transverse acoustic wave in an anisotropic medium can be described by the same principles as those set forth in Sec. 20.1. The incident optical wave is diffracted from the acoustic wave if the Bragg condition for constructive interference is satisfied. The analysis is a bit more complex than that for the scalar theory, however, since the incident and reflected waves travel with different velocities; thus, the angles of reflection and incidence need not be equal.

The condition for Bragg diffraction is the conservation-of-momentum (phase-matching) condition provided in (20.1-16),

$$\mathbf{k}_r = \mathbf{k} + \mathbf{q}. \tag{20.3-15}$$

Now, however, the magnitudes of these wavevectors are $k = (2\pi/\lambda_o)n$, $k_r = (2\pi/\lambda_o)n_r$, and $q = (2\pi/\Lambda)$, where λ_o and Λ are the optical and acoustic wavelengths,

and n and n_r are the refractive indices of the incident and reflected optical waves, respectively.

As illustrated in Fig. 20.3-3, if θ and θ_r are the angles of incidence and reflection, respectively, the vector equation (20.3-15) may be replaced with two scalar equations relating the z and x components of the wavevectors in the plane of incidence:

$$\frac{2\pi}{\lambda_o} n_r \cos\theta_r = \frac{2\pi}{\lambda_o} n \cos\theta \tag{20.3-16}$$

$$\frac{2\pi}{\lambda_o} n_r \sin\theta_r + \frac{2\pi}{\lambda_o} n \sin\theta = \frac{2\pi}{\Lambda}, \tag{20.3-17}$$

from which

$$n_r \cos\theta_r = n \cos\theta \tag{20.3-18a}$$

$$n_r \sin\theta_r + n \sin\theta = \frac{\lambda_o}{\Lambda}. \tag{20.3-18b}$$

Given the wavelengths λ_o and Λ, the angles θ and θ_r may be determined by solving (20.3-18). Note that n and n_r are generally functions of θ and θ_r that may be determined from the index ellipsoid of the unperturbed crystal.

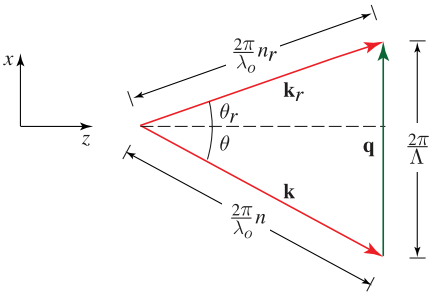
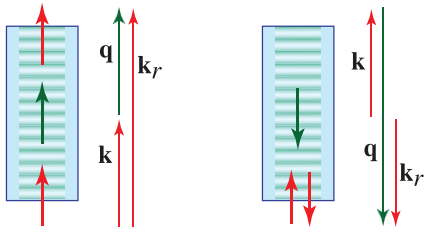


Figure 20.3-3 Bragg condition (conservation of momentum or phase-matching condition) in an anisotropic medium.

For collinear optical and acoustic waves, the phase-matching conditions are readily inferred from the vector configurations displayed in Fig. 20.3-4. For front-reflection we have $k_r = k + q$, where k_r, k, and q are the magnitudes of the respective vectors with no signs attached. This gives rise to $n_r - n = \lambda_o/\Lambda$. For back-reflection we have $k_r + k = q$, which leads to $n_r + n = \lambda_o/\Lambda$. Combining both results in the form of a single equation yields

$$n_r \pm n = \frac{\lambda_o}{\Lambda}. \tag{20.3-19}$$

For back reflection (+ sign), Λ must be smaller than λ_o, which is unlikely except for very-high-frequency acoustic waves. For front reflection (− sign), the incident and reflected waves must have different polarizations so that $n_r \neq n$.



(*a*) Front reflection (*b*) Back reflection

Figure 20.3-4 Wavevector diagrams for front and back reflections of an optical wave from an acoustic wave.

READING LIST

Acousto-Optics

See also the reading list on crystals and tensor analysis in Chapter 6.

T.-C. Poon and T. Kim, *Engineering Optics with MATLAB*, World Scientific, 2nd ed. 2018, Chapter 4.

N. A. Riza, *Photonic Signals and Systems: An Introduction*, McGraw–Hill, 2013, Chapters 5 and 7.

P. A. Daymier, ed., *Acoustic Metamaterials and Phononic Crystals*, Springer-Verlag, 2013.

J. C. Joshi, *Acousto-Optic Devices and Their Defence Applications*, Defence Scientific Information & Documentation Centre (Delhi), 2007.

J. P. Wolfe, *Imaging Phonons: Acoustic Wave Propagation in Solids*, Cambridge University Press, 2005.

A. Yariv and P. Yeh, *Optical Waves in Crystals: Propagation and Control of Laser Radiation*, Wiley, 2003, Chapters 9 and 10.

M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 7th expanded and corrected ed. 2002, Chapter 12.

D. Royer and E. Dieulesaint, *Elastic Waves in Solids II: Generation, Acousto-Optic Interaction, Applications*, Springer-Verlag, 2000.

A. Korpel, *Acousto-Optics*, CRC Press, 2nd ed. 1997.

J. Xu and R. Stroud, *Acousto-Optic Devices: Principles, Design and Applications*, Wiley, 1992.

C. S. Tsai, *Guided-Wave Acoustooptics: Interactions, Devices, and Applications*, Springer-Verlag, 1990, reprinted 2011.

A. Korpel, ed., *Selected Papers on Acousto-Optics*, SPIE Optical Engineering Press (Milestone Series Volume 16), 1990.

M. Gottlieb, C. L. M. Ireland, and J. M. Ley, *Electro-Optic and Acousto-Optic Scanning and Deflection*, CRC Press, 1983.

T. S. Narasimhamurty, *Photoelastic and Electro-Optic Properties of Crystals*, Springer-Verlag/Plenum, 1981.

D. F. Nelson, *Electric, Optic, and Acoustic Interactions in Dielectrics*, Wiley, 1979.

M. V. Berry, *The Diffraction of Light by Ultrasound*, Academic Press, 1966.

L. Brillouin, Diffusion de la lumière et des rayons X par un corps transparent homogène: Influence de l'agitation thermique [Diffusion of Light and X-rays by a Transparent Homogeneous Body: The Influence of Thermal Excitations], *Annales de Physique* (*Paris*), vol. 9, no. 17, pp. 88–122, 1922.

W. L. Bragg, The Diffraction of X-Rays by Crystals (Nobel Lecture in Physics, 1915), in *Nobel Lectures in Physics, 1901–1921*, World Scientific, 1998.

PROBLEMS

20.1-1 **Diffraction of Light from Various Periodic Structures.** Discuss the diffraction of an optical plane wave of wavelength λ from the following periodic structures, indicating in each case the geometrical configuration and the frequency shift(s):

(a) An acoustic traveling wave of wavelength Λ.

(b) An acoustic standing wave of wavelength Λ.

(c) A graded-index transparent medium with refractive index that varies sinusoidally with position (period Λ).

(d) A stratified medium comprising parallel layers of two materials of different refractive indices that alternate to form a periodic structure of period Λ (Sec. 7.1C).

*20.1-2 **Bragg Diffraction as a Scattering Process.** An incident optical wave of angular frequency ω, wavevector $\mathbf{k}$, and complex envelope A interacts with a medium perturbed by an acoustic wave of angular frequency Ω and wavevector $\mathbf{q}$, and creates a light source $\mathcal{S}$ described by (20.1-32). The angle θ corresponds to upshifted Bragg diffraction, so that the scattering light source is $\mathcal{S} = \text{Re}\{S_r(\mathbf{r}) \exp(j\omega_r t)\}$, where $S_r(\mathbf{r}) = -(\Delta n_0/n)\, k_r^2\, A \exp(-j\mathbf{k}_r \cdot \mathbf{r})$, $\omega_r = \omega + \Omega$, and $\mathbf{k}_r = \mathbf{k} + \mathbf{q}$. This source emits a scattered field E. Assuming that the incident wave is undepleted by the acousto-optic interaction (first Born approximation, i.e., A remains approximately constant), the scattered light may be obtained by solving the Helmholtz equation $\nabla^2 E + k^2 E = -S$. This equation has the far-field solution (Prob. 22.2-6)

$$E(\mathbf{r}) \approx \frac{\exp(-jkr)}{4\pi r} \int_V S_r(\mathbf{r}') \exp(jk\hat{\mathbf{r}} \cdot \mathbf{r}')\, d\mathbf{r}',$$

where $\hat{\mathbf{r}}$ is a unit vector in the direction of $\mathbf{r}$, $k = 2\pi/\lambda$, and V is the volume of the source. Use this equation to determine an expression for the intensity reflectance of the acousto-optic cell when the Bragg condition is satisfied. Compare the result with (20.1-22).

20.1-3 **Condition for Raman–Nath Diffraction.** Derive an expression for the maximum width D_s of an acoustic beam of wavelength Λ that permits Raman–Nath diffraction of light of wavelength λ (refer to Fig. 20.1-10).

*20.1-4 **Combined Acousto-Optic and Electro-Optic Modulation.** One end of a lithium niobate ($LiNbO_3$) crystal is placed inside a microwave cavity that contains an electromagnetic field at a frequency of 3 GHz. As a result of the piezoelectric effect (an electric field creates a strain in the material), an acoustic wave is launched in the crystal, which has a refractive index of $n = 2.3$ and in which the velocity of sound is $v_s = 7.4$ km/s. Light from a He–Ne laser ($\lambda_o = 633$ nm) is reflected from the acoustic wave. Determine the Bragg angle. Since lithium niobate is also an electro-optic material, the applied electric field modulates the refractive index, which in turn modulates the phase of the incident light. Sketch the spectrum of the reflected light. If the microwave electric field is a pulse of short duration, sketch the spectrum of the reflected light at different times, indicating the contributions of the electro-optic and acousto-optic effects.

20.2-4 **Choice of Materials for Acousto-Optic Modulators.** Consider the following materials for possible use as Bragg reflectors: 1) CaF_2; 2) fused silica; 3) TiO_2; 4) $LiNbO_3$; 5) Si; and 6) Ge.

(a) For each material, determine the wavelength region of transparency (see Fig. 5.5-1); the acousto-optic figure of merit $\mathcal{M}$ specified in (20.1-7); and the amplitude of the refractive-index wave Δn_0 launched by an acoustic wave of intensity 10 W/cm^2. Compare your results with those for extra-dense flint glass provided in Example 20.1-1.

(b) If each of these Bragg cells is placed in air, determine the internal and external Bragg angles for the diffraction of an optical wave of free-space wavelength $\lambda_o = 2.5$ μm from a sound wave of frequency $f = 100$ MHz. Compare your results with those for extra-dense flint glass provided in Example 20.1-2.

(c) For light of wavelength $\lambda_o = 2.5$ μm, a sound intensity of $I_s = 100$ W/cm^2, and a penetration length of the light through the sound of $L/\sin\theta = 1$ mm, determine the intensity reflectance $\mathcal{R}_e$ for each material. Compare your results with those for extra-dense flint glass provided in Example 20.1-3.

(d) Which materials promise the best performance at $\lambda_o = 0.532$ μm? At 1.06 μm? At 10.6 μm?

20.2-5 **Frequency Shifting with a Bragg Reflector.** Devise a system for converting a monochromatic optical wave with complex wavefunction $U(t) = A \exp(j\omega t)$ into a modulated wave of complex wavefunction $A \cos(\Omega t) \exp(j\omega t)$ by making use of an acousto-optic cell with an acoustic wave $s(x,t) = S_0 \cos(\Omega t - qx)$. *Hint:* Consider the use of upshifted and downshifted Bragg reflections.

20.2-6 **Frequency-Shift-Free Bragg Reflector.** Design an acousto-optic system that deflects light without imparting a frequency shift. *Hint:* Use two Bragg cells.

*20.3-2 **Front Bragg Diffraction.** A transverse acoustic wave of wavelength Λ travels along the x direction in a uniaxial crystal whose refractive indices are n_o and n_e and whose optic axis in the z direction. Derive an expression for the wavelength λ_o of an incident optical

wave, traveling in the x direction and polarized in the z direction, that satisfies the Bragg-diffraction condition. What is the polarization of the front reflected wave? Determine Λ if $\lambda_o = 633\ \text{nm}$, $n_e = 2.200$, and $n_o = 2.286$.

CHAPTER
21

ELECTRO-OPTICS

21.1 PRINCIPLES OF ELECTRO-OPTICS 977
- A. Pockels and Kerr Effects
- B. Electro-Optic Modulators and Switches
- C. Scanners
- D. Directional Couplers
- E. Spatial Light Modulators

*21.2 ELECTRO-OPTICS OF ANISOTROPIC MEDIA 989
- A. Pockels and Kerr Effects
- B. Modulators

21.3 ELECTRO-OPTICS OF LIQUID CRYSTALS 996
- A. Wave Retarders and Modulators
- B. Spatial Light Modulators and Displays

*21.4 PHOTOREFRACTIVITY 1005

21.5 ELECTROABSORPTION 1010

Friedrich Pockels (1865–1913), a German physicist, described the linear electro-optic effect in 1884.

John Kerr (1824–1907), a Scottish physicist, discovered the quadratic electro-optic effect in 1875.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Much as with acousto-optics (Chapter 20), electro-optics is a branch of photonics that deals with the modulation, switching, deflection, scanning, and redirection of optical beams. However, in electro-optics attention is directed to the implementation of these operations by means of transparent materials whose optical properties are altered when subjected to an electric field, rather than to an acoustic wave. The electric field distorts the positions, orientations, and/or shapes of the molecules that constitute the material. The **electro-optic effect** represents a change in the *refractive index* of the material that results from the application of a steady or low-frequency electric field (Fig. 21.0-1). In particular, an electric field applied to an *anisotropic* optical material modifies its refractive indices and thereby the effect that the material has on polarized light passing through it.

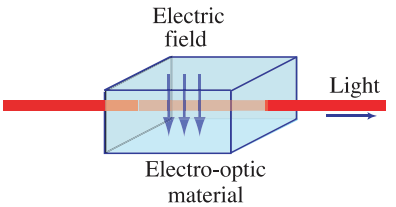


Figure 21.0-1 A steady electric field applied to an electro-optic material changes its refractive index. This in turn changes the effect of the material on light traveling through it. The electric field therefore controls the light.

The dependence of the refractive index on the applied electric field usually assumes one of two forms:

- The refractive index changes in proportion to the applied electric field, an effect known as the **linear electro-optic effect** or **Pockels effect**.
- The refractive index changes in proportion to the square of the applied electric field, an effect known as the **quadratic electro-optic effect** or **Kerr effect**.

The change in the refractive index is typically small. Nevertheless, the phase of an optical wave propagating through an electro-optic medium can be modified significantly if the distance of travel substantially exceeds the wavelength of the light. As an example, if the refractive index is increased by 10^{-5} by virtue of the presence of the electric field, an optical wave propagating a distance of 10^5 wavelengths will experience an additional phase shift of 2π.

Materials whose refractive index can be modified by means of an applied electric field are useful for producing electrically controllable optical devices, as indicated by the following examples:

- A lens made of a material whose refractive index can be varied is a lens of controllable focal length.
- A prism whose beam-bending capability is controllable can be used as an optical scanning device.
- Light transmitted through a transparent plate of controllable refractive index undergoes a controllable phase shift so that the plate can be used as an optical phase modulator.
- An anisotropic crystal whose refractive indices can be changed serves as a wave retarder of controllable retardation; it may be used to change the polarization properties of light.
- A wave retarder placed between two crossed polarizers results in transmitted light whose intensity is dependent on the phase retardation (Sec. 6.6B). The transmittance of such a device is therefore electrically controllable so that it can be used as an optical intensity modulator or an optical switch.

Controllable components such as these find substantial use in optical signal-processing and in optical fiber communications applications.

Alternatively, an electric field can be used to modify the optical properties of a material via *absorption*. A semiconductor material is normally optically transparent to light whose wavelength is longer than the bandgap wavelength (Sec. 17.2B). However, an applied electric field can reduce the bandgap of the material, thereby facilitating absorption and converting the material from transparent to opaque. This effect, known as **electroabsorption**, is useful for making optical modulators and switches.

This Chapter

We begin with a description of the electro-optic effect and the principles of electro-optic modulation and scanning. The initial presentation in Sec. 21.1 is simplified by deferring the detailed consideration of anisotropic effects to Sec. 21.2.

Section 21.3 is devoted to the electro-optic properties of liquid crystals. An electric field applied to the molecules of a liquid crystal causes them to alter their orientations. This in turn leads to changes in the optical properties of the medium, i.e., it exhibits an electro-optic effect. The molecules of a twisted nematic liquid crystal, for example, are organized in a helical pattern so that they normally act as a polarization rotator. An applied electric field can be used to remove this helical pattern, thereby deactivating the rotatory power of the material. Turning the electric field off results in the material regaining its original helical structure and therefore its rotatory power. Hence, the device acts as a dynamic polarization rotator. The use of fixed polarizers in conjunction with such a polarization rotator permits it to serve as an intensity modulator or switch. This behavior is the basis of most liquid-crystal display devices.

The electro-optic properties of photorefractive media are considered in Sec. 21.4. These are materials in which the absorption of light creates an internal electric field, which in turn initiates an electro-optic effect that alters the optical properties of the medium. Thus, the optical properties of the medium are indirectly controlled by the light incident on it. Photorefractive devices therefore permit *light to control light*. Finally, a brief introduction to electroabsorption is provided in Sec. 21.5.

21.1 PRINCIPLES OF ELECTRO-OPTICS

A. Pockels and Kerr Effects

The refractive index of an electro-optic medium is a function $n(E)$ of an applied electric field E that is steady (or slowly varying in comparison with optical frequencies). The function $n(E)$ varies only slightly with E so that it can be expanded in a Taylor series about $E = 0$,

$$n(E) \approx n + a_1 E + \tfrac{1}{2} a_2 E^2 + \cdots, \tag{21.1-1}$$

where the coefficients of expansion are $n = n(0)$, $a_1 = (dn/dE)|_{E=0}$, and $a_2 = (d^2n/dE^2)|_{E=0}$. For reasons that will become apparent below, it is conventional to write (21.1-1) in terms of two new coefficients, $\mathrm{r} = -2a_1/n^3$ and $\mathrm{s} = -a_2/n^3$, known as the electro-optic coefficients, so that

$$n(E) \approx n - \tfrac{1}{2}\mathrm{r}\, n^3 E - \tfrac{1}{2}\mathrm{s}\, n^3 E^2 + \cdots. \tag{21.1-2}$$

The second- and higher-order terms of this series are typically many orders of magnitude smaller than n. Terms higher than the third can safely be neglected.

For future use in connection with the optical properties of anisotropic media (Sec. 6.3A), it is convenient to derive an expression for the electric impermeability of the electro-optic medium, $\eta = \epsilon_o/\epsilon = 1/n^2$, as a function of E. The incremental

change $\Delta\eta \approx (d\eta/dn)\Delta n = (-2/n^3)(-\frac{1}{2}\mathrm{r}\,n^3E - \frac{1}{2}\mathrm{s}\,n^3E^2) = \mathrm{r}\,E + \mathrm{s}\,E^2$, so that

$$\eta(E) \approx \eta + \mathrm{r}\,E + \mathrm{s}\,E^2, \tag{21.1-3}$$

where $\eta = \eta(0)$. The electro-optic coefficients r and s are therefore simply the coefficients of proportionality of the two terms of $\Delta\eta$ with E and E^2, respectively. This explains the seemingly odd definitions of r and s used in (21.1-2). The values of the coefficients r and s depend on the direction of the applied electric field and the polarization of the light, as will be discussed in Sec. 21.2.

Pockels Effect

There is a large class of materials for which the third term of (21.1-2) is negligible in comparison with the second, whereupon

$$\boxed{n(E) \approx n - \tfrac{1}{2}\mathrm{r}\,n^3E,} \tag{21.1-4}$$
Pockels Effect

as illustrated in Fig. 21.1-1(a). The medium is then known as a Pockels medium (or a Pockels cell) and the coefficient r is called the **Pockels coefficient** or the linear electro-optic coefficient. The change in refractive index induced by the electric field in the Pockel's effect [(21.1-4)] is analogous to that induced by strain in the acousto-optic effect [(20.1-3)]. Typical values of r lie in the range 10^{-12} to 10^{-10} m/V (or 1 to 100 pm/V). For $E = 10^6$ V/m (e.g., 10 kV applied across a 1-cm-thick cell), for example, the term $\frac{1}{2}\mathrm{r}\,n^3E$ in (21.1-4) is on the order of 10^{-6} to 10^{-4}. Changes in the refractive index induced by electric fields are indeed very small. Crystals commonly used as Pockels cells include $KTiOPO_4$ (KTP), β-BaB_2O_4 (BBO), KH_2PO_4 (KDP), $NH_4H_2PO_4$ (ADP), $LiNbO_3$, and $LiTaO_3$. Specially designed polymers can also serve this purpose.

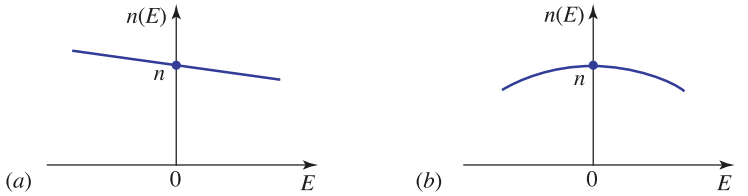


Figure 21.1-1 Dependence of the refractive index on the electric field: (a) Pockels medium; (b) Kerr medium.

Kerr Effect

If the material is centrosymmetric, as is the case for gases, liquids, and certain crystals, $n(E)$ must be an even symmetric function [Fig. 21.1-1(b)] since it must be invariant to the reversal of E. The linear term then vanishes when $\mathrm{r} = 0$, whereupon

$$\boxed{n(E) \approx n - \tfrac{1}{2}\mathrm{s}\,n^3E^2.} \tag{21.1-5}$$
Kerr Effect

The material is then known as a Kerr medium (or a Kerr cell) and the parameter s is called the **Kerr coefficient** or the quadratic electro-optic coefficient. Typical values of s lie between 10^{-18} and 10^{-14} m^2/V^2 in crystals, and between 10^{-22} and 10^{-19} m^2/V^2 in liquids. For $E = 10^6$ V/m, for example, the term $\frac{1}{2}\mathrm{s}\,n^3E^2$ in (21.1-5) is on the order of 10^{-6} to 10^{-2} in crystals and 10^{-10} to 10^{-7} in liquids.

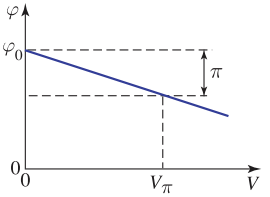
B. Electro-Optic Modulators and Switches

Phase Modulators

A beam of light traversing a Pockels cell of length L to which an electric field E is applied undergoes a phase shift $\varphi = n(E)\, k_o L = 2\pi n(E) L/\lambda_o$, where λ_o is the free-space wavelength. Using (21.1-4), we therefore have

$$\varphi \approx \varphi_0 - \pi \frac{\mathrm{r}\, n^3 E L}{\lambda_o}, \tag{21.1-6}$$

where $\varphi_0 = 2\pi n L/\lambda_o$. If the electric field is obtained by applying a voltage V across two faces of the cell separated by a distance d, then $E = V/d$ and (21.1-6) provides



$$\boxed{\varphi = \varphi_0 - \pi \frac{V}{V_\pi},} \tag{21.1-7}$$

Phase Modulation

where

$$\boxed{V_\pi = \frac{d}{L} \frac{\lambda_o}{\mathrm{r}\, n^3}.} \tag{21.1-8}$$

Half-Wave Voltage

The parameter V_π, known as the **half-wave voltage**, is the applied voltage at which the phase shift changes by π.

Equation (21.1-7) expresses a linear relation between the optical phase shift and the voltage. One can therefore modulate the phase of an optical wave by varying the voltage V applied across a material through which the light passes. The parameter V_π is an important characteristic of the modulator. It depends on the material properties (n and r), on the wavelength λ_o, and on the aspect ratio d/L. The value of the electro-optic coefficient r depends on the directions of propagation and the applied field since the crystal is, in general, anisotropic, as explained in Sec. 21.2. As illustrated in Fig. 21.1-2, the electric field may be applied in a direction parallel to the direction of light propagation (longitudinal modulator), in which case $d = L$, or perpendicular thereto (transverse modulator).

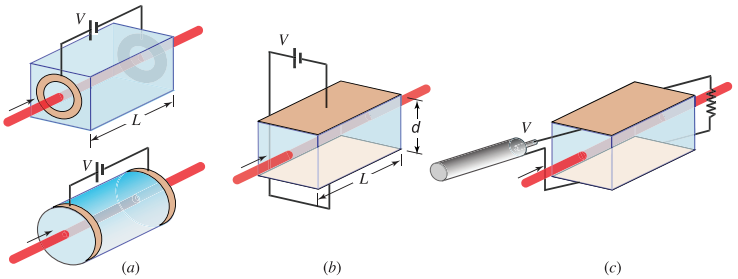


Figure 21.1-2 (a) Longitudinal modulator. The electrodes may take the shape of washers or bands, or may be transparent conductors. (b) Transverse modulator. (c) Traveling-wave transverse modulator. Typical values of the half-wave voltage are in the vicinity of 1 to a few kilovolts for longitudinal modulators, and hundreds of volts for transverse modulators.

The speed at which an electro-optic modulator operates is limited by electrical capacitive effects and by the transit time of the light through the material. If the slowly varying electric field $E(t)$ varies significantly within the light transit time T, the traveling optical wave will be subjected to different electric fields as it traverses the crystal. The modulated phase at a given time t will then be proportional to the average electric field $E(t)$ at times from $t - T$ to t. As a result, the transit-time-limited modulation bandwidth is $\approx 1/T$. One method of reducing this time is to apply the voltage V at one end of the crystal while the electrodes serve as a transmission line, as illustrated in Fig. 21.1-2(c). If the velocity of the traveling electrical wave matches that of the optical wave, transit time effects can, in principle, be eliminated. Commercial modulators in forms such as those shown in Fig. 21.1-2 generally operate up to the GHz domain.

As illustrated in Fig. 21.1-3, electro-optic modulators can also be constructed in the form of integrated-photonic devices, which operate at higher speeds and at lower voltages than bulk devices. An optical waveguide is fabricated on an electro-optic substrate (often $LiNbO_3$) by indiffusing a material (such as Ti) to increase the refractive index, and the electric field is applied to the waveguide using electrodes. Because the configuration is transverse, and the width of the waveguide is much smaller than its length ($d \ll L$), the half-wave voltage can be as small as several volts. Modulators such as these can be operated at speeds in excess of 100 GHz. Light can be conveniently coupled into, and out of, such devices by the use of optical fibers.

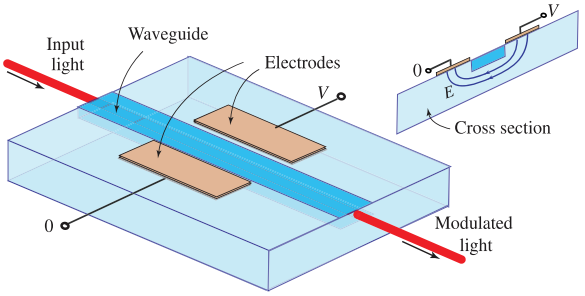


Figure 21.1-3 An integrated-photonic phase modulator using the electro-optic effect.

Dynamic Wave Retarders

An anisotropic medium has two linearly polarized normal modes that propagate with different velocities, say c_o/n_1 and c_o/n_2 (Sec. 6.3B). If the medium exhibits the Pockels effect, then in the presence of a steady electric field E the two refractive indices are modified in accordance with (21.1-4), so that

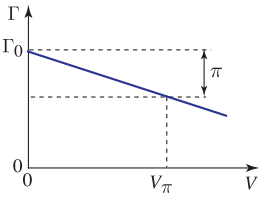
$$n_1(E) \approx n_1 - \tfrac{1}{2}\mathrm{r}_1 n_1^3 E \tag{21.1-9}$$

$$n_2(E) \approx n_2 - \tfrac{1}{2}\mathrm{r}_2 n_2^3 E, \tag{21.1-10}$$

where r_1 and r_2 are the appropriate Pockels coefficients (anisotropic effects are examined in detail in Sec. 21.2). After propagating a distance L, the two modes thus undergo a relative phase retardation given by

$$\Gamma = k_o[n_1(E) - n_2(E)]L = k_o(n_1 - n_2)L - \tfrac{1}{2}k_o(\mathrm{r}_1 n_1^3 - \mathrm{r}_2 n_2^3)EL. \tag{21.1-11}$$

If E results from the application of a voltage V across two surfaces of the medium separated by a distance d, (21.1-11) can be written in compact form as



$$\Gamma = \Gamma_0 - \pi \frac{V}{V_\pi}, \tag{21.1-12}$$

Phase Retardation

where $\Gamma_0 = k_o(n_1 - n_2)L$ is the phase retardation in the absence of the electric field. The applied voltage V_π necessary to obtain a phase retardation π is then given by

$$V_\pi = \frac{d}{L} \frac{\lambda_o}{\mathrm{r}_1 n_1^3 - \mathrm{r}_2 n_2^3}. \tag{21.1-13}$$

Retardation Half-Wave Voltage

Equation (21.1-12) indicates that the phase retardation is linearly related to the applied voltage so that the medium behaves as an electrically controllable dynamic wave retarder.

Intensity Modulators: Use of a Phase Modulator in an Interferometer

Phase delay (or retardation) alone does not affect the intensity of a light beam. However, a phase modulator placed in one branch of an interferometer can function as an intensity modulator. Consider, for example, the Mach–Zehnder interferometer portrayed in Fig. 21.1-4. If the beamsplitters divide the optical power equally, the intensity transmitted through one output port of the interferometer I_o is related to the incident intensity I_i by

$$I_o = \tfrac{1}{2} I_i + \tfrac{1}{2} I_i \cos\varphi = I_i \cos^2(\varphi/2), \tag{21.1-14}$$

where $\varphi = \varphi_1 - \varphi_2$ is the difference between the phase shifts encountered by the light as it travels through the two interferometer branches (Sec. 2.5A). The transmittance of the interferometer is thus $\mathcal{T} = I_o/I_i = \cos^2(\varphi/2)$.

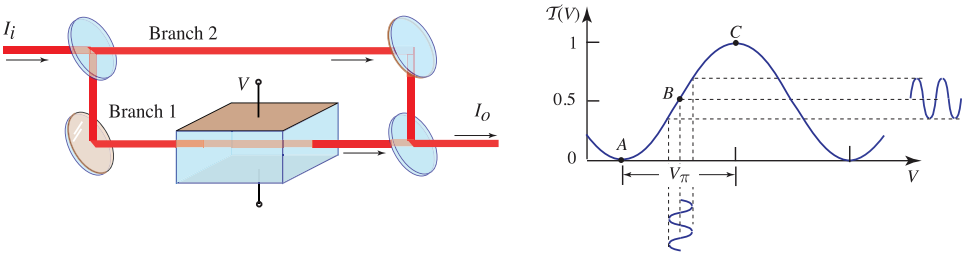


Figure 21.1-4 A phase modulator placed in one branch of a Mach–Zehnder interferometer can serve as an intensity modulator. The transmittance of the interferometer $\mathcal{T}(V) = I_o/I_i$ varies periodically with the applied voltage V. By operating over a limited region of voltage near point B, the device acts as a linear intensity modulator. If V is switched between points A and C, the device serves as an optical switch.

Because of the presence of the phase modulator in branch 1, in accordance with (21.1-7) we have $\varphi_1 = \varphi_{1_0} - \pi V/V_\pi$, so that φ is controlled by the applied voltage V in accordance with the linear relation $\varphi = \varphi_1 - \varphi_2 = \varphi_0 - \pi V/V_\pi$, where the constant

$\varphi_0 = \varphi_{1_0} - \varphi_2$ depends on the optical path difference. The transmittance of the device is therefore a function of the applied voltage V,

$$\mathcal{T}(V) = \cos^2\left(\frac{\varphi_0}{2} - \frac{\pi}{2}\frac{V}{V_\pi}\right). \tag{21.1-15}$$

Transmittance

This function is plotted in Fig. 21.1-4 for an arbitrary value of φ_0. The device may be operated as a linear intensity modulator by adjusting the optical path difference so that $\varphi_0 = \pi/2$ and operating in the nearly linear region around $\mathcal{T} = 0.5$. Alternatively, the optical path difference may be adjusted so that φ_0 is a multiple of 2π. In that case $\mathcal{T}(0) = 1$ and $\mathcal{T}(V_\pi) = 0$, so the modulator switches the light on and off as V is switched between 0 and V_π.

A Mach–Zehnder intensity modulator may also be constructed in the form of an integrated-photonic device. Waveguides are placed on a substrate in the geometry shown in Fig. 21.1-5. The beamsplitters are implemented using waveguide Y's while the optical input and output are carried on optical fibers. Commercially available integrated-photonic modulators generally operate at speeds of tens of GHz.

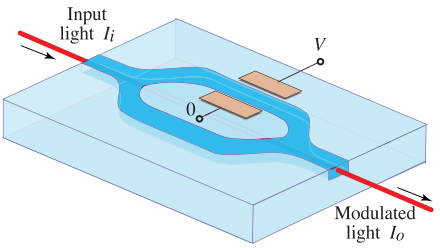


Figure 21.1-5 An integrated-photonic intensity modulator (or optical switch). A Mach–Zehnder interferometer and an electro-optic phase modulator are implemented using optical waveguides fabricated from a material such as $LiNbO_3$ indiffused with Ti.

Intensity Modulators: Use of a Retarder Between Crossed Polarizers

As described in Sec. 6.6B, a wave retarder (of retardation Γ) sandwiched between two crossed polarizers placed at $45°$ with respect to the retarder's axes (Fig. 6.6-4) exhibits an intensity transmittance $\mathcal{T} = \sin^2(\Gamma/2)$. If the retarder is a Pockels cell, then Γ is linearly dependent on the applied voltage V, as provided in (21.1-12). The transmittance of the device is then a periodic function of V,

$$\mathcal{T}(V) = \sin^2\left(\frac{\Gamma_0}{2} - \frac{\pi}{2}\frac{V}{V_\pi}\right), \tag{21.1-16}$$

Transmittance

as displayed in Fig. 21.1-6. By changing V, the transmittance can be varied between 0 (shutter closed) and 1 (shutter open). The device can also be used as a linear modulator if the system is operated in the region near $\mathcal{T}(V) = 0.5$. By selecting $\Gamma_0 = \pi/2$ and $V \ll V_\pi$, we have

$$\mathcal{T}(V) = \sin^2\left(\frac{\pi}{4} - \frac{\pi}{2}\frac{V}{V_\pi}\right) \approx \mathcal{T}(0) + V\left.\frac{d\mathcal{T}}{dV}\right|_{V=0} = \frac{1}{2} - \frac{\pi}{2}\frac{V}{V_\pi}, \tag{21.1-17}$$

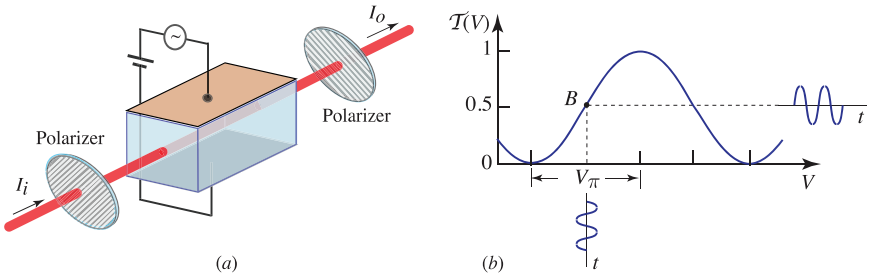


Figure 21.1-6 (a) An optical intensity modulator using a Pockels cell placed between two crossed polarizers. (b) Optical transmittance versus applied voltage for an arbitrary value of Γ_0; for linear operation the cell is biased near point B.

so that $\mathcal{T}(V)$ is a linear function with slope $\pi/2V_\pi$, representing the sensitivity of the modulator. The phase retardation Γ_0 can be adjusted either optically (by incorporating an external phase retarder) or electrically (by incorporating a constant bias voltage).

In practice, the maximum transmittance of the modulator is smaller than unity because of losses associated with reflection, absorption, and scattering. Moreover, the minimum transmittance is greater than zero because of inadvertent misalignments of the propagation direction and the polarization directions relative to the crystal axes and polarizers. Nevertheless, the ratio between the maximum and minimum transmittances, called the **extinction ratio**, can exceed 30 dB (1000:1).

C. Scanners

An optical beam can be deflected dynamically by using a prism with an electrically controlled refractive index. In accordance with (1.2-7), the angle of deflection introduced by a prism of small apex angle α and refractive index n is $\theta \approx (n-1)\alpha$. An incremental change of the refractive index Δn caused by an applied electric field E via the Pockels effect results in an incremental change of the deflection angle,

$$\Delta\theta \approx \alpha\Delta n = -\tfrac{1}{2}\alpha r\, n^3 E = -\tfrac{1}{2}\alpha r\, n^3 V/d, \tag{21.1-18}$$

where V is the applied voltage and d is the prism width [Fig. 21.1-7(a)]. The angle $\Delta\theta$ is proportional to the applied voltage V so that the incident light is scanned. It is sometimes more convenient to place triangularly shaped electrodes that define a prism on a rectangular crystal. Two, or several, prisms can be cascaded by alternating the direction of the electric field, as illustrated in Fig. 21.1-7(b).

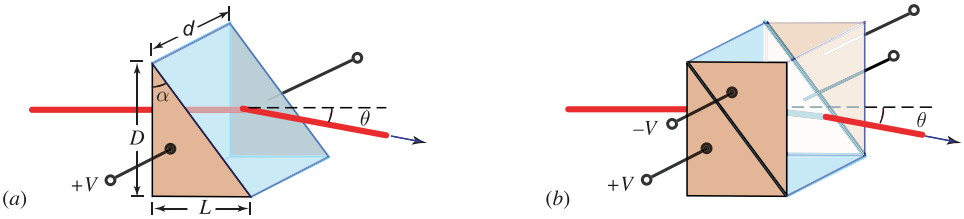


Figure 21.1-7 (a) An electro-optic prism. The deflection angle θ is controlled by the applied voltage. (b) An electro-optic double prism.

An important parameter that characterizes a scanner is its resolution, i.e., the number of independent spots it can scan and it turns out that electro-optic scanning is inefficient

in this respect. An optical beam of width D and wavelength λ_o has an angular divergence $\delta\theta \approx \lambda_o/D$ [see (4.3-7)]. To minimize that angle, the beam should be as wide as possible, ideally covering the entire width of the prism itself. For a given maximum voltage V corresponding to a scanned angle $\Delta\theta$, the number of independent spots is given by

$$N \approx \frac{|\Delta\theta|}{\delta\theta} = \frac{\frac{1}{2}\alpha \mathrm{r}\, n^3 V/d}{\lambda_o/D}. \tag{21.1-19}$$

Substituting $\alpha \approx L/D$ and $V_\pi = (d/L)(\lambda_o/\mathrm{r}\, n^3)$ then leads to

$$N \approx \frac{V}{2V_\pi}, \tag{21.1-20}$$

from which we conclude that $V \approx 2NV_\pi$. This is a discouraging result because it indicates that the scanning of N independent spots requires a voltage $2N$ times greater than the half-wave voltage V_π, which is generally large to begin with. As a consequence, acousto-optic and mechanical scanners (Secs. 20.2B and 24.3B, respectively) are of greater use than electro-optic scanners.

Lateral beam shift, rather than deflection, can be effected by making use of the process of double refraction in an anisotropic crystal (Sec. 6.3E). An incident beam is shifted parallel to itself for one polarization while undergoing no shift for the orthogonal polarization. This process is implemented by first passing a linearly polarized optical beam through an electro-optic wave retarder that acts as a polarization rotator and then passing it through the birefringent crystal, as illustrated in Fig. 21.1-8. The electrically controlled polarization determines whether the beam is, or is not, shifted laterally.

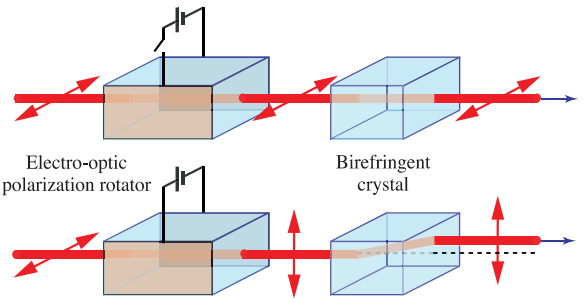


Figure 21.1-8 A position switch based on electro-optic phase retardation and double refraction.

D. Directional Couplers

An important application of the electro-optic effect is in controlling the coupling between two parallel waveguides in an integrated-photonic device. An electric field can be used to transfer light from one waveguide to the other, so that the device serves as an electrically controlled directional coupler.

The coupling of light between two parallel single-mode planar waveguides, as displayed in Fig. 21.1-9(a), was examined in Sec. 9.4B. It was shown there that the optical powers carried by the two waveguides, $P_1(z)$ and $P_2(z)$, are periodically exchanged along the direction of propagation z. Two parameters govern the strength of the coupling process: the coupling coefficient $\mathcal{C}$ (which depends on the dimensions,

wavelength, and refractive indices), and the mismatch of the propagation constants $\Delta\beta = \beta_1 - \beta_2 = 2\pi\Delta n/\lambda_o$, where Δn is the difference between the refractive indices of the waveguides. If $P_2(0) = 0$ and the waveguides are identical with $\Delta\beta = 0$, then at a distance $z = L_0 = \pi/2\mathcal{C}$, known as the **transfer distance** or **coupling length**, the power is totally transferred from waveguide 1 into waveguide 2, i.e., $P_1(L_0) = 0$ and $P_2(L_0) = P_1(0)$, as illustrated in Fig. 21.1-9(a).

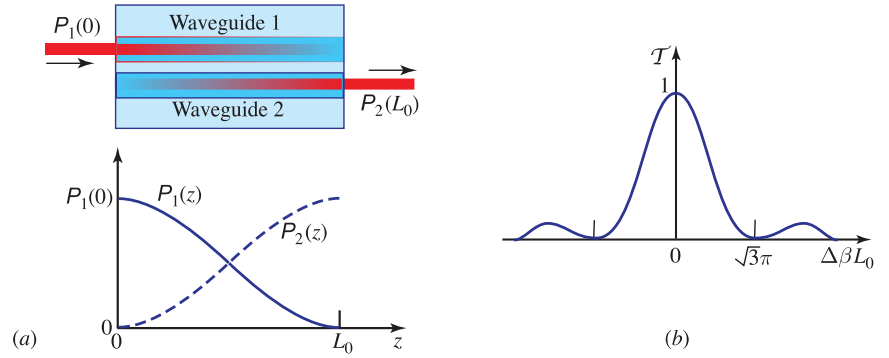


Figure 21.1-9 (a) Exchange of power between two parallel, weakly coupled, identical waveguides with the same propagation constant β ($\Delta\beta = 0$). At $z = 0$ all of the power resides in waveguide 1; at $z = L_0$ all of the power is transferred to waveguide 2. (b) Dependence of the power-transfer ratio $\mathcal{T} = P_2(L_0)/P_1(0)$ on the phase-mismatch parameter $\Delta\beta\, L_0$ for $\Delta\beta \neq 0$.

For waveguides of lengths L_0 with different propagation constants ($\Delta\beta \neq 0$), the power-transfer ratio $\mathcal{T} = P_2(L_0)/P_1(0)$ is a function of the phase mismatch, as provided in (9.4-13),

$$\mathcal{T} = \frac{\pi^2}{4}\,\text{sinc}^2\left[\frac{1}{2}\sqrt{1+\left(\frac{\Delta\beta\, L_0}{\pi}\right)^2}\right], \tag{21.1-21}$$

where $\text{sinc}(x) \equiv \sin(\pi x)/(\pi x)$. Figure 21.1-9($b$) illustrates this dependence. The ratio assumes a maximum value of unity at $\Delta\beta\, L_0 = 0$, decreases with increasing $\Delta\beta\, L_0$, and vanishes when $\Delta\beta\, L_0 = \sqrt{3}\,\pi$, at which point no optical power is transferred to waveguide 2.

The dependence of the coupled power on the phase mismatch is the key to fabricating electrically activated directional couplers. If the mismatch $\Delta\beta\, L_0$ is switched from 0 to $\sqrt{3}\pi$, the power-transfer ratio switches from unity to zero. Electrical control of $\Delta\beta$ is readily achieved by making use of the Pockels electro-optic effect. An electric field E applied to one of two, otherwise identical, waveguides alters its refractive index by $\Delta n = -\frac{1}{2}n^3\mathrm{r}\,E$, where r is the Pockels coefficient. This results in a phase shift $\Delta\beta\, L_0 = \Delta n(2\pi L_0/\lambda_o) = -(\pi/\lambda_o)n^3\mathrm{r}\,L_0 E$.

A typical electro-optic directional coupler has the geometry displayed in Fig. 21.1-10. The electrodes are placed over two waveguides separated by a distance d. An applied voltage V creates an electric field $E \approx V/d$ in one waveguide and $-V/d$ in the other, where d is an effective distance determined by solving the electrostatics problem (the electric-field lines go downward at one waveguide and upward at the other, as portrayed in the inset of Fig. 21.1-3). The refractive index is thus incremented in one waveguide and decremented in the other. The result is a net refractive index difference $2\Delta n = -n^3\mathrm{r}\,(V/d)$, corresponding to a phase mismatch $\Delta\beta\, L_0 = -(2\pi/\lambda_o)n^3\mathrm{r}\,(L_0/d)V$, which is proportional to the applied voltage V.

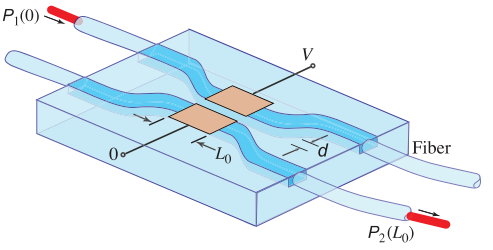


Figure 21.1-10 An integrated electro-optic directional coupler.

The voltage V_0 required to switch the optical power from one waveguide to the other is that for which $|\Delta\beta\, L_0| = \sqrt{3}\,\pi$, namely

$$V_0 = \sqrt{3}\frac{d}{L_0}\frac{\lambda_o}{2n^3\mathrm{r}} = \frac{\sqrt{3}}{\pi}\frac{\mathcal{C}\lambda_o d}{n^3\mathrm{r}}, \tag{21.1-22}$$

where $L_0 = \pi/2\mathcal{C}$ and $\mathcal{C}$ is the coupling coefficient. This is called the **switching voltage**. Since $|\Delta\beta\, L_0| = \sqrt{3}\,\pi\, V/V_0$, (21.1-21) yields

$$\mathcal{T} = \frac{\pi^2}{4}\,\mathrm{sinc}^2\left[\frac{1}{2}\sqrt{1+3\left(\frac{V}{V_0}\right)^2}\right]. \tag{21.1-23}$$

Coupling Efficiency

This equation, plotted in Fig. 21.1-11, governs the coupling of optical power in the directional coupler as a function of the applied voltage V.

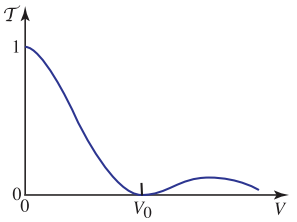


Figure 21.1-11 Dependence of the coupling efficiency of the directional coupler on the applied voltage V. When $V = 0$, all of the optical power is coupled from waveguide 1 into waveguide 2; when $V = V_0$, all of the optical power remains in waveguide 1.

An electro-optic directional coupler is characterized by its coupling length L_0, which is inversely proportional to the coupling coefficient $\mathcal{C}$, and its switching voltage V_0, which is directly proportional to $\mathcal{C}$. The key parameter is thus $\mathcal{C}$, which is governed by the geometry and refractive indices of the device (Sec. 9.4B). An integrated-photonic directional coupler may be fabricated, for example, by indiffusing titanium into a high-purity $LiNbO_3$ substrate. The light beams are focused to spot sizes of a few μm and the waveguide ends can be permanently fixed to single-mode polarization-maintaining optical fibers (Sec. 10.2B). The switching voltage V_0 is typically less than 10 V, and operating speeds reach tens of GHz. Increased bandwidths may be attained by making use of traveling-wave versions of such devices.

EXERCISE 21.1-1

Coupling-Efficiency Spectral Response. Equation (21.1-22) indicates that the switching voltage V_0 is proportional to the wavelength. Assume that the applied voltage $V = V_0$ for a particular value of the wavelength $\overline{\lambda}_o$, so that the coupling efficiency $\mathcal{T} = 0$ at $\overline{\lambda}_o$. If, instead, the incident wave has wavelength λ_o, plot the coupling efficiency $\mathcal{T}$ as a function of $(\lambda_o - \overline{\lambda}_o)/\overline{\lambda}_o$. Assume that the coupling coefficient $\mathcal{C}$ and the material parameters n and r are independent of wavelength.

E. Spatial Light Modulators

A spatial light modulator (SLM) is a device that modulates the intensity of light at different positions by prescribed factors (Fig. 21.1-12). It is a planar optical element of controllable intensity transmittance $\mathcal{T}(x, y)$ such that the transmitted light intensity $I_o(x, y)$ is related to the incident light intensity $I_i(x, y)$ by the product $I_o(x, y) = I_i(x, y)\mathcal{T}(x, y)$. If the incident light is uniform (i.e., if $I_i(x, y)$ is constant), the transmitted light intensity is proportional to $\mathcal{T}(x, y)$. The "image" $\mathcal{T}(x, y)$ is then imparted to the transmitted light, much as the image stored in a transparency is "read" by uniformly illuminating it in a slide projector. In a spatial light modulator, however, $\mathcal{T}(x, y)$ is controllable. In an electro-optic spatial light modulator the control is electrical.

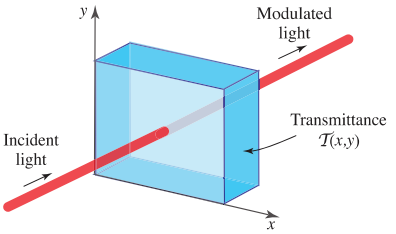


Figure 21.1-12 The spatial light modulator.

To construct a spatial light modulator using the electro-optic effect, some mechanism must be devised for creating an electric field $E(x, y)$ that is proportional to the desired transmittance $\mathcal{T}(x, y)$ at each position. This is not easy. One approach is to place an array of transparent electrodes on small plates of electro-optic material set between crossed polarizers and to apply to each electrode an appropriate voltage (Fig. 21.1-13). The voltage applied to the electrode centered at position (x_i, y_i), $i = 1, 2, \ldots$, is made proportional to the desired value of $\mathcal{T}(x_i, y_i)$ (see Fig. 21.1-6). If the number of electrodes is sufficiently large, the transmittance approximates $\mathcal{T}(x, y)$. This system is in effect a parallel array of longitudinal electro-optic modulators operated as intensity modulators. Though it is not practical to address a large number of such electrodes independently, it will become clear in Sec. 21.3B that this scheme is in fact practical for liquid-crystal spatial light modulators used for display, since the required voltages are low.

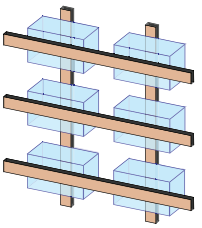


Figure 21.1-13 An electrically addressable array of longitudinal electro-optic modulators.

Optically Addressed Electro-Optic Spatial Light Modulators

One method of optically addressing an electro-optic spatial light modulator is based on the use of a thin layer of photoconductive material to create the electric field required to operate the modulator (Fig. 21.1-14). The conductivity of a photoconductive material is proportional to the intensity of light to which it is exposed (Sec. 19.2). When illuminated by a write image with an intensity distribution $I_W(x, y)$, a spatial pattern of conductance $G(x, y) \propto I_W(x, y)$ is created. The photoconductive layer is placed between two electrodes that act as a capacitor. The capacitor is initially charged and the electrical charge leakage at position (x, y) is proportional to the local conductance $G(x, y)$. Hence, the charge on the capacitor is reduced in those regions where the conductance is high. The local voltage is therefore proportional to $1/G(x, y)$ and the corresponding electric field $E(x, y) \propto 1/G(x, y) \propto 1/I_W(x, y)$. If the transmittance $\mathcal{T}(x, y)$ [or the reflectance $\mathcal{R}(x, y)$] of the modulator is proportional to the applied field, it must be inversely proportional to the initial light intensity $I_W(x, y)$.

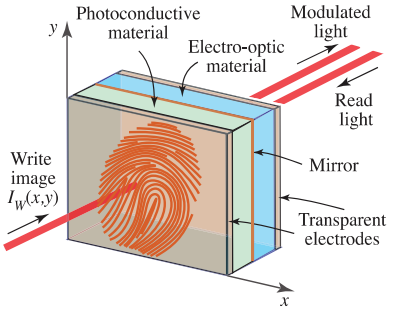


Figure 21.1-14 An optically addressed electro-optic spatial light modulator uses a photoconductive material to create a spatial distribution of electric field that controls an electro-optic material.

The Pockels Readout Optical Modulator

An ingenious implementation of this principle is the **Pockels readout optical modulator** (**PROM**). One implementation makes use of a crystal of bismuth silicon oxide, $Bi_{12}SiO_{20}$ (BSO), which has an unusual combination of optical and electrical properties: (1) it exhibits the Pockels electro-optic effect; (2) it is photoconductive for blue, but not for red light; and (3) it is a good insulator in the dark. The PROM displayed in Fig. 21.1-15 comprises a thin wafer of BSO sandwiched between two transparent electrodes. The light to be modulated (read light) is transmitted through a polarizer, enters the BSO layer, and is reflected by a dichroic reflector, whereupon it crosses a second polarizer. The reflector reflects red but is transparent to blue light. The PROM is operated as follows:

- *Priming:* A large potential difference ($\approx$ 4 kV) is applied to the electrodes and the capacitor is charged (with no leakage since the crystal is a good insulator in the dark).
- *Writing:* Intense blue write light with an intensity distribution $I_W(x, y)$ illuminates the crystal. As a result, a spatial pattern of conductance $G(x, y) \propto I_W(x, y)$ is created, the voltage across the crystal is selectively lowered, and the electric field decreases proportionally at each position, so that $E(x, y) \propto 1/G(x, y) \propto 1/I_W(x, y)$. As a result of the electro-optic effect, the refractive index of the BSO is altered, and a spatial pattern of refractive-index change $\Delta n(x, y) \propto 1/I_W(x, y)$ is created and stored in the crystal.
- *Reading:* Uniform red light is used to read $\Delta n(x, y)$ as with usual electro-optic intensity modulators [see Fig. 21.1-6(a)] with the polarizing beamsplitter playing the role of the crossed polarizers.
- *Erasing:* The refractive-index pattern is erased by illumination with a uniform flash of blue light. The crystal is again primed by applying 4 kV, and the device is ready for a new cycle.

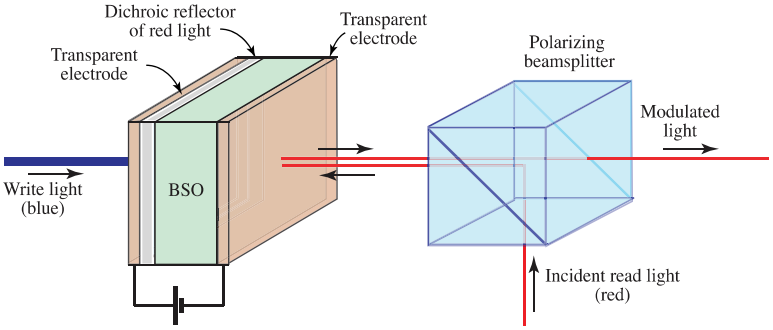


Figure 21.1-15 The Pockels readout optical modulator (PROM).

Incoherent-to-Coherent Optical Converters

In an optically addressed spatial light modulator, such as the PROM, the light used to write a spatial pattern into the modulator need not be coherent since photoconductive materials are sensitive to optical intensity. A spatial optical pattern (an image) may therefore be written using incoherent light, and read using coherent light. This process of real-time conversion of a spatial distribution of natural incoherent light into a proportional spatial distribution of coherent light is useful in a number of optical data- and image-processing applications.

*21.2 ELECTRO-OPTICS OF ANISOTROPIC MEDIA

The basic principles and applications of electro-optics were presented in a simplified fashion in Sec. 21.1; polarization and anisotropic effects were either ignored or introduced only generically. In this section a more complete analysis of the electro-optics of anisotropic media is presented. A brief refresher of some of the important properties of anisotropic media set forth in Sec. 6.3 is provided below.

Crystal Optics: A Brief Refresher

The optical properties of an anisotropic medium are characterized by a geometric construction called the **index ellipsoid**,

$$\sum_{ij} \eta_{ij}\, x_i x_j = 1, \qquad i,j = 1,2,3,$$

where $\eta_{ij} = \eta_{ji}$ are elements of the impermeability tensor $\boldsymbol{\eta} = \epsilon_o \boldsymbol{\epsilon}^{-1}$. If the axes of the ellipsoid correspond to the principal axes of the medium, its dimensions along these axes are the principal refractive indices n_1, n_2, and n_3 (Fig. 21.2-1):

$$x_1^2/n_1^2 + x_2^2/n_2^2 + x_3^2/n_3^2 = 1.$$

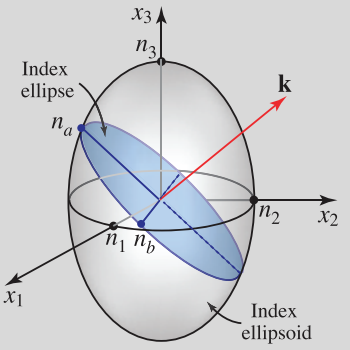


Figure 21.2-1 The index ellipsoid. The coordinates (x_1, x_2, x_3) are the principal axes and n_1, n_2, n_3 are the principal refractive indices. The refractive indices of the normal modes of a wave traveling in the direction **k** are n_a and n_b.

The index ellipsoid may be used to determine the polarizations and refractive indices n_a and n_b of the two normal modes of a wave traveling in an arbitrary direction in the anisotropic medium. This is accomplished by drawing a plane perpendicular to the direction of propagation that passes through the center of the ellipsoid. Its intersection with the ellipsoid is an ellipse whose major and minor axes have half-lengths equal to n_a and n_b, as described in Sec. 6.3C.

A. Pockels and Kerr Effects

When a steady electric field $\mathbf{E}$ with components (E_1, E_2, E_3) is applied to a crystal, the elements of the impermeability tensor $\boldsymbol{\eta}$ are altered. Each of the nine elements η_{ij} becomes a function of E_1, E_2, and E_3, i.e., $\eta_{ij} = \eta_{ij}(\mathbf{E})$, so that the index ellipsoid is modified (Fig. 21.2-2). Once we know the functions $\eta_{ij}(\mathbf{E})$, we can determine the index ellipsoid and the optical properties for an arbitrary applied electric field $\mathbf{E}$. The problem is simple in principle, but the implementation is often lengthy.

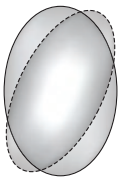


Figure 21.2-2 The index ellipsoid of a crystal is modified when a steady electric field is applied.

Each of the elements $\eta_{ij}(\mathbf{E})$ is a function of the three variables $\mathbf{E} = (E_1, E_2, E_3)$, which may be expanded in a Taylor series about $\mathbf{E} = \mathbf{0}$,

$$\eta_{ij}(\mathbf{E}) = \eta_{ij} + \sum_k \mathrm{r}_{ijk} E_k + \sum_{kl} \mathrm{s}_{ijkl} E_k E_l, \qquad i,j,k,l = 1,2,3, \tag{21.2-1}$$

where $\eta_{ij} = \eta_{ij}(\mathbf{0})$, $\mathrm{r}_{ijk} = \partial\eta_{ij}/\partial E_k$, $\mathrm{s}_{ijkl} = \frac{1}{2}\partial^2\eta_{ij}/\partial E_k \partial E_l$, and the derivatives are evaluated at $\mathbf{E} = \mathbf{0}$. Equation (21.2-1) is a generalization of (21.1-3), in which r is replaced by $3^3 = 27$ coefficients $\{\mathrm{r}_{ijk}\}$, and s is replaced by $3^4 = 81$ coefficients $\{\mathrm{s}_{ijkl}\}$. The quantities $\{\mathrm{r}_{ijk}\}$ are the coefficients of the (third-rank) **linear electro-optic (Pockels) tensor**, whereas the quantities $\{\mathrm{s}_{ijkl}\}$ represent the coefficients of the (fourth-rank) **quadratic electro-optic (Kerr) tensor**.

Symmetry

Because $\boldsymbol{\eta}$ is symmetric ($\eta_{ij} = \eta_{ji}$), r and s are invariant under permutations of the indices i and j, i.e., $\mathrm{r}_{ijk} = \mathrm{r}_{jik}$ and $\mathrm{s}_{ijkl} = \mathrm{s}_{jikl}$. Also, the coefficients $\mathrm{s}_{ijkl} = \frac{1}{2}\partial^2\eta_{ij}/\partial E_k \partial E_l$ are invariant to permutations of k and l (because of the invariance to the order of differentiation), so that $\mathrm{s}_{ijkl} = \mathrm{s}_{ijlk}$. Because of this permutation symmetry, the nine combinations of the indices i, j generate six instead of nine independent elements. The same reduction applies to the indices k, l. Consequently, r_{ijk} has 6×3 independent elements, whereas s_{ijkl} has 6×6 independent elements.

It is conventional to rename the pair of indices (i, j), $i, j = 1, 2, 3$, as a single index $I = 1, 2, \ldots, 6$ in accordance with Table 21.2-1. The pair (k, l) is similarly replaced by an index $K = 1, 2, \ldots, 6$, in accordance with the same rule. Thus, the elements r_{ijk} and s_{ijkl} are replaced by r_{Ik} and s_{IK}, respectively. For example, r_{12k} is denoted as r_{6k}, s_{1231} is renamed s_{65}, and so on. Hence, the third-rank tensor r is replaced by a 6×3 matrix and the fourth-rank tensor s is contracted to a 6×6 matrix.

$j\backslash i$	1	2	3
1	1	6	5
2	6	2	4
3	5	4	3

Table 21.2-1 Lookup table for the index I that represents the pair of indices (i, j).[a]

[a] The pair $(i, j) = (3, 2)$, for example, is labeled $I = 4$.

Crystal Symmetry

The symmetry of the crystal adds further constraints to the elements of the r and s matrices. Some entries must be zero and others must be equal, or equal in magnitude and opposite in sign, or related by some other rule. For centrosymmetric materials, as an example, r vanishes altogether and only the Kerr effect is exhibited. Tabulations of r and s, and their symmetry relations for the 32 crystallographic point groups (of which 11 are centrosymmetric), may be found in several books referenced in the reading list. Representative examples are provided in Tables 21.2-2 and 21.2-3.

Table 21.2-2 Pockels coefficients r_{Ik} for several representative crystal groups.

$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \mathrm{r}_{41} & 0 & 0 \\ 0 & \mathrm{r}_{41} & 0 \\ 0 & 0 & \mathrm{r}_{41} \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \mathrm{r}_{41} & 0 & 0 \\ 0 & \mathrm{r}_{41} & 0 \\ 0 & 0 & \mathrm{r}_{63} \end{bmatrix}$	$\begin{bmatrix} 0 & -\mathrm{r}_{22} & \mathrm{r}_{13} \\ 0 & \mathrm{r}_{22} & \mathrm{r}_{13} \\ 0 & 0 & \mathrm{r}_{33} \\ 0 & \mathrm{r}_{51} & 0 \\ \mathrm{r}_{51} & 0 & 0 \\ -\mathrm{r}_{22} & 0 & 0 \end{bmatrix}$
Cubic $\bar{4}3m$ (e.g., GaAs, CdTe, InAs)	Tetragonal $\bar{4}2m$ (e.g., KDP, ADP)	Trigonal $3m$ (e.g., BBO, $LiNbO_3$, $LiTaO_3$)

Table 21.2-3 Kerr coefficients s_{IK} for an isotropic medium. The form of this matrix is identical to that for the photoelasticity matrix p_{IK} for an isotropic medium, as displayed in (20.3-4).

$$\begin{bmatrix} \mathrm{s}_{11} & \mathrm{s}_{12} & \mathrm{s}_{12} & 0 & 0 & 0 \\ \mathrm{s}_{12} & \mathrm{s}_{11} & \mathrm{s}_{12} & 0 & 0 & 0 \\ \mathrm{s}_{12} & \mathrm{s}_{12} & \mathrm{s}_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{s}_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{s}_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{s}_{44} \end{bmatrix}, \qquad \mathrm{s}_{44} = \tfrac{1}{2}\left(\mathrm{s}_{11} - \mathrm{s}_{12}\right)$$

Pockels Effect

The following procedure is used to determine the optical properties of an anisotropic medium exhibiting the Pockels effect in the presence of an electric field $\mathbf{E}$:

1. Find the principal axes and principal refractive indices n_1, n_2, and n_3 in the absence of $\mathbf{E}$.
2. Find the coefficients $\{\mathrm{r}_{ijk}\}$ from the appropriate matrix for r_{Ik}, e.g., from Table 21.2-2, by using the rule that relates I to (i, j) provided in Table 21.2-1.
3. Determine the elements of the impermeability tensor

$$\eta_{ij}(\mathbf{E}) = \eta_{ij}(\mathbf{0}) + \sum_k \mathrm{r}_{ijk} E_k, \tag{21.2-2}$$

   where $\eta_{ij}(\mathbf{0})$ is a diagonal matrix with elements $1/n_1^2$, $1/n_2^2$, and $1/n_3^2$.
4. Write the equation for the modified index ellipsoid

$$\sum_{ij} \eta_{ij}(\mathbf{E})\, x_i x_j = 1. \tag{21.2-3}$$

5. Determine the principal axes of the modified index ellipsoid by diagonalization, and find the corresponding principal refractive indices $n_1(\mathbf{E})$, $n_2(\mathbf{E})$, and $n_3(\mathbf{E})$.
6. Given the direction of light propagation, find the normal modes and their associated refractive indices from this index ellipsoid.

EXAMPLE 21.2-1. ***Trigonal $3m$ Crystals ($LiNbO_3$ and $LiTaO_3$).*** Trigonal $3m$ crystals are uniaxial ($n_1 = n_2 = n_o$; $n_3 = n_e$) with the matrix r provided in Table 21.2-2. Assuming that $\mathbf{E} = (0, 0, E)$, i.e., that the electric field points along the optic axis (see Fig. 21.2-3), the modified index ellipsoid is readily shown to be

$$\left(\frac{1}{n_o^2} + \mathrm{r}_{13}E\right)(x_1^2 + x_2^2) + \left(\frac{1}{n_e^2} + \mathrm{r}_{33}E\right)x_3^2 = 1. \tag{21.2-4}$$

This is an ellipsoid of revolution whose principal axes are not altered when the electric field is applied. The ordinary and extraordinary indices, $n_o(E)$ and $n_e(E)$, respectively, are given by

$$\frac{1}{n_o^2(E)} = \frac{1}{n_o^2} + \mathrm{r}_{13}E \tag{21.2-5}$$

$$\frac{1}{n_e^2(E)} = \frac{1}{n_e^2} + \mathrm{r}_{33}E. \tag{21.2-6}$$

Because the terms $\mathrm{r}_{13}E$ and $\mathrm{r}_{33}E$ in (21.2-5) and (21.2-6) are small, we use the Taylor-series approximation $(1 + \Delta)^{-1/2} \approx 1 - \frac{1}{2}\Delta$, valid for $|\Delta| \ll 1$, to obtain

$$n_o(E) \approx n_o - \tfrac{1}{2}n_o^3\mathrm{r}_{13}E \tag{21.2-7}$$

$$n_e(E) \approx n_e - \tfrac{1}{2}n_e^3\mathrm{r}_{33}E. \tag{21.2-8}$$

Note the similarity between these equations and the generic equation (21.1-4). We conclude that when an electric field is applied along the optic axis of this uniaxial crystal it remains uniaxial with the same principal axes, as illustrated in Fig. 21.2-3, but its refractive indices are modified in accordance with (21.2-7) and (21.2-8).

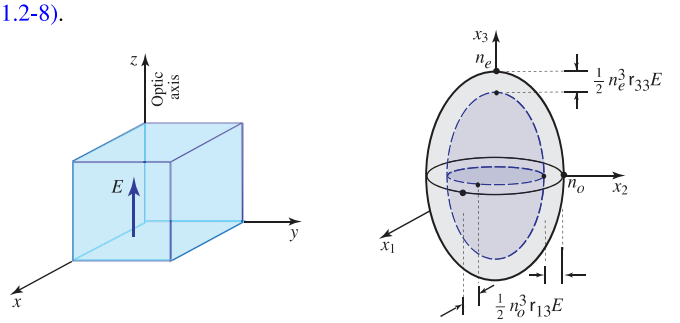


Figure 21.2-3 Modification of the index ellipsoid of a trigonal $3m$ crystal such as $LiNbO_3$ that results from the application of a steady electric field along the direction of the optic axis.

EXAMPLE 21.2-2. ***Tetragonal $\bar{4}2m$ Crystals (KDP and ADP).*** Carrying out the same process for this class of uniaxial crystals, and assuming that the electric field points along the optic axis (Fig. 21.2-4), we obtain the following equation for the index ellipsoid,

$$\frac{x_1^2 + x_2^2}{n_o^2} + \frac{x_3^2}{n_e^2} + 2\mathrm{r}_{63}Ex_1x_2 = 1, \tag{21.2-9}$$

where the factor of two results because $\{i, j = 1, 2\}$. The modified principal axes are obtained by rotating the coordinate system $45°$ about the z axis. Substituting $x_1' = (x_1 + x_2)/\sqrt{2}$, $x_2' = (x_1 - x_2)/\sqrt{2}$, $x_3' = x_3$ in (21.2-9), so that $x_1'^2 + x_2'^2 = x_1^2 + x_2^2$ and $x_1'^2 - x_2'^2 = 2x_1x_2$, and relabeling the new principal axes as (x_1, x_2, x_3), leads to

$$\frac{x_1^2}{n_1^2(E)} + \frac{x_2^2}{n_2^2(E)} + \frac{x_3^2}{n_3^2(E)} = 1, \tag{21.2-10}$$

where

$$\frac{1}{n_1^2(E)} = \frac{1}{n_o^2} + \mathrm{r}_{63}E, \qquad \frac{1}{n_2^2(E)} = \frac{1}{n_o^2} - \mathrm{r}_{63}E, \qquad n_3(E) = n_e. \tag{21.2-11}$$

Cross-multiplying and using the Taylor-series approximation $(1+\Delta)^{-1/2} \approx 1 - \frac{1}{2}\Delta$ yields

$$n_1(E) \approx n_o - \tfrac{1}{2}n_o^3 \mathrm{r}_{63}E \tag{21.2-12}$$

$$n_2(E) \approx n_o + \tfrac{1}{2}n_o^3 \mathrm{r}_{63}E \tag{21.2-13}$$

$$n_3(E) = n_e. \tag{21.2-14}$$

We conclude that in this case the originally uniaxial crystal takes on a biaxial character when subjected to an electric field in the direction of its optic axis, as illustrated in Fig. 21.2-4.

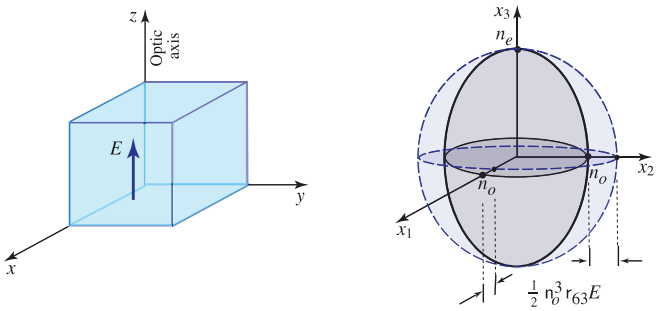


Figure 21.2-4 Modification of the index ellipsoid resulting from the application of a steady electric field E along the direction of the optic axis of a uniaxial tetragonal $\bar{4}2m$ crystal such as KDP.

EXAMPLE 21.2-3. ***Cubic $\bar{4}3m$ Crystals (GaAs, CdTe, and InAs).*** Assuming that the applied electric field points along a cubic axis of the material (taken as the z direction in Fig. 21.2-5), the index ellipsoid for these isotropic crystals $(n_1 = n_2 = n_3 = n)$ becomes

$$\frac{x_1^2 + x_2^2 + x_3^2}{n^2} + 2\mathrm{r}_{41}Ex_1x_2 = 1, \tag{21.2-15}$$

where r_{63} assumes the value r_{41} (see Table 21.2-2). As in Example 21.2-2, the new principal axes are rotated $45°$ about the z axis and the principal refractive indices turn out to be

$$n_1(E) \approx n - \tfrac{1}{2}n^3 \mathrm{r}_{41}E \tag{21.2-16}$$

$$n_2(E) \approx n + \tfrac{1}{2}n^3 \mathrm{r}_{41}E \tag{21.2-17}$$

$$n_3(E) \approx n. \tag{21.2-18}$$

The applied electric field thus causes the isotropic crystal to behave in biaxial fashion (Fig. 21.2-5).

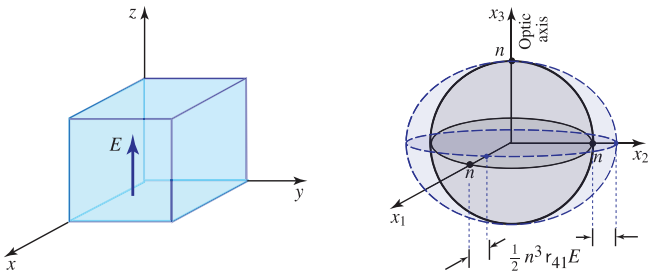


Figure 21.2-5 Modification of the index ellipsoid as a result of applying a steady electric field E along a cubic axis of a $\bar{4}3m$ crystal such as GaAs.

Though cubic crystals have isotropic linear optical properties, they have well-defined crystal axes. Two cubic crystals belonging to different point groups can thus exhibit different optical properties in the presence of a steady electric field. For example, cubic crystals with the $m3m$ diamond-structure, such as Si and Ge, are centrosymmetric and exhibit no Pockels effect whereas $\bar{4}3m$ cubic crystals, such as GaAs and InAs, exhibit anisotropic optical properties, as is clear from Example 21.2-3.

For initially anisotropic materials in which the applied electric field does not alter the principal axes, as in Example 21.2-1, the polarizations of the normal modes remain the same, but their associated refractive indices become dependent on E. The medium can then be conveniently used as a phase modulator, wave retarder, or intensity modulator, in accordance with the generic theory provided in Sec. 21.1B. This principle is described further in Sec. 21.2B.

Kerr Effect

The optical properties of a Kerr medium can be determined by making use of the same procedure used for the Pockels medium, except that the coefficients $\eta_{ij}(\mathbf{E})$ obey

$$\eta_{ij}(\mathbf{E}) = \eta_{ij}(\mathbf{0}) + \sum_{kl} \mathrm{s}_{ijkl} E_k E_l. \tag{21.2-19}$$

EXAMPLE 21.2-4. ***Kerr Effect in an Isotropic Medium.*** With a steady applied electric field E pointing along the z direction, we use the Kerr coefficients for an isotropic medium s_{IK} provided in Table 21.2-3 to find the equation for the index ellipsoid,

$$\left(\frac{1}{n^2} + \mathrm{s}_{12}E^2\right)(x_1^2 + x_2^2) + \left(\frac{1}{n^2} + \mathrm{s}_{11}E^2\right)x_3^2 = 1. \tag{21.2-20}$$

This is the equation of an ellipsoid of revolution whose optic axis is the z axis, along the direction of the applied electric field. The principal refractive indices $n_o(E)$ and $n_e(E)$ are determined from

$$\frac{1}{n_o^2(E)} = \frac{1}{n^2} + \mathrm{s}_{12}E^2 \tag{21.2-21}$$

$$\frac{1}{n_e^2(E)} = \frac{1}{n^2} + \mathrm{s}_{11}E^2. \tag{21.2-22}$$

Since the rightmost terms in (21.2-21) and (21.2-22) are small, we again make use of the approximation $(1+\Delta)^{-1/2} \approx 1 - \frac{1}{2}\Delta$ to obtain

$$n_o(E) \approx n - \tfrac{1}{2}n^3\mathrm{s}_{12}E^2 \tag{21.2-23}$$

$$n_e(E) \approx n - \tfrac{1}{2}n^3\mathrm{s}_{11}E^2. \tag{21.2-24}$$

Thus, a steady electric field E applied to an initially isotropic medium causes it to behave as a uniaxial crystal with the optic axis along the direction of the electric field. In this case, the ordinary and extraordinary indices are *quadratically* decreasing functions of E.

B. Modulators

The principles of phase and intensity modulation using the electro-optic effect were outlined in Sec. 21.1B. Anisotropic effects were introduced only generically. Using the anisotropic theory presented in this section, the generic parameters r and s used

in Sec. 21.1 can be explicitly determined for any crystal, for arbitrary electric-field and light-propagation directions. We restrict our discussion to Pockels modulators, but the same approach is applicable for Kerr modulators. For simplicity, we assume that the direction of the electric field is such that the principal axes of the crystal are not altered. We also assume that the direction of the wave relative to these axes is such that the planes of polarization of the normal modes are not altered by the presence of the electric field.

Phase Modulators

A normal mode is characterized by a refractive index $n(E) \approx n - \frac{1}{2}\mathrm{r}\, n^3 E$, where n and r are the appropriate refractive index and Pockels coefficient, respectively, and where $E = V/d$ is the electric field obtained by applying a voltage V across a distance d. A wave traveling a distance L undergoes a phase shift

$$\varphi = \varphi_o - \pi V/V_\pi \,, \tag{21.2-25}$$

where $\varphi_o = 2\pi n L/\lambda_o$, and the half-wave voltage is

$$V_\pi = \frac{d}{L}\frac{\lambda_o}{\mathrm{r}\, n^3} \,, \tag{21.2-26}$$

in agreement with (21.1-7) and (21.1-8), respectively. The appropriate coefficients generically referred to as n and r can be readily determined in specific cases, as demonstrated in the following example.

EXAMPLE 21.2-5. ***Trigonal*** $3m$ ***Crystals (LiNbO$_3$ and LiTaO$_3$).*** When an electric field is directed along the optic axis of this class of uniaxial crystal, the crystal remains uniaxial with the same principal axes, as shown in Example 21.2-1. The principal refractive indices are given by (21.2-7) and (21.2-8). The crystal can be used as a phase modulator in either of two configurations:

***Longitudinal Modulator*:** If a linearly polarized optical wave travels along the direction of the optic axis (parallel to the electric field), the appropriate parameters for the phase modulator are $n = n_o$, $\mathrm{r} = \mathrm{r}_{13}$, and $d = L$. For $LiNbO_3$, $\mathrm{r}_{13} = 9.6$ pm/V and $n_o = 2.3$ at $\lambda_o = 633$ nm. Equation (21.2-26) then yields $V_\pi = 5.41$ kV, the voltage required to change the phase by π.

***Transverse Modulator*:** If the wave travels in the x direction and is polarized in the z direction, the appropriate parameters are $n = n_e$ and $\mathrm{r} = \mathrm{r}_{33}$. The width d is generally not equal to the length L. For $LiNbO_3$ at $\lambda_o = 633$ nm, $r_{33} = 30.9$ pm/V and $n_e = 2.2$, which results in a half-wave voltage $V_\pi = 1.9(d/L)$ kV. If $d/L = 0.1$, we obtain $V_\pi \approx 190$ V, which is significantly lower than the half-wave voltage for the longitudinal modulator.

Intensity Modulators

The difference in the dependence of the refractive indices of the two normal modes of a Pockels cell on the applied field provides a voltage-dependent retardation,

$$\Gamma = \Gamma_0 - \pi V/V_\pi, \tag{21.2-27}$$

where

$$\Gamma_0 = 2\pi(n_1 - n_2)L/\lambda_o \tag{21.2-28}$$

$$V_\pi = \frac{d}{L}\frac{\lambda_o}{\mathrm{r}_1 n_1^3 - \mathrm{r}_2 n_2^3} \,, \tag{21.2-29}$$

in agreement with (21.1-12) and (21.1-13). If the cell is placed between crossed polarizers, the system serves as an intensity modulator (Sec. 21.1B). It is not difficult to determine the appropriate indices n_1 and n_2, and coefficients r_1 and r_2, as illustrated by Example 21.2-6.

EXAMPLE 21.2-6. ***Tetragonal $\bar{4}2m$ Crystals (KDP and ADP).*** As described in Example 21.2-2, when an electric field is applied along the optic axis of this uniaxial crystal, it behaves as a biaxial crystal. The new principal axes are the original axes rotated by $45°$ about the optic axis. Assume a longitudinal modulator configuration ($d/L = 1$) in which the wave travels along the optic axis. The two normal modes have refractive indices given by (21.2-12) and (21.2-13). The appropriate coefficients to be used in (21.2-29) are thus $n_1 = n_2 = n_o$, $\mathrm{r}_1 = \mathrm{r}_{63}$, $\mathrm{r}_2 = -\mathrm{r}_{63}$, and $d = L$, so that $\Gamma_0 = 0$ and

$$V_\pi = \frac{\lambda_o}{2\mathrm{r}_{63} n_o^3} . \tag{21.2-30}$$

For KDP at $\lambda_o = 633$ nm, $V_\pi = 8.4$ kV.

EXERCISE 21.2-1

Intensity Modulation Using the Kerr Effect. Use (21.2-23) and (21.2-24) to determine an expression for the phase shift φ and the phase retardation Γ in a longitudinal Kerr modulator made of an isotropic material, as functions of the applied voltage V. Derive expressions for the half-wave voltages V_π in each case.

21.3 ELECTRO-OPTICS OF LIQUID CRYSTALS

As described in Sec. 6.5, the elongated molecules of nematic liquid crystals tend to have ordered orientations that are altered when the material is subjected to mechanical or electric forces. Because of their unique anisotropic properties, liquid crystals can serve as dynamic wave retarders or polarization rotators. The presence of an electric field modifies their molecular orientation, so that their effect on polarized light is altered. Liquid crystals can therefore be used as electrically controlled optical wave retarders, modulators, and switches. These devices are particularly useful in display technology.

A. Wave Retarders and Modulators

Electrical Properties of Nematic Liquid Crystals

The liquid crystals used to make electro-optic devices are usually of sufficiently low conductivity that they can be regarded as ideal dielectric materials. Because of the elongated shape of the constituent molecules and their ordered orientation, liquid crystals have anisotropic dielectric properties with uniaxial symmetry. The electric permittivity is defined as $\epsilon_\parallel$ for electric fields pointing along the long axes of the molecules and as $\epsilon_\perp$ for fields pointing in the perpendicular direction. Liquid crystals for which $\epsilon_\parallel > \epsilon_\perp$ (positive uniaxial) are usually selected for electro-optic applications.

The application of a steady (or low frequency) electric field to a liquid crystal induces electric dipoles in its molecules and the resultant electric forces exert torques on them. The molecules rotate in a direction such that the free electrostatic energy, $-\frac{1}{2}\mathbf{E}\cdot\mathbf{D} = -\frac{1}{2}[\epsilon_\perp E_1^2 + \epsilon_\perp E_2^2 + \epsilon_\parallel E_3^2]$, is minimized (here, E_1, E_2, and E_3 are the components of $\mathbf{E}$ in the directions of the principal axes). Since $\epsilon_\parallel > \epsilon_\perp$, for a given direction of the electric field, minimum energy is achieved when the molecules are

aligned with the field. In that case $E_1 = E_2 = 0$, so that $\mathbf{E} = (0, 0, E)$, and the free energy is $-\frac{1}{2}\epsilon_{\|}E^2$. When alignment is complete, the long molecular axes point in the direction of the electric field, as portrayed in Fig. 21.3-1. A reversal of the electric field results in the same molecular rotation and an alternating field generated by an AC voltage also has the same effect.

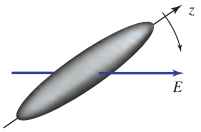
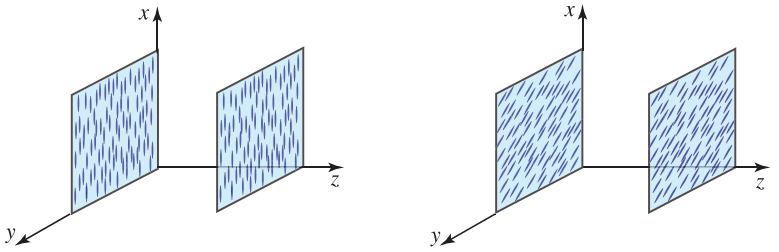


Figure 21.3-1 The molecules of a positive uniaxial liquid crystal rotate so their long molecular axes align with the applied electric field.

Nematic Liquid-Crystal Wave Retarders and Modulators

A nematic liquid-crystal cell is a thin layer of nematic liquid crystal placed between two parallel glass plates and rubbed so that the molecules are parallel to each other. The material then acts as a uniaxial crystal with the optic axis parallel to the molecular orientation. For waves traveling in the z direction (perpendicular to the glass plates), the normal modes are linearly polarized in the x and y directions (parallel and perpendicular to the molecular directions, respectively), as illustrated in Fig. 21.3-2(a). The principal refractive indices are then the extraordinary and ordinary indices, n_e and n_o, respectively. A cell of thickness d provides a wave retardation $\Gamma = 2\pi(n_e - n_o)d/\lambda_o$.



(a) Untilted state (b) Tilted state

Figure 21.3-2 Molecular orientation of a liquid-crystal cell: (a) in the absence of a steady electric field; (b) in the presence of a steady electric field. The optic axis lies along the direction of the long axes of the molecules.

Now, if an electric field is applied *in the z direction* (by applying a voltage V across transparent conductive electrodes coated on the insides of the glass plates), the resultant electric forces tend to tilt the molecules toward alignment with the field, but the elastic forces at the surfaces of the glass plates resist this motion. When the applied electric field is sufficiently strong, however, most of the molecules (except for those adjacent to the glass surfaces) tilt toward the z axis. The equilibrium tilt angle θ (with respect to the x–y) plane for these molecules is a monotonically increasing function of V characterized by[†]

$$\theta = \begin{cases} 0 & V \le V_c \\ \dfrac{\pi}{2} - 2\tan^{-1}\left[\exp\left(-\dfrac{V - V_c}{V_0}\right)\right] & V > V_c\,, \end{cases} \tag{21.3-1}$$

where V is the applied RMS voltage, V_c is a critical voltage at which the tilting process begins, and V_0 is a constant. When $(V - V_c)/V_0 = 1$, we find that $\theta \approx 50°$. As $(V - V_c)/V_0$ increases beyond unity, θ approaches 90° as shown in Fig. 21.3-3(a).

[†] See, e.g., P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Oxford University Press, 2nd ed. 1993.

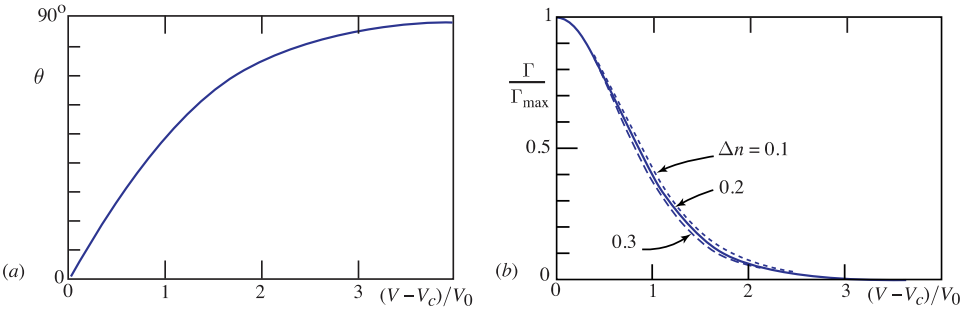


Figure 21.3-3 (a) Dependence of the tilt angle θ of the liquid-crystal molecules on the normalized RMS voltage applied to the liquid-crystal cell. (b) Dependence of the normalized retardation $\Gamma/\Gamma_{\max} = [n(\theta) - n_o]/(n_e - n_o)$ on the normalized RMS voltage when $n_o = 1.5$, for the values of $\Delta n = n_e - n_o$ indicated. This plot is obtained from (21.3-1) and (21.3-2).

When the electric field is removed, the orientations of the molecules near the glass plates are reasserted and all of the molecules tilt back to their original orientations, in planes parallel to the plates. In a sense, the liquid-crystal material may be viewed as a *liquid with memory*.

For a tilt angle θ, the normal modes of an optical wave traveling in the z direction are polarized in the x and y directions and have refractive indices $n(\theta)$ and n_o, where

$$\frac{1}{n^2(\theta)} = \frac{\cos^2\theta}{n_e^2} + \frac{\sin^2\theta}{n_o^2}. \tag{21.3-2}$$

Note that for a tilt angle θ, the direction between the optic axis and the direction of propagation is $90^\circ - \theta$, which explains why (21.3-2) differs from (6.3-15).

The liquid-crystal cell finds use in a number of applications:

- **Phase modulator**. For an optical wave traveling in the z direction that is linearly polarized in the x direction (parallel to the untilted molecular orientation), the phase shift is $\varphi = 2\pi n(\theta)d/\lambda_o$, where $n(\theta)$ is given by (21.3-2). Since θ is controlled by the voltage applied to the cell in accordance with (21.3-1), the cell can be readily used as a voltage-controlled phase modulator.
- **Variable wave retarder**. The retardation $\Gamma = 2\pi[n(\theta) - n_o]d/\lambda_o$ also depends on the tilt angle θ. It achieves its maximum value $\Gamma_{\max} = 2\pi(n_e - n_o)d/\lambda_o$ when the molecules are not tilted ($\theta = 0$) and decreases monotonically toward 0 when the tilt angle reaches 90°. Using (21.3-2) and (21.3-1), the dependence of Γ on the applied voltage is plotted in Fig. 21.3-3(b). The liquid-crystal cell therefore serves as a voltage-controlled wave retarder with principal axes along the x and y directions.
- **Intensity modulator**. When the liquid-crystal cell is placed between two crossed polarizers (at $\pm 45^\circ$ with respect to the x axis in the x–y plane), the transmittance of the device is $\mathcal{T} = \sin^2(\Gamma/2)$, which is a function of the voltage-controlled retardation Γ and has a maximum value of unity when the retardation is π. Since the retardation is voltage dependent, so too is the transmittance so that the device functions as a voltage-controlled intensity modulator. Intensity modulation may also be implemented in reflection mode by placing the cell between a mirror and a polarizer oriented at 45° with respect to the x-axis, as illustrated in Fig. 21.3-4(a), in which case the reflectance is controlled by the applied voltage. A sketch of the reflectance versus the applied voltage is provided in Fig. 21.3-4(b).

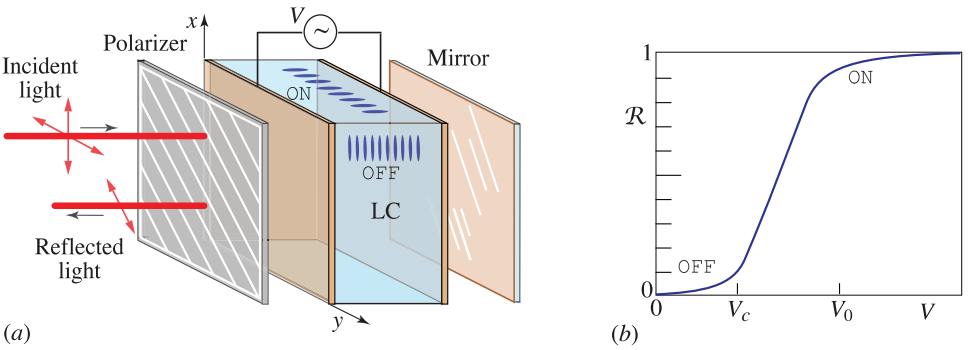


Figure 21.3-4 (a) A nematic liquid-crystal cell placed between a polarizer and a mirror functions as a reflector with voltage-controllable reflectance. The retardation Γ varies between $\pi/2$ and 0 as the voltage is tuned between 0 ("OFF" state) and the saturation voltage V_0 ("ON" state). After reflection from the mirror and a round trip through the crystal, the plane of polarization rotates 90° in the "OFF" state, so that the light is blocked. In the "ON" state, there is no rotation, and the reflected light is transmitted by the polarizer. (b) Dependence of the intensity reflectance $\mathcal{R}$ on the applied voltage V.

Liquid-Crystal Cell Parameters

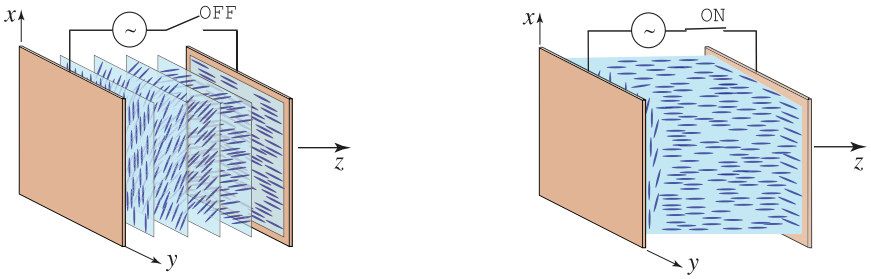
Liquid-crystal cells are usually sealed between antireflection-coated, optically flat, glass windows. Liquid-crystal layers have typical thicknesses $d \approx 10$ μm and a refractive-index difference $\Delta n = n_e - n_o$ in the vicinity of 0.1–0.3. The *retardation* $\Gamma = 2\pi\varrho/\lambda_o$ is often expressed in terms of the *retardance* $\varrho = (n_e - n_o)d$; retardances of several hundred nanometers are typical (e.g., a retardance of 300 nm corresponds to a retardation of π at $\lambda_o = 600$ nm). The response time of a liquid crystal layer depends on its thickness, as well as on the viscosity of the material, its temperature, and the nature of the applied drive voltage. Liquid-crystal cells are slow devices; the rise time is of the order of tens of milliseconds if the operating voltage is near the critical voltage V_c and roughly a few milliseconds at higher voltages. The critical voltage V_c is typically a few volts RMS. The decay time is insensitive to the operating voltage but can be reduced by using cells of smaller thickness. The voltage is usually applied as a square waveform with a frequency that ranges between tens of Hz and a few kHz. Operation at lower frequencies tends to cause electromechanical effects that disrupt molecular alignment and reduce the lifetime of the device, while frequencies higher than 100 Hz entail greater power consumption because of increased conductivity.

Twisted Nematic Liquid-Crystal Modulators

A *twisted* nematic liquid-crystal cell is a thin layer of nematic liquid crystal placed between two parallel glass plates and rubbed in such a way that the molecular orientation rotates helically about an axis normal to the plates (the axis of twist). If the angle of twist is 90°, for example, the molecules point in the x direction at one plate and in the y direction at the other [Fig. 21.3-5(a)]. Consecutive transverse layers of the material act as uniaxial crystals with an optic axis that rotates helically about the axis of twist. It was demonstrated in Sec. 6.5 that the plane of polarization for linearly polarized light traveling along the direction of the axis of twist rotates with the molecules, so that the cell acts as a polarization rotator.

When an electric field is applied along the direction of the axis of twist (the z direction) the molecules tilt toward the field [Fig. 21.3-5(b)]. When the tilt is 90°, the molecules lose their twisted character (except for those adjacent to the glass surfaces), whereupon the polarization rotatory power is deactivated. If the electric field is removed, the orientations of the layers near the glass surfaces dominate, causing the

molecules to return to their original twisted state and the polarization rotatory power to be regained.



(*a*) Twisted state (*b*) Tilted (untwisted) state

Figure 21.3-5 (*a*) A twisted nematic liquid-crystal cell in its twisted state. (*b*) In the presence of a sufficiently strong electric field, the molecules tilt in the direction of the field and lose their twisted character.

Since the polarization rotatory power may be turned off and on by switching the electric field on and off, a shutter can be designed by placing a cell providing a 90° twist between two crossed polarizers. The system then transmits light in the absence of an electric field and blocks it when the electric field is applied, as portrayed in Fig. 21.3-6.

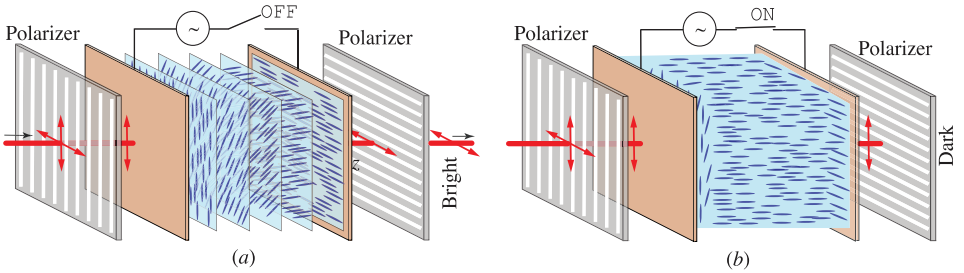


Figure 21.3-6 A twisted nematic liquid-crystal modulator. (*a*) When the electric field is absent, the liquid-crystal cell acts as a polarization rotator and the light is transmitted. (*b*) When the electric field is present, the cell's rotatory power is suspended and the light is blocked. At intermediate values of the applied voltage, the device provides partial intensity transmission and therefore operates as an analog intensity modulator.

The twisted liquid-crystal cell placed between crossed polarizers may also be operated as an analog modulator. Intermediate tilt angles result in a combination of polarization rotation and wave retardation. The analysis of the transmission of polarized light through tilted and twisted molecules is rather complex, but the overall effect is partial intensity transmittance. There is an approximately linear range of transition between the total transmission of the fully twisted (untilted) state and the total blocking of the fully tilted (untwisted) state, although the dynamic range is rather limited.

Operation in the reflective mode is also possible, as illustrated in Fig. 21.3-7. Here, the twist angle is 45°; a mirror is placed on one side of the cell and a polarizer on the other side. When the electric field is absent the polarization plane rotates a total of 90° as the wave propagates a round trip through the cell; the reflected light is therefore blocked by the polarizer. When the electric field is present, the polarization rotatory power is suspended and the reflected light is transmitted through the polarizer. Other reflective and transmissive modes of operation with different angles of twist are also possible.

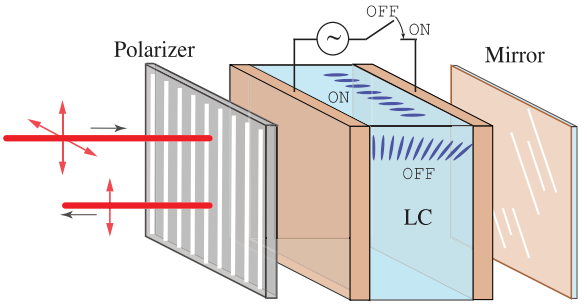


Figure 21.3-7 A twisted nematic liquid-crystal cell with a 45° twist angle and a mirror provides a round-trip polarization rotation of 90° in the absence of the electric field (blocked state) and no rotation when the field is applied (unblocked state). The device is a reflective switch. Much as with the nematic liquid-crystal cell illustrated in Fig. 21.3-4, this device can be operated as a reflective analog intensity modulator for intermediate values of the field.

Ferroelectric Liquid Crystals

Smectic liquid crystals are organized in layers, as displayed in Fig. 6.5-1(b). In the smectic-C phase, the molecular orientation is tilted by an angle θ with respect to the normal to the parallel smectic layers (the x axis), as illustrated in Fig. 21.3-8. This material has ferroelectric properties. When placed between two closely spaced glass plates the surface interactions permit only two stable states of molecular orientation: at the angles $\pm\theta$, as shown in Fig. 21.3-8. When an electric field $+E$ is applied in the z direction, a torque is produced that switches the molecular orientation into the stable state $+\theta$ [Fig. 21.3-8(a)]. The molecules can be switched into the state $-\theta$ by applying an electric field $-E$ of the opposite polarity [Fig. 21.3-8(b)]. The cell therefore acts as a uniaxial crystal whose optic axis may be switched between two orientations.

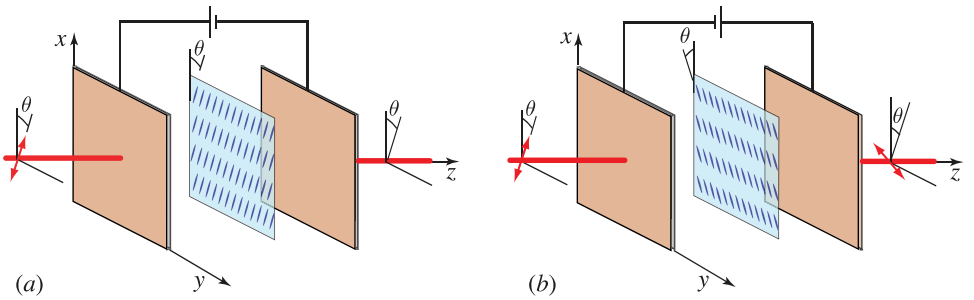


Figure 21.3-8 The two allowed states of a ferroelectric liquid-crystal cell.

In the geometry of Fig. 21.3-8, the incident light is linearly polarized at an angle θ with respect to the x axis. In the $+\theta$ state at left, the polarization is parallel to the optic axis and the wave travels with the extraordinary refractive index n_e without retardation. In the $-\theta$ state at right, the polarization makes an angle 2θ with the optic axis. If $2\theta = 45°$, the wave undergoes a retardation $\Gamma = 2\pi(n_e - n_o)d/\lambda_o$, where d is the thickness of the cell and n_o is the ordinary refractive index. If d is selected such that $\Gamma = \pi$, the device acts as a half-wave retarder and the plane of polarization is rotated by 90°, as illustrated in Fig. 6.1-8(b). Hence, reversing the applied electric field has the effect of rotating the plane of polarization by 90°. A switch is therefore readily constructed by placing this cell between two crossed polarizers. The principal merit of ferroelectric liquid-crystal switches lies in their fast response times at room temperature. Their μs response times are far faster than the ms response times associated with nematic liquid crystals. The switching voltage is on the order of ± 10 V.

B. Spatial Light Modulators and Displays

Spatial light modulators (**SLMs**) modify the spatial distribution of the wavefront (phase), intensity, or polarization of an optical wave by making use of an array of optical modulators arranged in a particular spatial configuration. Wavefront modulation is used for beam steering and focusing, as well as in adaptive optics. The principal application of intensity modulation is for the display of spatial patterns, such as numbers, letters, graphics, and images. Liquid-crystal technology is widely used in SLMs and **liquid-crystal displays** (**LCDs**) are dominant in applications such as mobile phones, digital watches, laptop and desktop computers, and flat panel and high-definition television receivers.

LCDs are designed to operate either in a transmissive mode (**T-mode**), as in the configuration depicted in Fig. 21.3-6, or in a reflective mode (**R-mode**), as exemplified in Fig. 21.3-7. T-mode LCDs rely on a **backlight** placed at the rear of the device; the emitted light is transmitted through the liquid-crystal cell and is viewed from the front of the device. The backlight generally consists of a panel of white LEDs or WOLEDs (Sec. 18.1F), or a cold-cathode fluorescent lamp; the emitted light is rendered spatially uniform with the help of a diffuser. The images in R-mode LCDs are usually generated using the reflection of ambient light from the device. R-mode devices thus have the dual merits of low power consumption and superior readability at high ambient light levels (such as outdoors), but they cannot be used in the dark. Transreflective devices that can operate in either mode are available.

LCDs make use of nematic, twisted-nematic, or ferroelectric liquid-crystal cells (Sec. 21.3A). The configuration most commonly used in laptop computers and high-definition television receivers is the 90° twisted-nematic transmissive configuration (Fig. 21.3-6), principally because it offers high contrast. The contrast of LCDs typically degrades significantly as the angle of view increases, in part because oblique rays undergo different retardation at different angles of incidence/reflection, particularly when the molecules are rotated in an out-of-plane direction. Special compensation filters have been developed for wide-angle viewing.

Segmented LCD

A segmented LCD is constructed by placing transparent electrodes in a particular configuration on the glass plate of a reflective liquid-crystal cell. Applying a voltage to selected electrodes (with respect to a common electrode at the rear of the device) results in the desired pattern of reflection and nonreflection. As an example, a seven-segment electrode configuration suitable for displaying the numerals 0 to 9 is illustrated in Fig. 21.3-9. Larger numbers of electrodes may be addressed sequentially, such as by the use of a charge-coupled configuration (Sec. 19.5).

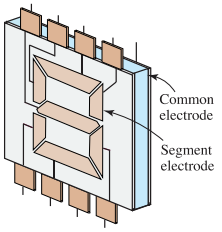


Figure 21.3-9 Electrodes of a seven-bar-segment reflective-mode LCD.

Matrix LCD

A matrix LCD [Fig. 21.3-10(a)] makes use of transparent electrodes arranged in the form of rows and columns, with pixels defined by the locations of their intersections. Each pixel functions much like the single liquid-crystal modulator illustrated in Fig. 21.3-6. The image information is imparted via voltages applied across the

transparent electrodes of each pixel, which in turn determine the tilt angles of their liquid-crystal layers and thence the intensities of the light they transmit. The resolution of the device is determined by the pixel density.

Given a passive-matrix LCD with N rows and M columns, an external wire is required to address each of the NM pixels, at least in principle. Since this is not feasible for large arrays, in practice only the rows and columns are addressed, requiring $N + M$ external wires. Crosstalk among the pixels is minimized by entering the image sequentially over a time period T that is shorter than the response time of the viewer's visual system. This period is divided into N intervals; the N rows are sequentially scanned with an applied voltage, while the image data are entered by applying a positive or negative voltage to each of the M columns, depending on the desired image. A schematic of the system and its operation is presented in Fig. 21.3-10.

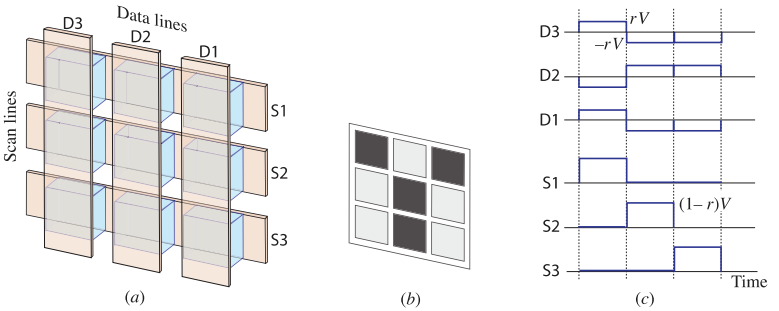


Figure 21.3-10 (a) A matrix LCD is addressed by applying sequential scanning (S) voltage pulses to the rows, and data (D) voltage pulses to the columns. (b) Image obtained by applying to the rows and columns the voltage-pulse sequences displayed in (c).

A binary image with dark and bright pixels, such as that displayed in Fig. 21.3-10(b) may be entered into a matrix LCD by means of the following scheme: During the nth time interval, a voltage $(1 - r)V$ is applied to the nth row (V is a constant voltage and $r < 1$), while all other rows are maintained at 0 V. Concurrently, a voltage $-rV$ is applied to the mth column if the (m, n) pixel is to be bright, and $+rV$ is applied if it is to be dark. The voltage applied to the (m, n) pixel during the nth time interval is therefore $(1-r)V-(-rV) = V$ if the pixel is bright, and $(1-r)V-(rV) = (1-2r)V$ if it is dark. During the other $N - 1$ time intervals, the voltage applied to the (m, n) pixel is either $0 - (-rV) = rV$ or $0 - (rV) = -rV$. The LCD is responsive to the RMS voltage V_{RMS} averaged over the sequence of N intervals. For the (m, n) pixel, $V_{RMS}^2 = (1/N)[V^2 + (N - 1)r^2V^2]$ if the pixel is bright, and $V_{RMS}^2 = (1/N)[(1 - 2r)^2V^2 + (N - 1)r^2V^2]$ if it is dark. The ratio of these two voltages is maximum if $r = 1/(\sqrt{N} + 1)$, in which case V_{RMS} in the bright state is greater than that in the dark states by the factor $[(\sqrt{N} + 1)/(\sqrt{N} - 1)]^{1/2}$. Though this ratio is only slightly greater than unity, small changes in the applied voltage can switch the pixel from dark to bright because of the nonlinear relation between the transmittance and the applied voltage. Nevertheless, this type of passive addressing system is clearly not adequate if N is too large.

Active-Matrix LCD (AMLCD)

The AMLCD eliminates pixel-to-pixel crosstalk by making use of thin-film transistor (TFT) circuitry. As illustrated in Fig. 21.3-11, each pixel is coupled to its own TFT, gated by a signal from its row wire, while the data are entered via its column wire. When the gate signal is present, the TFT applies a voltage V between the indium-tin-oxide (ITO) pixel electrode and an ITO common electrode during a time duration

T/N, where T is the frame period and N is the number of rows. Since the data are entered row by row, crosstalk is eliminated. Though the voltage applied on each pixel appears only during the brief frame time, the viewer perceives the image in its totality since the integration time of the human visual system is long (of the order of ms) in comparison with the frame time.

Figure 21.3-11 illustrates the structure and operation of an AMLCD in the T-mode configuration. The liquid-crystal (LC) material is sandwiched between two thin glass plates with a single ITO common electrode on one side of it and the ITO pixel electrode array on the other side. The stack is placed between a pair of crossed polarizer sheets that form the display screen. Color display is implemented by segmenting each pixel into three independently addressed adjacent segments (sub-pixels) with red, green, and blue color filters in front, as shown in Fig. 21.3-11(b). With white backlight illuminating the array, the color displayed by each pixel is determined by the superposition of the light transmitted through its three sub-pixels. Metameric white light is produced via additive color mixing, as described in Sec. 18.1F.

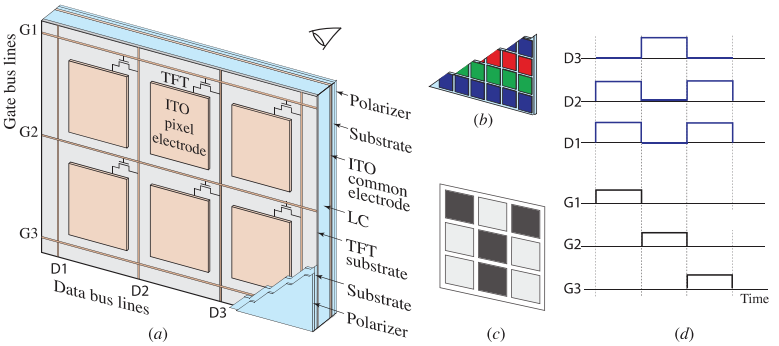


Figure 21.3-11 An active-matrix liquid-crystal display (AMLCD). (a) Device structure viewed from the rear. Pixels are addressed via thin-film transistors (TFTs) gated by signals (G) applied from the rows and data (D) entered via the columns. The white backlight is not shown. (b) Adjacent color filters placed between the common electrode and substrate provide color sub-pixels. (c) Black-and-white image obtained by applying the gating and data waveforms displayed in (d). The electrodes are fabricated from indium tin oxide (ITO), a transparent conductor.

Optical phase-modulation SLMs, widely used for applications such as wavefront modification and the reshaping of optical beams, often make use of liquid-crystal-on-silicon technology, and are typically configured in the reflective mode.

A competing display technology makes use of TFT circuitry to control arrays of organic light-emitting diodes. The color sub-pixels in an active-matrix organic light-emitting display (AMOLED) are created by OLEDs that directly emit red, green, and blue light (Sec. 18.1E), or by WOLEDs that emit white OLED light that is passed through color filters. AMOLEDs have a number of salutary features in comparison with AMLCDs: they are 1) thinner, 2) flexible, 3) superior in color rendition, 4) superior in contrast, 5) insensitive to viewing angle, and 6) faster. However, they are more expensive in the current state of their development.

Optically Addressed Spatial Light Modulators

Most LCDs are addressed electrically. However, optically addressed spatial light modulators are attractive for applications involving image and optical data processing. Light with an intensity distribution $I_W(x, y)$, the "write" image, is converted by an optoelectronic sensor into an electric-field distribution $E(x, y)$, which controls the reflectance $\mathcal{R}(x, y)$ of a liquid-crystal cell operated in the reflective mode. A separate optical wave of uniform intensity is reflected from the device and creates the "read"

image $I(x, y) \propto \mathcal{R}(x, y)$. Thus, the "read" image is controlled by the "write" image (Fig. 21.1-14).

If the write image is carried by incoherent light, and the read image is formed by coherent light, the device serves as a *spatial incoherent-to-coherent light converter*, much like the PROM device discussed earlier (Sec. 21.1E). Furthermore, the wavelengths of the write and read beams need not be the same. The read light may also be more intense than the write light, so that the device may serve as an image intensifier.

There are several ways of converting the write image $I_W(x, y)$ into a pattern of electric field $E(x, y)$ for application to the liquid-crystal cell. A layer of photoconductive material, e.g., cadmium sulfide (CdS), placed between the electrodes of a capacitor may be used. When illuminated by the distribution $I_W(x, y)$, the conductance $G(x, y)$ is altered proportionally. The capacitor is then discharged at each position in accordance with the local conductance, so that the resultant voltage and electric field $E(x, y) \propto 1/I_W(x, y)$ is a negative of the original image. An alternative makes use of a sheet photodiode [e.g., a p–i–n photodiode of hydrogenated amorphous silicon (α-Si:H)]. The reverse-biased photodiode conducts in the presence of light, thereby creating a potential difference proportional to the local light intensity.

An example of a commercially available liquid-crystal spatial light modulator (SLM) is the **Parallel-Aligned Spatial Light Modulator (PAL-SLM)** illustrated in Fig. 21.3-12. This device uses α-Si:H as the write medium and a nematic LC with molecules in parallel alignment as a phase modulator. At each point, the impedance of the amorphous silicon layer is altered by the write light and a voltage proportional to the optical intensity is applied to the corresponding point in the LC layer. This results in rotation of the anisotropic LC molecules to align with the applied electric field. Consequently, the read light beam undergoes a proportional phase shift as it travels through the LC layer. The PAL-SLM is a continuous modulator (i.e., it is not pixelated). It has high spatial resolution, corresponding to 480×480 points over its active area of 2×2 cm^2, and its rise (fall) time is 30 (40) ms.

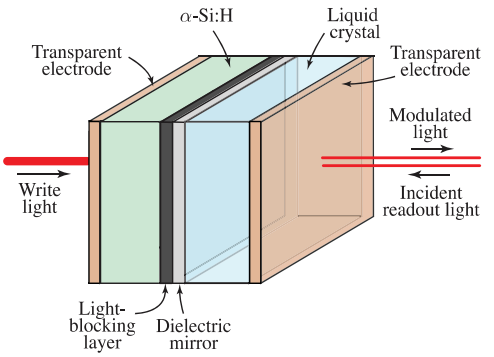


Figure 21.3-12 Schematic of the Hamamatsu Parallel-Aligned Spatial Light Modulator (PAL-SLM). This optically addressed SLM has two principal layers: an amorphous silicon layer, which senses the write light intensity, and a liquid-crystal (LC) layer that serves as a reflective phase modulator for the read light. These layers are separated by a light-blocking dielectric material. The device is encased in glass substrates (not shown).

*21.4 PHOTOREFRACTIVITY

Photorefractive materials exhibit photoconductive and electro-optic behavior, and have the ability to detect and store spatial distributions of optical intensity in the form of spatial patterns of altered refractive index. Photoinduced charges create a space-charge distribution that produces an internal electric field, which in turn alters the refractive index by means of the electro-optic effect.

Ordinary *photoconductive* materials are often good insulators in the dark. Upon illumination, photons are absorbed, free charge carriers (electron–hole pairs) are generated, and the conductivity of the material increases. When the light is removed, the process of charge photogeneration ceases, and the conductivity returns to its dark value

as the excess electrons and holes recombine. Photoconductors are most often used as *photodetectors* (Sec. 19.2).

When a *photorefractive* material is exposed to light, free charge carriers (electrons or holes) are generated by excitation from impurity energy levels to an energy band at a rate proportional to the optical intensity. This process is much like that in an extrinsic semiconductor photoconductor (Sec. 19.2B). These carriers then diffuse away from the positions of high intensity where they were generated, leaving behind fixed charges of the opposite sign (associated with the impurity ions). The free carriers can be trapped by ionized impurities at other locations, depositing their charge there as they recombine. The result is the creation of an inhomogeneous space-charge distribution that remains in place for a period of time after the light is removed. This charge distribution creates an internal electric-field pattern that modulates the local refractive index of the material by virtue of the (Pockels) electro-optic effect. The image may then be accessed optically by monitoring the spatial pattern of the refractive index using a probe optical wave. The material can be brought back to its original state (erased) by illumination with uniform light, or by heating. Thus, a photorefractive material can be used to record and store images, much as with a photographic emulsion. The process is illustrated in Fig. 21.4-1 for iron-doped lithium niobate ($Fe^{2+/3+}$:$LiNbO_3$).

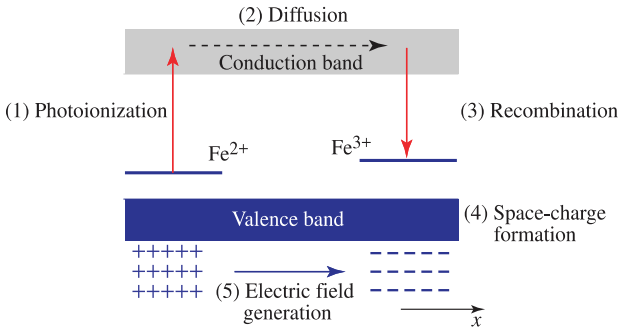


Figure 21.4-1 Energy-level diagram of $LiNbO_3$ doped with Fe ions that illustrates the processes of (1) photoionization, (2) diffusion, (3) recombination, (4) space-charge formation, and (5) electric-field generation. The Fe^{2+} impurity centers act as donors and become Fe^{3+} after photoionization, whereas the Fe^{3+} centers act as traps and revert to Fe^{2+} after recombination.

Important photorefractive materials include lithium niobate ($LiNbO_3$), potassium niobate ($KNbO_3$), barium titanate ($BaTiO_3$), bismuth silicon oxide ($Bi_{12}SiO_{20}$ or BSO), strontium barium niobate ($Sr_xBa_{1-x}Nb_2O_6$ or SBN), and gallium arsenide (GaAs).

Simplified Theory of Photorefractivity

When a photorefractive material is illuminated by light of intensity $I(x)$ that varies in the x direction, the refractive index changes by $\Delta n(x)$. The following is a step-by-step description of the processes that mediate this effect, as illustrated in Fig. 21.4-1, together with a simplified set of one-dimensional equations that govern these processes:

1. *Photoionization.* The absorption of a photon at position x raises an electron from the donor level to the conduction band. The rate of photoionization $G(x)$ is proportional both to the optical intensity and to the number density of nonionized donors. Thus,

$$G(x) = s\,(N_D - N_D^+)\,I(x), \tag{21.4-1}$$

where N_D is the overall number density of donors, N_D^+ is the number density of ionized donors, and s is a constant proportional to the *photoionization cross section*.

2. *Diffusion.* Since $I(x)$ is nonuniform, the number density of excited electrons $\mathrm{n}(x)$ is also nonuniform. As a result, electrons diffuse from locations of high concentration to locations of low concentration.
3. *Recombination.* The electrons recombine at a rate $R(x)$ proportional to their number density $\mathrm{n}(x)$, and to the number density of ionized donors (traps) N_D^+, so that

$$R(x) = \gamma_R \, \mathrm{n}(x) \, N_D^+, \tag{21.4-2}$$

where γ_R is a constant. In equilibrium, the rate of recombination equals the rate of photoionization, i.e., $R(x) = G(x)$, so that

$$s \, I(x) \, (N_D - N_D^+) = \gamma_R \, \mathrm{n}(x) \, N_D^+, \tag{21.4-3}$$

from which

$$\mathrm{n}(x) = \frac{s}{\gamma_R} \frac{N_D - N_D^+}{N_D^+} \, I(x). \tag{21.4-4}$$

4. *Space-charge formation.* Each photogenerated electron leaves behind a positive ionic charge. When the electron is trapped (recombines), its negative charge is deposited at a different site. As a result, a nonuniform space-charge distribution is formed.
5. *Electric-field generation.* This nonuniform space charge generates a position-dependent electric field $E(x)$, which may be determined by observing that in steady state the drift and diffusion electric-current densities must be of equal magnitude and opposite sign, so that the total current density vanishes, i.e.,

$$J = e\mu_e \, \mathrm{n}(x) \, E(x) - kT \, \mu_e \frac{d\mathrm{n}}{dx} = 0 \, , \tag{21.4-5}$$

where μ_e is the electron mobility, k is Boltzmann's constant, and T is the temperature. Thus,

$$E(x) = \frac{kT}{e} \frac{1}{\mathrm{n}(x)} \frac{d\mathrm{n}}{dx}. \tag{21.4-6}$$

6. *Refractive-index modification.* Since the material is electro-optic, the internal electric field $E(x)$ locally modifies its refractive index in accordance with

$$\Delta n(x) = -\tfrac{1}{2} n^3 \mathrm{r} \, E(x), \tag{21.4-7}$$

where n and r are the appropriate values of refractive index and Pockels electro-optic coefficient for the material [see (21.1-4)].
7. *Photorefractive image storage.* The relation between the incident light intensity $I(x)$ and the resultant refractive index change $\Delta n(x)$ may be readily estimated if we assume that the quantity $(N_D/N_D^+ - 1)$ in (21.4-4) is approximately constant, independent of x. In that case $\mathrm{n}(x)$ is proportional to $I(x)$, whereupon (21.4-6) yields

$$E(x) = \frac{kT}{e} \frac{1}{I(x)} \frac{dI}{dx}. \tag{21.4-8}$$

Finally, substituting (21.4-8) into (21.4-7) provides an expression for the position-dependent refractive-index change as a function of intensity,

$$\Delta n(x) = -\tfrac{1}{2} n^3 \mathrm{r} \frac{kT}{e} \frac{1}{I(x)} \frac{dI}{dx} . \tag{21.4-9}$$

Refractive-Index Change

This result is readily generalized to two dimensions, whereupon it governs the operation of a photorefractive material as an image storage device.

Many assumptions have been made in an attempt to keep the foregoing theory simple: In deriving (21.4-8) from (21.4-6) it was assumed that the ratio of number densities of unionized to ionized donors is approximately uniform, despite the spatial variation of the photoionization process. This assumption is approximately applicable when the ionization is caused by other more effective processes that are position independent in addition to the light pattern $I(x)$. Dark conductivity and volume photovoltaic effects were neglected. Holes were ignored. It was assumed that no external electric field was applied, when in fact this can be useful in certain applications. The theory is valid only in the steady state even though the time dynamics of the photorefractive process are clearly important since they determine the speed with which the photorefractive material responds to the applied light. Yet, in spite of all these assumptions, the simplified theory captures the essence of the behavior of photorefractive materials.

EXAMPLE 21.4-1. ***Photorefractive Sinusoidal Spatial Intensity Pattern.*** Consider an intensity distribution that takes the form of a sinusoidal function of period Λ, contrast m, and mean intensity I_0,

$$I(x) = I_0 \left(1 + m \cos \frac{2\pi x}{\Lambda}\right), \tag{21.4-10}$$

as illustrated in Fig. 21.4-2. Substituting this expression into (21.4-8) and (21.4-9) yields the internal electric-field and refractive-index patterns

$$E(x) = E_{\max} \frac{-\sin(2\pi x/\Lambda)}{1 + m\cos(2\pi x/\Lambda)} \quad \text{and} \quad \Delta n(x) = \Delta n_{\max} \frac{\sin(2\pi x/\Lambda)}{1 + m\cos(2\pi x/\Lambda)}, \tag{21.4-11}$$

where $E_{\max} = 2\pi(kT/e\Lambda)\, m$ and $\Delta n_{\max} = \frac{1}{2} n^3 \mathrm{r}\, E_{\max}$ are the maximum values of $E(x)$ and $\Delta n(x)$, respectively.

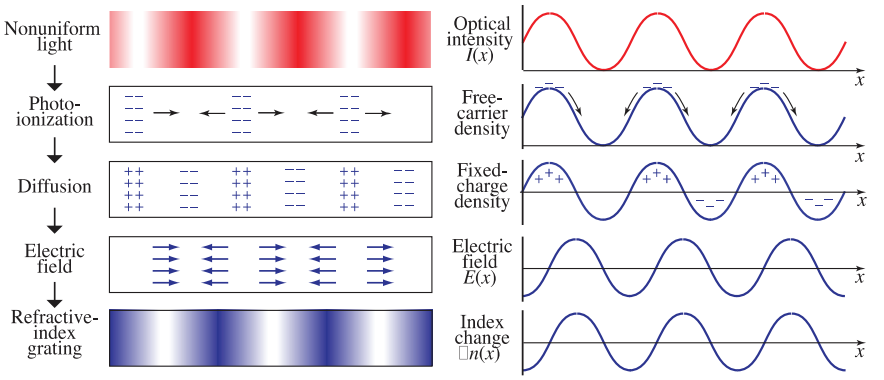


Figure 21.4-2 Response of a photorefractive material to a sinusoidal spatial light pattern.

As a numerical example, if $\Lambda = 1$ μm, $m = 1$, and $T = 300°$ K, we have $E_{\max} = 1.6 \times 10^5$ V/m, an internal field is equivalent to applying 1.6 kV across a crystal of 1-cm width. The

maximum refractive-index change $\Delta n_{\max}$ is directly proportional to the contrast m and the electro-optic coefficient r, and inversely proportional to the spatial period Λ. The grating pattern $\Delta n(x)$ is insensitive to the average intensity I_0. When the image contrast m is small, the second terms in the denominators of (21.4-11) may be neglected, in which case the internal electric field and refractive-index change are simple sinusoidal patterns shifted by 90° relative to the incident light pattern, i.e.,

$$\Delta n(x) \approx \Delta n_{\max} \sin(2\pi x/\Lambda), \tag{21.4-12}$$

as illustrated in Fig. 21.4-2.

Applications of the Photorefractive Effect

An image $I(x, y)$ may be stored in a photorefractive crystal in the form of a refractive-index distribution $\Delta n(x, y)$. The image can be read by using the crystal as a spatial phase modulator to encode the information on a uniform optical plane wave that acts as a probe. Phase modulation may be converted to intensity modulation, for example, by placing the cell in an interferometer (Fig. 21.1-4).

Because of the capability of recording images, photorefractive materials are attractive for use in real-time holography (holography is discussed in Sec. 4.5). As illustrated in Fig. 21.4-3 for two plane waves, an object wave is holographically recorded by mixing it with a reference wave. The intensity of the sum of two such waves forms a sinusoidal interference pattern, which is recorded in the photorefractive crystal in the form of refractive-index variations. The crystal then serves as a volume phase hologram (Fig. 4.5-10). To reconstruct the stored object wave, the crystal is illuminated with the reference wave. Acting as a volume diffraction grating, the crystal reflects the reference wave and reproduces the object wave.

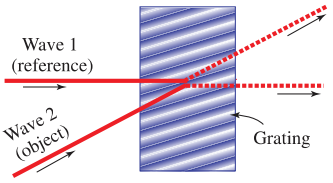


Figure 21.4-3 Two-wave mixing is a form of dynamic holography.

Since the recording process is relatively fast, the processes of recording and reconstruction can be carried out simultaneously. The object and reference waves travel together in the medium and exchange energy via reflection from the created grating, a process called **two-wave mixing**. As shown in Fig. 21.4-3 (see also Fig. 4.5-8), waves 1 and 2 interfere and form a volume grating. Wave 1 reflects from the grating and adds to wave 2; wave 2 reflects from the grating and adds to wave 1. Thus, the two waves are coupled together by the grating they create in the medium. Consequently, the transmission of wave 1 through the medium is controlled by the presence of wave 2, and *vice versa*. For example, wave 1 may be amplified at the expense of wave 2.

The mixing of multiple waves also occurs in other nonlinear optical materials with light-dependent optical properties, as discussed in Chapter 22. Wave mixing has numerous applications in optical data processing (Chapters 22 and 24), including image amplification, the removal of image aberrations, the cross-correlation of images, and optical interconnections.

Electrochromism

Another electro-optic effect involving the transport of charge is **electrochromism**, wherein an electric field causes a change in the absorption spectrum (color) of a material. Applications include displays and electrically controlled windows. Both inorganic and organic materials can exhibit electrochromism.

21.5 ELECTROABSORPTION

Electroabsorption is a change of the absorption characteristics of a medium in response to an externally applied electric field. In a bulk semiconductor, the application of an external electric field results in electron tunneling, which extends the absorption edge into the forbidden gap. The bandgap energy of the material E_g is thus reduced below that provided by the band tail and the Urbach tail, so that $h\nu_2 < h\nu_1$ when the field is ON, as illustrated in Fig. 21.5-1(a). This phenomenon, known as the **Franz–Keldysh effect**, therefore shifts the absorption spectrum to longer wavelengths [Fig. 21.5-1(b)]. The applied electric field also results in the broadening, and ultimate disappearance of, the exciton absorption peaks (Sec. 17.2C).

This effect may be used in optical **electroabsorption modulators** and **electroabsorption switches**, which are technologically simple to implement. In the absence of the electric field (OFF), an incident beam at the operating wavelength, which is longer than the normal bandgap wavelength, is transmitted without absorption [Fig. 21.5-1(b)]. However, upon application of the electric field (ON), the light is absorbed. Such modulators are often constructed in the form of waveguides, with the electric field applied in a direction perpendicular to the direction of propagation of the light beam, as portrayed in Fig. 21.5-1(c). In comparison with electro-optic modulators, which operate on the basis of refractive-index change in response to an externally applied electric field (Secs. 21.1–21.3), electroabsorption modulators typically operate at greater speeds and at lower voltages. Moreover, they can be integrated on the same chip as semiconductor light sources so they are convenient for use in optical fiber communication systems. They also exhibit less chirp than directly modulated laser diodes (Sec. 25.1B).

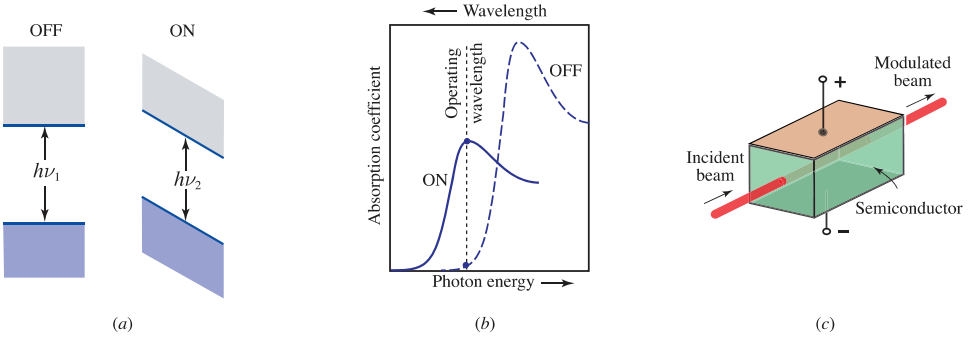


Figure 21.5-1 The Franz–Keldysh effect. (a) The bandgap energy E_g in the absence of an external electric field (OFF) is reduced in its presence (ON). (b) Change in the absorption spectrum caused by the presence of an electric field, which moves the absorption peak toward longer wavelengths. (c) Electroabsorption modulator in a waveguide configuration.

The electroabsorption effect is more pronounced in semiconductor multiquantum-well (MQW) structures (Secs. 17.1G and 18.2D). An electric field applied in the plane of a quantum well gives rise to behavior similar to the Franz–Keldysh effect, including a shift of the absorption edge to a longer wavelength and exciton dissociation. However, an electric field applied in the direction of confinement gives rise to additional phenomena, known collectively as the **quantum-confined Stark effect** (**QCSE**), as illustrated in Fig. 21.5-2:

- The energy difference between the conduction- and valence-band energy levels decreases with increasing electric field ($h\nu_2 < h\nu_1$).
- The band tilt causes the locations of the wavefunctions to shift toward the edges of the well.

- Exciton ionization is inhibited and exciton energy levels remain unbroadened even at high field levels, since the electron and hole remain in proximity by virtue of the confinement.

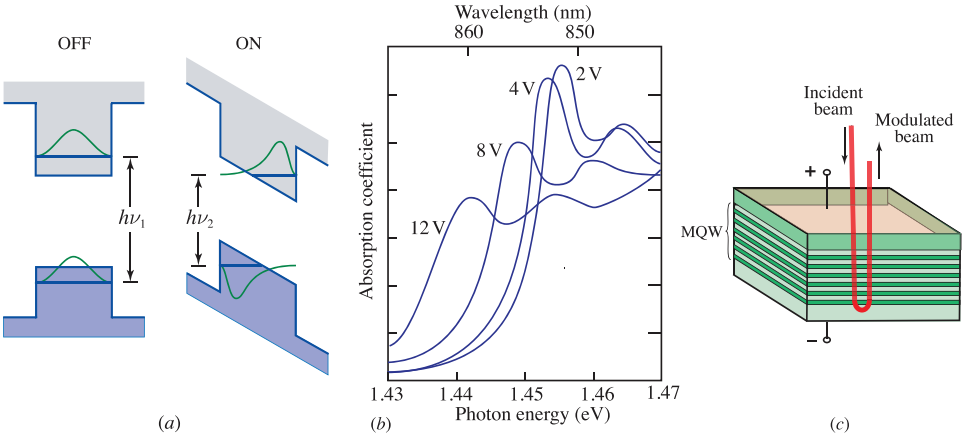


Figure 21.5-2 (a) Energy-band diagrams of a quantum well in the absence (OFF) and presence (ON) of an external electric field applied in the direction of confinement. The field causes the interband energy difference to decrease and the wavefunctions to shift from the centers of the wells toward opposite edges. (b) Absorption spectrum in an AlGaAs/GaAs multiquantum-well structure for various values of the applied voltage (field). The exciton absorption peak moves toward longer wavelengths as the voltage is increased. (Adapted from D. A. B. Miller, D. S. Chemla, T. C. Damen, T. H. Wood, C. A. Burrus, Jr., A. C. Gossard, and W. Wiegmann, The Quantum Well Self-Electrooptic Effect Device: Optoelectronic Bistability and Oscillation, and Self-Linearized Modulation, *IEEE Journal of Quantum Electronics*, vol. 21, pp. 1462–1476, Fig. 1 ©1985 IEEE.) (c) Schematic of a MQW electroabsorption modulator operated in a surface-normal architecture.

As a result of these MQW characteristics, the wavelength shift of the absorption peak is greater, and the absorption edge is more abrupt, than in bulk semiconductors. Electroabsorption modulators based on the QCSE have excellent characteristics, including

- High speeds
- Large extinction ratios
- Low drive voltages
- Low chirp

The simplest transmission implementation directs light through an intrinsic MQW structure sandwiched between p and n regions across which a voltage is applied. Switching is accomplished by simply turning the voltage on and off. A device of this sort can also be fabricated in a waveguide configuration and can be integrated with a distributed-feedback (DFB) laser on a single chip. QCSE modulators and switches can also be fabricated in the form of arrays operated in a double-pass surface-normal architecture, as illustrated in Fig. 21.5-2(c).

EXAMPLE 21.5-1. ***GeSi Electroabsorption Modulator.*** Optical modulators are important elements in integrated photonics. Though silicon has a small electro-optic coefficient, SiGe is suitable for the fabrication of electroabsorption modulators (EAMs). A GeSi EAM with an active region of 1 μm $\times$ 50 μm, integrated on a 3-μm-high silicon-on-insulator waveguide, operates over a 35-nm wavelength range in the vicinity of 1550 nm. The device has a 3-dB bandwidth of 35 GHz and a swing voltage of 2.5 V.

READING LIST

Electro-Optics

See also the reading list on crystals and tensor analysis in Chapter 6 and the reading list in Chapter 22.

T.-C. Poon and T. Kim, *Engineering Optics with MATLAB*, World Scientific, 2nd ed. 2018, Chapter 5.

L. R. Dalton, P. Günter, M. Jazbinsek, O.-P. Kwon, and P. A. Sullivan, *Organic Electro-Optics and Photonics: Molecules, Polymers, and Crystals*, Cambridge University Press, 2015.

R. Dinu, E. Miller, G. Yu, B. Chen, A. Scarpaci, H. Chen, and C. Pilgrim, High-Speed Polymer Optical Modulators, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

G. D. Boreman, *Basic Electro-Optics for Electrical Engineers*, SPIE Optical Engineering Press, 1998.

F. Agulló-López, J. M. Cabrera, and F. Agulló-Rueda, *Electrooptics: Phenomena, Materials and Applications*, Academic Press, 1994.

I. P. Kaminow, *An Introduction to Electrooptic Devices*, Academic Press, 1974.

Liquid-Crystal Devices and Displays

See also the reading list on liquid crystals in Chapter 6.

V. C. Coffey, The Age of OLED Displays, *Optics & Photonics News*, vol. 28, no. 11, pp. 34–41, 2017.

D.-K. Yang and S.-T. Wu, *Fundamentals of Liquid Crystal Devices*, Wiley, 2nd ed. 2015.

S. R. Restaino and S. W. Teare, *Introduction to Liquid Crystals for Optical Design and Engineering*, SPIE Optical Engineering Press, 2015.

R. H. Chen, *Liquid Crystal Displays: Fundamental Physics and Technology*, Wiley, 2011.

P. Yeh and C. Gu, *Optics of Liquid Crystal Displays*, Wiley, 2nd ed. 2010.

W. den Boer, *Active Matrix Liquid Crystal Displays: Fundamentals and Applications*, Elsevier, 2010.

L. Vicari, *Optical Applications of Liquid Crystals*, Institute of Physics, 2003.

V. G. Chigrinov, *Liquid Crystal Devices: Physics and Applications*, Artech, 1999.

U. Efron, ed., *Spatial Light Modulator Technology: Materials, Devices, and Applications*, CRC Press, 1995.

P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Oxford University Press, 2nd ed. 1993.

M. A. Karim, ed., *Electro-Optical Displays*, CRC Press, 1992.

Photorefractive Materials

J. Frejlich, *Photorefractive Materials: Fundamental Concepts, Holographic Recording and Materials Characterization*, Wiley, 2007.

P. Günter and J.-P. Huignard, eds., *Photorefractive Materials and Their Applications. 3: Applications*, Springer-Verlag, 2007.

P. Günter and J.-P. Huignard, eds., *Photorefractive Materials and Their Applications. 2: Materials*, Springer-Verlag, 2007, paperback ed. 2011.

P. Günter and J.-P. Huignard, eds., *Photorefractive Materials and Their Applications. 1: Basic Effects*, Springer-Verlag, 2006, paperback ed. 2010.

F. T. S. Yu and S. Yin, eds., *Photorefractive Optics: Materials, Properties, and Applications*, Academic Press, 2000.

L. Solymar, D. J. Webb, and A. Grunnet-Jepsen, *The Physics and Applications of Photorefractive Materials*, Oxford University Press, 1996.

P. Yeh and C. Gu, eds., *Landmark Papers on Photorefractive Nonlinear Optics*, World Scientific, 1995.

F. M. Davidson, ed., *Selected Papers on Photorefractive Materials*, SPIE Optical Engineering Press (Milestone Series Volume 86), 1994.

P. Yeh, *Introduction to Photorefractive Nonlinear Optics*, Wiley, 1993.

D. M. Pepper, J. Feinberg, and N. K. Kukhtarev, The Photorefractive Effect, *Scientific American*, vol. 263, no. 4, pp. 62–74, 1990.

Electroabsorption

D. Feng, W. Qian, H. Liang, C.-C. Kung, Z. Zhou, Z. Li, J. S. Levy, R. Shafiiha, J. Fong, B. J. Luff, and M. Asghari, High-Speed GeSi Electroabsorption Modulator on the SOI Waveguide Platform, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, 3401710, 2013.

Y.-H. Kuo, Y. K. Lee, Y. Ge, S. Ren, J. E. Roth, T. I. Kamins, D. A. B. Miller, and J. S. Harris, Strong Quantum-Confined Stark Effect in Germanium Quantum-Well Structures on Silicon, *Nature*, vol. 437, pp. 1334–1336, 2005.

PROBLEMS

21.1-2 **Response Time of a Phase Modulator.** A GaAs crystal with refractive index $n = 3.6$ and electro-optic coefficient $\mathrm{r} = 1.6$ pm/V is used as an electro-optic phase modulator operating at $\lambda_o = 1.3$ μm in the longitudinal configuration. The crystal is 3 cm long and has a 1-cm^2 cross-sectional area. Determine the half-wave voltage V_π, the transit time of light through the crystal, and the electrical capacitance of the device (the low-frequency relative permittivity of GaAs is $\epsilon/\epsilon_o = 13.5$). The voltage is applied with a source of 50-Ω resistance. Which factor limits the speed of the device, the transit time of the light through the crystal or the response time of the electrical circuit?

21.1-3 **Sensitivity of an Interferometric Electro-Optic Intensity Modulator.** An integrated-photonic intensity modulator in a Mach–Zehnder configuration, such as that illustrated in Fig. 21.1-5, is used as a linear analog modulator. If the half-wave voltage is $V_\pi = 10$ V, what is the sensitivity of the device (the incremental change of the intensity transmittance per unit incremental change of the applied voltage)?

21.1-4 **An Elasto-Optic Strain Sensor.** An elasto-optic material exhibits a change of the refractive index proportional to the strain. Design a strain sensor based on this effect in the context of an integrated-photonic implementation. If the material is also electro-optic, consider a design based on compensating the elasto-optic and electro-optic refractive index changes against each other, and determining the electric field that nulls the reading of the photodetector in a Mach–Zehnder interferometer configuration.

21.1-5 **Magneto-Optic Modulator.** Describe how a Faraday rotator (see Sec. 6.4B) may be used as an optical intensity modulator.

*21.2-2 **Silica Integrated-Photonic Phase Modulator.** Since bulk fused silica is centrosymmetric, it does not ordinarily exhibit the linear electro-optic (Pockels) effect. However, thermally poled silica has Pockels coefficients that are sufficiently large for use as optical modulators. Determine the phase shift introduced by a poled-silica integrated-photonic phase modulator in a configuration such as that shown in Fig. 21.1-3. Assume that the electrode length is $L = 25$ mm, the electrode separation is $d = 30$ μm, and the wavelength is $\lambda = 1.55$ μm. Assume also that the optical wave is polarized in the y direction, the electric field is created by an applied voltage $V = 400$ V and points in the y direction, and the wave travels along the electrodes in the z direction. The material is poled in a direction such that its principal axes (x_1, x_2, x_3) point in the z, x, and y directions, respectively. The refractive index of the poled material is $n = 1.445$ and the Pockels coefficients are characterized by the matrix

$$\begin{bmatrix} 0 & 0 & \mathrm{r}_{13} \\ 0 & 0 & \mathrm{r}_{13} \\ 0 & 0 & \mathrm{r}_{33} \\ 0 & \mathrm{r}_{13} & 0 \\ \mathrm{r}_{13} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{with } \mathrm{r}_{33} = 0.15 \text{ pm/V}.$$

*21.2-3 **Cascaded Phase Modulators.**

(a) A KDP crystal ($r_{41} = 8$ pm/V, $r_{63} = 11$ pm/V; $n_o = 1.507$, $n_e = 1.467$ at $\lambda_o = 633$ nm) is used as a longitudinal phase modulator. The orientation of the crystal axes and the applied electric field are as shown in Examples 21.2-2 and 21.2-6. Determine the half-wave voltage V_π at $\lambda_o = 633$ nm.

(b) An electro-optic phase modulator consists of 9 KDP crystals separated by electrodes that are biased as shown in Fig. P21.2-3. How should the plates be oriented relative to each other so that the total phase modulation is maximized? Calculate V_π for the composite modulator.

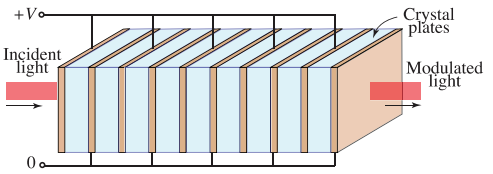


Figure P21.2-3

*21.2-4 **The "Push–Pull" Intensity Modulator.** An optical intensity modulator uses two integrated electro-optic phase modulators and a 3-dB directional coupler, as shown in Fig. P21.2-4. The input wave is split into two waves of equal amplitudes, each of which is phase modulated, reflected from a mirror, phase modulated once again, and the two returning waves are added by the directional coupler to form the output wave. Derive an expression for the intensity transmittance of the device in terms of the applied voltage, wavelength, dimensions, and physical parameters of the phase modulator.

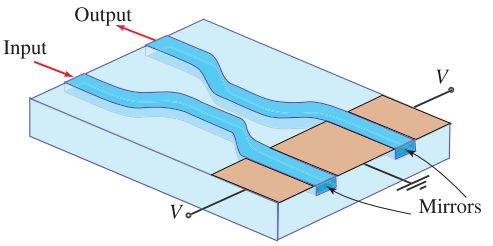


Figure P21.2-4

*21.2-5 **A $LiNbO_3$ Integrated-Photonic Intensity Modulator.** Design a $LiNbO_3$ integrated-photonic intensity modulator using the Mach–Zehnder interferometer shown in Fig. 21.1-5. Select the orientation of the crystal and the polarization of the guided wave to achieve the smallest half-wave voltage V_π. Assume that the active region has length $L = 1$ mm and width $d = 5$ μm. The wavelength is $\lambda_o = 0.85$ μm, the refractive indices are $n_o = 2.29$ and $n_e = 2.17$, and the electro-optic coefficients are $r_{33} = 30.9$, $r_{13} = 8.6$, $r_{22} = 3.4$, and $r_{42} = 28$ pm/V.

*21.2-6 **Double Refraction in an Electro-Optic Crystal.**

(a) An unpolarized He–Ne laser beam ($\lambda_o = 633$ nm) is transmitted through a 1-cm-thick $LiNbO_3$ plate ($n_e = 2.17$, $n_o = 2.29$, $r_{33} = 30.9$ pm/V, $r_{13} = 8.6$ pm/V). The beam is orthogonal to the plate and the optic axis lies in the plane of incidence of the light at $45°$ with the beam. The beam is double refracted (see Sec. 6.3E). Determine the lateral displacement and the retardation between the ordinary and extraordinary beams.

(b) If an electric field $E = 30$ V/m is applied in a direction parallel to the optic axis, what is the effect on the transmitted beams? What are possible applications of this device?

CHAPTER

22

NONLINEAR OPTICS

22.1 NONLINEAR OPTICAL MEDIA 1017

22.2 SECOND-ORDER NONLINEAR OPTICS 1021

A. Second-Harmonic Generation (SHG) and Rectification
B. The Electro-Optic Effect
C. Three-Wave Mixing
D. Phase Matching and Tuning Curves
E. Quasi-Phase Matching

22.3 THIRD-ORDER NONLINEAR OPTICS 1036

A. Third-Harmonic Generation (THG) and Optical Kerr Effect
B. Self-Phase Modulation (SPM), Self-Focusing, and Spatial Solitons
C. Cross-Phase Modulation (XPM)
D. Four-Wave Mixing (FWM)
E. Optical Phase Conjugation (OPC)

*22.4 SECOND-ORDER NONLINEAR OPTICS: COUPLED WAVES 1047

A. Second-Harmonic Generation (SHG)
B. Optical Frequency Conversion (OFC)
C. Optical Parametric Amplification (OPA) and Oscillation (OPO)

*22.5 THIRD-ORDER NONLINEAR OPTICS: COUPLED WAVES 1059

A. Four-Wave Mixing (FWM)
B. Three-Wave Mixing and Third-Harmonic Generation (THG)
C. Optical Phase Conjugation (OPC)

*22.6 ANISOTROPIC NONLINEAR MEDIA 1066

*22.7 DISPERSIVE NONLINEAR MEDIA 1069

Nicolaas Bloembergen (1920–2017) began his seminal studies in nonlinear optics in the early 1960s. He shared the 1981 Nobel Prize with Arthur Schawlow (pictured on p. 657).

Peter A. Franken (1928–1999) provided the first demonstration of optical second-harmonic generation by converting red laser light into ultraviolet light in a quartz crystal.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Throughout the long history of optics, and indeed until relatively recently, it was thought that all optical media were linear. The consequences of this assumption are far-reaching:

- The optical properties of materials, such as refractive index and absorption coefficient, are independent of light intensity.
- The principle of superposition, a fundamental tenet of classical optics, is applicable.
- The frequency of light is never altered by its passage through a medium.
- Two beams of light in the same region of a medium have no effect on each other so that light cannot be used to control light.

The operation of the first laser in 1960 enabled us to examine the behavior of light in optical materials at higher intensities than previously possible. Experiments carried out in the post-laser era clearly demonstrate that optical media do in fact exhibit nonlinear behavior, as exemplified by the following observations:

- The refractive index, and consequently the speed of light in a nonlinear optical medium, does depend on light intensity.
- The principle of superposition is violated in a nonlinear optical medium.
- The frequency of light is altered as it passes through a nonlinear optical medium; the light can change from red to blue, for example.
- Photons do interact within the confines of a nonlinear optical medium so that light can indeed be used to control light.

The field of nonlinear optics offers a host of fascinating phenomena, many of which are also eminently useful.

Nonlinear optical behavior is not observed when light travels in free space. The "nonlinearity" resides in the medium through which the light travels, rather than in the light itself. The interaction of light with light is therefore mediated by the nonlinear medium: the presence of an optical field modifies the properties of the medium, which in turn causes another optical field, or even the original field itself, to be modified.

As discussed in Chapter 5, the properties of a dielectric medium through which an optical electromagnetic wave propagates are described by the relation between the polarization-density vector $\mathcal{P}(\mathbf{r}, t)$ and the electric-field vector $\mathcal{E}(\mathbf{r}, t)$. Indeed it is useful to view $\mathcal{P}(\mathbf{r}, t)$ as the output of a system whose input is $\mathcal{E}(\mathbf{r}, t)$. The mathematical relation between the vector functions $\mathcal{P}(\mathbf{r}, t)$ and $\mathcal{E}(\mathbf{r}, t)$, which is governed by the characteristics of the medium, defines the system. The medium is said to be nonlinear if this relation is nonlinear (see Sec. 5.2).

This Chapter

In Chapter 5, dielectric media were further classified with respect to their dispersiveness, inhomogeneity, and anisotropy (see Sec. 5.2). To focus on the principal effect of interest — nonlinearity — the first portion of our exposition is restricted to a medium that is nondispersive, homogeneous, and isotropic. The vectors $\mathcal{P}$ and $\mathcal{E}$ are consequently parallel at every position and time and may therefore be examined on a component-by-component basis.

The theory of nonlinear optics and its applications is presented at two levels. A simplified approach is provided in Secs. 22.1–22.3. This is followed by a more detailed analysis of the same phenomena in Sec. 22.4 and Sec. 22.5.

The propagation of light in media characterized by a second-order (quadratic) nonlinear relation between $\mathcal{P}$ and $\mathcal{E}$ is described in Sec. 22.2 and Sec. 22.4. Applications include the frequency doubling of a monochromatic wave (*second-harmonic generation*), the mixing of two monochromatic waves to generate a third wave at their sum or difference frequencies (*frequency conversion*), the use of two monochromatic waves

to amplify a third wave (*parametric amplification*), and the incorporation of feedback in a parametric-amplification device to create an oscillator (*parametric oscillation*). Wave propagation in a medium with a third-order (cubic) relation between $\mathcal{P}$ and $\mathcal{E}$ is discussed in Secs. 22.3 and 22.5. Applications include *third-harmonic generation*, *self-phase modulation*, *self-focusing*, *four-wave mixing*, and *optical phase conjugation*. The behavior of anisotropic and dispersive nonlinear optical media is briefly considered in Secs. 22.6 and 22.7, respectively.

Nonlinear Optics in Other Chapters

A principal assumption of the treatment provided in this chapter is that the nonlinear optical medium is passive, i.e., it does not exchange energy with the light wave(s). Waves of different frequencies may exchange energy with each another via the nonlinear property of the medium, but their total energy is conserved. This class of nonlinear phenomena, known as **parametric interactions**, are so-named because a *parameter* of the system is varied periodically in time; a strong electric field, for example, can cause the electric susceptibility to oscillate in time. Several nonlinear phenomena involving **nonparametric interactions** are described in other chapters of this book:

- *Laser interactions*. The interactions of light with a medium at frequencies near atomic or molecular resonances involve phenomena such as absorption, and stimulated and spontaneous emission, as described in Sec. 14.3. These interactions become nonlinear when the light is sufficiently intense so that the populations of the various energy levels are significantly altered. Nonlinear optical effects are manifested in the saturation of laser amplifiers and saturable absorbers (Sec. 15.4).
- *Multiphoton absorption*. Intense light can induce the absorption of multiple photons whose total energy matches that of an atomic transition. For k-photon absorption, the rate of absorption is proportional to I^k, where I is the optical intensity. This nonlinear optical phenomenon is described briefly in Sec. 14.5B.
- *Nonlinear scattering*. Nonlinear inelastic scattering involves the interaction of light with the vibrational or acoustic modes of a medium. Examples include stimulated Raman and stimulated Brillouin scattering, as described in Secs. 14.5C and 15.3D.

It is also assumed throughout this chapter that the light is described by stationary continuous waves. Nonstationary nonlinear optical phenomena include:

- *Nonlinear optics of pulsed light*. The parametric interaction of optical pulses with a nonlinear medium is described in Sec. 23.5.
- *Optical solitons* are light pulses that travel over exceptionally long distances through nonlinear dispersive media without changing their width or shape. This nonlinear phenomenon is the result of a balance between dispersion and nonlinear self-phase modulation, as described in Sec. 23.5B. The use of solitons in optical fiber communication systems is described in Sec. 25.2E.

Yet another nonlinear optical effect is *optical bistability*. This involves nonlinear optical effects together with feedback. The use of optical bistability in photonic logic gates is described in Sec. 24.4.

22.1 NONLINEAR OPTICAL MEDIA

A linear dielectric medium is characterized by a linear relation between the polarization density and the electric field, $\mathcal{P} = \epsilon_o \chi \mathcal{E}$, where ϵ_o is the permittivity of free space and χ is the electric susceptibility of the medium (see Sec. 5.2A). A nonlinear dielectric medium, on the other hand, is characterized by a nonlinear relation between $\mathcal{P}$ and $\mathcal{E}$ (see Sec. 5.2B), as illustrated in Fig. 22.1-1.

The nonlinearity may be of microscopic or macroscopic origin. The polarization

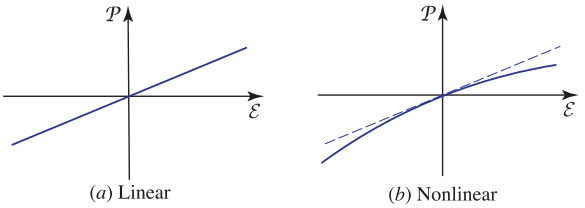


Figure 22.1-1 The $\mathcal{P}$–$\mathcal{E}$ relation for (a) a linear dielectric medium, and (b) a nonlinear medium.

density $\mathcal{P} = N\mathrm{p}$ is a product of the individual dipole moment p induced by the applied electric field $\mathcal{E}$ and the number density of dipole moments N. The nonlinear behavior may reside either in p or in N.

The relation between p and $\mathcal{E}$ is linear when $\mathcal{E}$ is small, but becomes nonlinear when $\mathcal{E}$ acquires values comparable to interatomic electric fields, which are typically $\sim 10^5$–10^8 V/m. This may be understood in terms of a simple Lorentz oscillator model in which the dipole moment is $\mathrm{p} = -ex$, where x is the displacement of a mass of charge $-e$ to which an electric force $-e\mathcal{E}$ is applied (see Sec. 5.5C). If the restraining elastic force is proportional to the displacement (i.e., if Hooke's law is satisfied), the equilibrium displacement x is proportional to $\mathcal{E}$. In that case $\mathcal{P}$ is proportional to $\mathcal{E}$ and the medium is linear. However, if the restraining force is a nonlinear function of the displacement, the equilibrium displacement x and the polarization density $\mathcal{P}$ are nonlinear functions of $\mathcal{E}$ and, consequently, the medium is nonlinear. The time dynamics of an anharmonic oscillator model describing a dielectric medium with these features is discussed in Sec. 22.7.

Another possible origin of a nonlinear response of an optical material to light is the dependence of the number density N on the optical field. An example is provided by a laser medium in which the number of atoms occupying the energy levels involved in the absorption and emission of light are dependent on the intensity of the light itself (see Sec. 15.4).

Since externally applied optical electric fields are typically small in comparison with characteristic interatomic or crystalline fields, even when focused laser light is used, the nonlinearity is usually weak. The relation between $\mathcal{P}$ and $\mathcal{E}$ is then approximately linear for small $\mathcal{E}$, deviating only slightly from linearity as $\mathcal{E}$ increases (see Fig. 22.1-1). Under these circumstances, the function that relates $\mathcal{P}$ to $\mathcal{E}$ can be expanded in a Taylor series about $\mathcal{E} = 0$,

$$\mathcal{P} = a_1\mathcal{E} + \tfrac{1}{2}a_2\mathcal{E}^2 + \tfrac{1}{6}a_3\mathcal{E}^3 + \cdots, \tag{22.1-1}$$

and it suffices to use only a few terms. The coefficients a_1, a_2, and a_3 are, respectively, the first, second, and third derivatives of $\mathcal{P}$ with respect to $\mathcal{E}$, evaluated at $\mathcal{E} = 0$. These coefficients are characteristic constants of the medium. The first term, which is linear, dominates at small $\mathcal{E}$. Clearly, $a_1 = \epsilon_o\chi$, where χ is the linear susceptibility, which is related to the relative permittivity and the refractive index of the material via $n^2 = \epsilon/\epsilon_o = 1 + \chi$ [see (5.2-13)]. The second term represents a quadratic or second-order nonlinearity, while the third term represents a third-order nonlinearity, and so on.

It is customary to write (22.1-1) in the form[†]

$$\boxed{\mathcal{P} = \epsilon_o\chi\mathcal{E} + 2\mathrm{d}\mathcal{E}^2 + 4\chi^{(3)}\mathcal{E}^3 + \cdots,} \tag{22.1-2}$$

[†] An alternative form of this relation, $\mathcal{P} = \epsilon_o(\chi\mathcal{E} + \chi^{(2)}\mathcal{E}^2 + \chi^{(3)}\mathcal{E}^3)$, is also widely used.

where $\mathrm{d} = \frac{1}{4}a_2$ and $\chi^{(3)} = \frac{1}{24}a_3$ are **nonlinear optical coefficients** that serve to describe the strength of the second- and third-order nonlinear effects, respectively.

Equation (22.1-2) provides the essential mathematical characterization of a nonlinear optical medium. Material dispersion, inhomogeneity, and anisotropy have not been accommodated, both for the sake of simplicity and to enable us to focus on the essential features of nonlinear optical behavior. Sections 22.6 and 22.7 are devoted to anisotropic and dispersive nonlinear media, respectively.

In centrosymmetric media, which have inversion symmetry so that the properties of the medium are not altered by the transformation $\mathbf{r} \rightarrow -\mathbf{r}$, the $\mathcal{P}$–$\mathcal{E}$ function must have odd symmetry, so that the reversal of $\mathcal{E}$ results in the reversal of $\mathcal{P}$ without any other change. The second-order nonlinear optical coefficient d must then vanish, and the lowest order nonlinearity is of third order.

Typical values of the second-order nonlinear optical coefficient d for dielectric crystals, semiconductors, and organic materials used in photonics applications lie in the range $\mathrm{d} = 10^{-24}$–10^{-21} ($\mathrm{C/V^2}$ in MKS units). Typical values of the third-order nonlinear optical coefficient $\chi^{(3)}$ for glasses, crystals, semiconductors, semiconductor-doped glasses, and organic materials of interest in photonics are in the vicinity of $\chi^{(3)} = 10^{-34}$–10^{-29} ($\mathrm{Cm/V^3}$ in MKS units). Biased or asymmetric quantum wells offer large nonlinearities in the mid and far infrared.

EXERCISE 22.1-1

Intensity of Light Required to Elicit Nonlinear Effects.

(a) Determine the light intensity (in $\mathrm{W/cm^2}$) at which the ratio of the second term to the first term in (22.1-2) is 1% in an ADP ($\mathrm{NH_4H_2PO_4}$) crystal for which $n = 1.5$ and $\mathrm{d} = 6.8 \times 10^{-24}$ $\mathrm{C/V^2}$ at $\lambda_o = 1.06$ μm.

(b) Determine the light intensity at which the third term in (22.1-2) is 1% of the first term in carbon disulfide ($\mathrm{CS_2}$) for which $n = 1.6$, $\mathrm{d} = 0$, and $\chi^{(3)} = 4.4 \times 10^{-32}$ $\mathrm{Cm/V^3}$ at $\lambda_o = 694$ nm.

Note: In accordance with (5.4-8), the light intensity is $I = |E_0|^2/2\eta = \langle \mathcal{E}^2 \rangle / \eta$, where $\eta = \eta_o/n$ is the impedance of the medium and $\eta_o = \sqrt{\mu_o/\epsilon_o} \approx 377\ \Omega$ is the impedance of free space (see Sec. 5.4).

The Nonlinear Wave Equation

The propagation of light in a nonlinear medium is governed by the wave equation (5.2-25), which was derived from Maxwell's equations for an arbitrary homogeneous, isotropic dielectric medium. The isotropy of the medium ensures that the vectors $\mathcal{P}$ and $\mathcal{E}$ are always parallel so that they may be examined on a component-by-component basis, which allows us to write (5.2-25) as

$$\nabla^2 \mathcal{E} - \frac{1}{c_o^2} \frac{\partial^2 \mathcal{E}}{\partial t^2} = \mu_o \frac{\partial^2 \mathcal{P}}{\partial t^2}. \tag{22.1-3}$$

It is convenient to write the polarization density in (22.1-2) as a sum of linear ($\epsilon_o \chi \mathcal{E}$) and nonlinear ($\mathcal{P}_{\mathrm{NL}}$) parts,

$$\mathcal{P} = \epsilon_o \chi \mathcal{E} + \mathcal{P}_{\mathrm{NL}}, \tag{22.1-4}$$

$$\mathcal{P}_{\mathrm{NL}} = 2\mathrm{d}\mathcal{E}^2 + 4\chi^{(3)}\mathcal{E}^3 + \cdots. \tag{22.1-5}$$

Using (22.1-4), along with the relations $c = c_o/n$, $n^2 = 1 + \chi$, and $c_o = 1/\sqrt{\epsilon_o \mu_o}$ provided in (5.2-12) and (5.2-13), allows (22.1-3) to be written as

$$\nabla^2 \mathcal{E} - \frac{1}{c^2}\frac{\partial^2 \mathcal{E}}{\partial t^2} = -\mathcal{S} \tag{22.1-6}$$

$$\mathcal{S} = -\mu_o \frac{\partial^2 \mathcal{P}_{\text{NL}}}{\partial t^2}. \tag{22.1-7}$$

Wave Equation in Nonlinear Medium

It is convenient to regard (22.1-6) as a wave equation in which the term $\mathcal{S}(t)$ is regarded as a source that radiates in a linear medium of refractive index n. Because $\mathcal{P}_{\text{NL}}$ (and therefore $\mathcal{S}$) is a nonlinear function of $\mathcal{E}$, (22.1-6) is a nonlinear partial differential equation in $\mathcal{E}$. This is the basic equation that underlies the theory of nonlinear optics.

Two approximate approaches to solving this nonlinear wave equation can be called upon. The first is the iterative approach known as the Born approximation. This approximation underlies the simplified introduction to nonlinear optics presented in Secs. 22.2 and 22.3. The second approach is a coupled-wave theory in which the nonlinear wave equation is used to derive approximate linear coupled partial differential equations that govern the interacting waves. This is the basis of the more advanced study of wave interactions in nonlinear media presented in Secs. 22.4 and 22.5.

Scattering Theory of Nonlinear Optics: The Born Approximation

The radiation source $\mathcal{S}$ in (22.1-6) is a function of the field $\mathcal{E}$ that it, itself, radiates. To emphasize this point we write $\mathcal{S} = \mathcal{S}(\mathcal{E})$ and illustrate the process by the simple block diagram in Fig. 22.1-2. Suppose that an optical field $\mathcal{E}_0$ is incident on a nonlinear medium confined to some volume, as shown in the figure. This field creates a radiation source $\mathcal{S}(\mathcal{E}_0)$ that radiates an optical field $\mathcal{E}_1$. The corresponding radiation source $\mathcal{S}(\mathcal{E}_1)$ radiates a field $\mathcal{E}_2$, and so on. This process suggests an iterative solution, the first step of which is known as the **first Born approximation**. The second Born approximation carries the process an additional step, and so on. The first Born approximation is

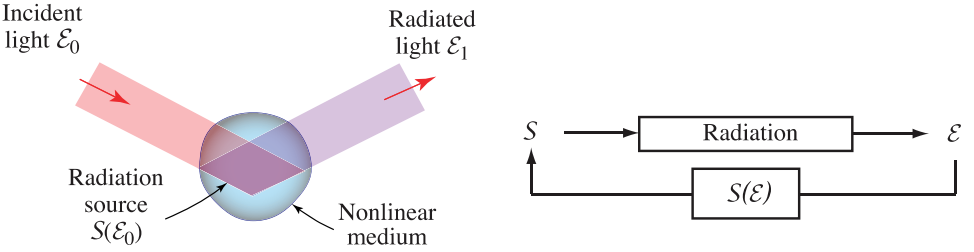


Figure 22.1-2 The first Born approximation. An incident optical field $\mathcal{E}_0$ creates a source $\mathcal{S}(\mathcal{E}_0)$, which radiates an optical field $\mathcal{E}_1$.

adequate when the light intensity is sufficiently weak so that the nonlinearity is small. In this approximation, light propagation through the nonlinear medium is regarded as a scattering process in which the incident field is scattered by the medium. The scattered light is determined from the incident light in two steps:

1. The incident field $\mathcal{E}_0$ is used to determine the nonlinear polarization density $\mathcal{P}_{\text{NL}}$, from which the radiation source $\mathcal{S}(\mathcal{E}_0)$ is determined.
2. The radiated (scattered) field $\mathcal{E}_1$ is determined from the radiation source by adding the spherical waves associated with the different source points (as in the theory of scattering discussed in Sec. 5.6).

The development presented in Secs. 22.2 and 22.3 are based on the first Born approximation. The initial field $\mathcal{E}_0$ is assumed to contain one or several monochromatic waves of different frequencies. The corresponding nonlinear polarization density $\mathcal{P}_{\text{NL}}$ is then determined using (22.1-5) and the source function $\mathcal{S}(\mathcal{E}_0)$ is evaluated using (22.1-7). Since $\mathcal{S}(\mathcal{E}_0)$ is a nonlinear function, new frequencies are created and the source emits an optical field $\mathcal{E}_1$ with frequencies not present in the original wave $\mathcal{E}_0$. This leads to numerous interesting phenomena that have been utilized to make useful nonlinear optics devices.

22.2 SECOND-ORDER NONLINEAR OPTICS

In this section we examine the optical properties of a nonlinear medium in which nonlinearities of order higher than the second are negligible, so that

$$\boxed{\mathcal{P}_{\text{NL}} = 2\text{d}\mathcal{E}^2.} \tag{22.2-1}$$

A material for which (22.2-1) is applicable is called a **second-order nonlinear medium**; more colloquially it is also known as a **chi-two medium** since $\mathcal{P}_{\text{NL}} = 2\text{d}\mathcal{E}^2 = \epsilon_o\chi^{(2)}\mathcal{E}^2$ (see footnote on page 1018).

We proceed to consider an electric field $\mathcal{E}$ comprising one or two harmonic components and determine the spectral components of $\mathcal{P}_{\text{NL}}$. In accordance with the first Born approximation, the radiation source $\mathcal{S}$ contains the same spectral components as $\mathcal{P}_{\text{NL}}$ and, therefore, so too does the emitted (scattered) field.

A. Second-Harmonic Generation (SHG) and Rectification

Consider first the response of this nonlinear medium to a single harmonic electric field of angular frequency ω (wavelength $\lambda_o = 2\pi c_o/\omega$) and complex amplitude $E(\omega)$:

$$\mathcal{E}(t) = \text{Re}\{E(\omega)\exp(j\omega t)\} = \tfrac{1}{2}[E(\omega)\exp(j\omega t) + E^*(\omega)\exp(-j\omega t)]. \tag{22.2-2}$$

The corresponding nonlinear polarization density $\mathcal{P}_{\text{NL}}$ is obtained by substituting (22.2-2) into (22.2-1),

$$\mathcal{P}_{\text{NL}}(t) = P_{\text{NL}}(0) + \text{Re}\{P_{\text{NL}}(2\omega)\ \exp(j2\omega t)\} \tag{22.2-3}$$

where

$$P_{\text{NL}}(0) = \text{d}\,E(\omega)E^*(\omega) \tag{22.2-4}$$

$$P_{\text{NL}}(2\omega) = \text{d}\,E^2(\omega). \tag{22.2-5}$$

This process is graphically illustrated in Fig. 22.2-1.

Second-Harmonic Generation (SHG)

The source $\mathcal{S}(t) = -\mu_o\partial^2\mathcal{P}_{\text{NL}}/\partial t^2$ corresponding to (22.2-3) has a component at frequency 2ω with complex amplitude $S(2\omega) = 4\mu_o\omega^2\,\text{d}E(\omega)E(\omega)$, which radiates an optical field at frequency 2ω (wavelength $\lambda_o/2$). Thus, the scattered optical field has a component at the second harmonic of the incident optical field. Since the amplitude of the emitted second-harmonic light is proportional to $S(2\omega)$, its intensity $I(2\omega)$ is proportional to $|S(2\omega)|^2$, which in turn is proportional to the square of the intensity of the incident wave $I(\omega) = |E(\omega)|^2/2\eta$ and to the square of the nonlinear coefficient d. Also, since the emissions are added coherently, the intensity of the second-harmonic wave is proportional to the square of the length of the interaction region L.

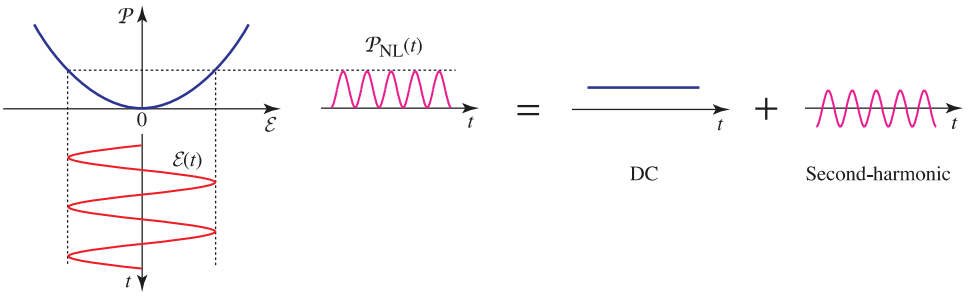


Figure 22.2-1 A sinusoidal electric field of angular frequency ω in a second-order nonlinear optical medium creates a polarization density with a component at 2ω (second-harmonic) and a steady (DC) component.

The efficiency of second-harmonic generation $\eta_{\mathrm{SHG}} = I(2\omega)/I(\omega)$ is therefore proportional to $L^2 I(\omega)$. Since $I(\omega) = P/A$, where P is the incident power and A is the cross-sectional area of the interaction volume, the SHG efficiency is often expressed in the form

$$\boxed{\eta_{\mathrm{SHG}} = C^2 \frac{L^2}{A} P,} \qquad \begin{matrix}(22.2\text{-}6)\\ \text{SHG Efficiency}\end{matrix}$$

where C^2 is a constant (units of W^{-1}) proportional to d^2 and ω^2. An expression for C^2 will be provided in (22.4-36).

In accordance with (22.2-6), to maximize the SHG efficiency it is essential that the incident wave have the largest possible power P. This is accomplished by the use of pulsed lasers for which the energy is confined in time, so that large peak powers are obtained. Additionally, to maximize the ratio L^2/A, the wave must be focused to the smallest possible area A and experience the longest possible interaction length L. For a thin crystal, L is determined by the length of the crystal so that the beam should be focused to the smallest spot area A [Fig. 22.2-2(a)]. If the dimensions of the nonlinear medium are not limiting factors, however, the maximum value of L for a given area A is limited by beam diffraction. For example, a Gaussian beam focused to a beam width W_0 maintains a beam cross-sectional area $A \approx \pi W_0^2$ over a depth of focus $L = 2z_0 = 2\pi W_0^2/\lambda$ [see (3.1-22)] so that the ratio $L^2/A = 2L/\lambda = 4A/\lambda^2$. In this case, the beam should be focused to the largest spot size, corresponding to the largest depth of focus. The efficiency is then proportional to L. For a thick crystal, therefore, the beam should be focused to the largest spot that fits within the cross-sectional area of the crystal [Fig. 22.2-2(b)].

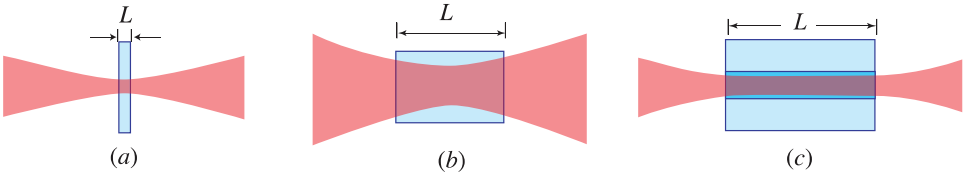


Figure 22.2-2 SHG interaction volumes. (a) For a thin crystal, minimize A. (b) For a thick crystal, maximize A. (c) For an optical waveguide, maximize L.

Guided-wave structures offer the advantage of light confinement in a small cross-sectional area over long lengths. Since A is determined by the size of the guided mode, the efficiency is proportional to L^2 [Fig. 22.2-2(c)]. Optical waveguides take the form of planar or channel waveguides (Chapter 9) or fibers (Chapter 10). Though silica-glass fibers were initially ruled out for second-harmonic generation since glass is centrosymmetric (and therefore presumably has $\mathrm{d} = 0$), second-harmonic generation is in fact observed in silica-glass fibers, an effect attributed to electric-quadrupole and magnetic-dipole interactions and to defects and color centers in the fiber core.

Figure 22.2-3 displays several experimental configurations for generating optical second-harmonic generation in bulk materials and in waveguides, in which visible light is converted to ultraviolet light and infrared light is converted to visible light.

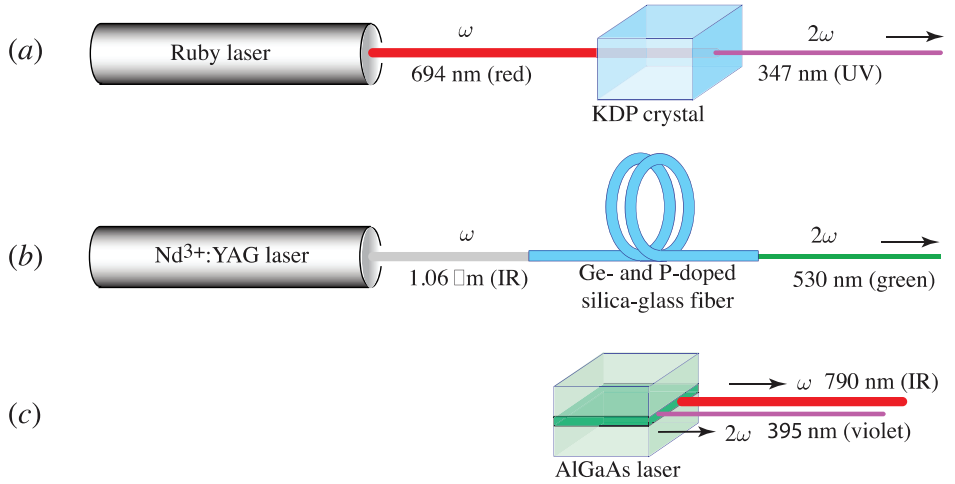


Figure 22.2-3 Optical second-harmonic generation (a) in a bulk crystal; (b) in a doped silica-glass fiber; (c) within the cavity of a laser diode.

Optical Rectification

The component $P_{\mathrm{NL}}(0)$ in (22.2-3) and (22.2-4) corresponds to a steady (non-time-varying) polarization density that creates a DC potential difference across the plates of a capacitor within which the nonlinear material is placed (Fig. 22.2-4). The generation of a DC voltage as a result of an intense optical field represents optical rectification (in analogy with the conversion of a sinusoidal AC voltage into a DC voltage in an ordinary electronic rectifier). An optical pulse with a peak power of several MW, for example, may generate a voltage of several hundred μV.

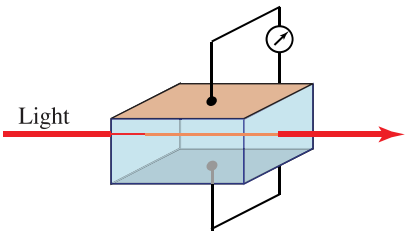


Figure 22.2-4 The transmission of an intense beam of light through a second-order nonlinear crystal generates a DC voltage across it.

B. The Electro-Optic Effect

We now consider an electric field $\mathcal{E}(t)$ comprising a harmonic component at an optical frequency ω together with a steady component (at $\omega = 0$),

$$\mathcal{E}(t) = E(0) + \text{Re}\{E(\omega)\ \exp(j\omega t)\}. \tag{22.2-7}$$

We distinguish between these two components by denoting the electric field $E(0)$ and the optical field $E(\omega)$. In fact, both components are electric fields.

Substituting (22.2-7) into (22.2-1), we obtain

$$\mathcal{P}_{\text{NL}}(t) = P_{\text{NL}}(0) + \text{Re}\{P_{\text{NL}}(\omega)\exp(j\omega t)\} + \text{Re}\{P_{\text{NL}}(2\omega)\exp(j2\omega t)\}, \tag{22.2-8}$$

where

$$P_{\text{NL}}(0) = \text{d}\left[2E^2(0) + |E(\omega)|^2\right] \tag{22.2-9a}$$

$$P_{\text{NL}}(\omega) = 4\text{d}\,E(0)E(\omega) \tag{22.2-9b}$$

$$P_{\text{NL}}(2\omega) = \text{d}\,E^2(\omega), \tag{22.2-9c}$$

so that the polarization density contains components at the angular frequencies 0, ω, and 2ω.

If the optical field is substantially smaller in magnitude than the electric field, i.e., $|E(\omega)|^2 \ll |E(0)|^2$, the second-harmonic polarization-density component $P_{\text{NL}}(2\omega)$ is negligible in comparison with the components $P_{\text{NL}}(0)$ and $P_{\text{NL}}(\omega)$. This is equivalent to the linearization of $\mathcal{P}_{\text{NL}}$ as a function of $\mathcal{E}$, i.e., approximating it by a straight line with a slope equal to the derivative at $\mathcal{E} = E(0)$, as illustrated in Fig. 22.2-5.

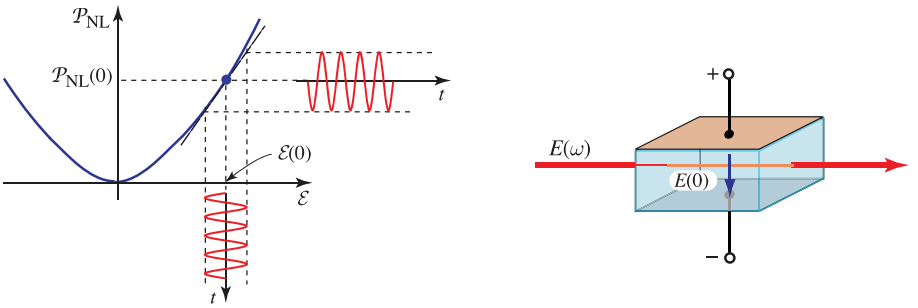


Figure 22.2-5 Linearization of the second-order nonlinear relation $\mathcal{P}_{\text{NL}} = 2\text{d}\mathcal{E}^2$ in the presence of a strong electric field $E(0)$ and a weak optical field $E(\omega)$.

Equation (22.2-9b) provides a linear relation between $P_{\text{NL}}(\omega)$ and $E(\omega)$, which we write in the form $P_{\text{NL}}(\omega) = \epsilon_o \Delta\chi E(\omega)$, where $\Delta\chi = (4\text{d}/\epsilon_o)E(0)$ represents an increase in the susceptibility proportional to the electric field $E(0)$. The corresponding incremental change of the refractive index is obtained by differentiating the relation $n^2 = 1 + \chi$, to obtain $2n\,\Delta n = \Delta\chi$, from which

$$\Delta n = \frac{2\text{d}}{n\epsilon_o}E(0). \tag{22.2-10}$$

The medium is then effectively linear with a refractive index $n + \Delta n$ that is linearly controlled by the electric field $E(0)$.

The nonlinear nature of the medium creates a coupling between the electric field $E(0)$ and the optical field $E(\omega)$, causing one to control the other, so that the nonlinear medium exhibits the linear electro-optic effect (Pockels effect) discussed in Chapter 21.

This effect is characterized by the relation $\Delta n = -\frac{1}{2}n^3 \mathrm{r} E(0)$, where r is the Pockels coefficient. Comparing this formula with (22.2-10), we conclude that the Pockels coefficient r is related to the second-order nonlinear optical coefficient d by

$$\mathrm{r} \approx -\frac{4}{\epsilon_o n^4}\mathrm{d}. \tag{22.2-11}$$

Though this expression reveals the common underlying origin of the Pockels effect and the medium nonlinearity, it is not consistent with experimentally observed values of r and d. This is because we have made the implicit assumption that the medium is nondispersive (i.e., that its response is insensitive to frequency). This assumption is clearly not satisfied when one of the components of the field is at the optical frequency ω and the other is a steady field with zero frequency. The role of dispersion is discussed in Sec. 22.7.

C. Three-Wave Mixing

We now consider the case of a field $\mathcal{E}(t)$ comprising two harmonic components at optical frequencies ω_1 and ω_2,

$$\mathcal{E}(t) = \mathrm{Re}\{E(\omega_1)\,\exp(j\omega_1 t) + E(\omega_2)\,\exp(j\omega_2 t)\}. \tag{22.2-12}$$

(The spatial features of these waves will be considered shortly.) The nonlinear component of the polarization density $\mathcal{P}_{\mathrm{NL}} = 2\mathrm{d}\mathcal{E}^2$ then contains components at five frequencies, 0, $2\omega_1$, $2\omega_2$, $\omega_+ = \omega_1 + \omega_2$, and $\omega_- = \omega_1 - \omega_2$, with amplitudes

$$P_{\mathrm{NL}}(0) = \mathrm{d}\left[|E(\omega_1)|^2 + |E(\omega_2)|^2\right] \tag{22.2-13a}$$

$$P_{\mathrm{NL}}(2\omega_1) = \mathrm{d}\,E(\omega_1)E(\omega_1) \tag{22.2-13b}$$

$$P_{\mathrm{NL}}(2\omega_2) = \mathrm{d}\,E(\omega_2)E(\omega_2) \tag{22.2-13c}$$

$$P_{\mathrm{NL}}(\omega_+) = 2\mathrm{d}\,E(\omega_1)E(\omega_2) \tag{22.2-13d}$$

$$P_{\mathrm{NL}}(\omega_-) = 2\mathrm{d}\,E(\omega_1)E^*(\omega_2). \tag{22.2-13e}$$

Thus, the second-order nonlinear medium can be used to mix two optical waves of different frequencies and generate (among other things) a third wave at the difference frequency or at the sum frequency. The former process is called **frequency downconversion** whereas the latter is known as **frequency up-conversion** or **sum-frequency generation**. An example of frequency up-conversion is provided in Fig. 22.2-6: the light from two lasers with free-space wavelengths $\lambda_{o1} = 1.06\,\mu$m and $\lambda_{o2} = 10.6\,\mu$m enter a proustite crystal and generate a third wave with wavelength $\lambda_{o3} = 0.96\,\mu$m (where $\lambda_{o3}^{-1} = \lambda_{o1}^{-1} + \lambda_{o2}^{-1}$).

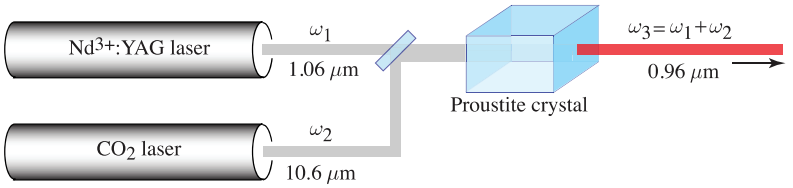


Figure 22.2-6 An example of sum-frequency generation (SFG), also called frequency up-conversion, in a nonlinear crystal.

Though the incident pair of waves at frequencies ω_1 and ω_2 produce polarization densities at frequencies 0, $2\omega_1$, $2\omega_2$, $\omega_1+\omega_2$, and $\omega_1-\omega_2$, all of these waves are not necessarily generated, since certain additional conditions (phase matching) must be satisfied, as explained presently.

Frequency and Phase Matching

If waves 1 and 2 are plane waves with wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$, so that $E(\omega_1) = A_1\exp(-j\mathbf{k}_1\cdot\mathbf{r})$ and $E(\omega_2) = A_2\exp(-j\mathbf{k}_2\cdot\mathbf{r})$, then in accordance with (22.2-13d), $P_{\text{NL}}(\omega_3) = 2\text{d}E(\omega_1)E(\omega_2) = 2\text{d}A_1A_2\exp(-j\mathbf{k}_3\cdot\mathbf{r})$, where

$$\boxed{\omega_1+\omega_2=\omega_3} \qquad \begin{matrix}(22.2\text{-}14)\\ \text{Frequency-Matching Condition}\end{matrix}$$

and

$$\boxed{\mathbf{k}_1+\mathbf{k}_2=\mathbf{k}_3.} \qquad \begin{matrix}(22.2\text{-}15)\\ \text{Phase-Matching Condition}\end{matrix}$$

The medium therefore acts as a light source of frequency $\omega_3=\omega_1+\omega_2$, with a complex amplitude proportional to $\exp(-j\mathbf{k}_3\cdot\mathbf{r})$, so that it radiates a wave of wavevector $\mathbf{k}_3 = \mathbf{k}_1+\mathbf{k}_2$, as illustrated in Fig. 22.2-7. Equation (22.2-15) can be regarded as a condition of phase matching among the wavefronts of the three waves that is analogous to the frequency-matching condition $\omega_1+\omega_2=\omega_3$. Since the argument of the complex wavefunction is $\omega t-\mathbf{k}\cdot\mathbf{r}$, these two conditions ensure both the temporal and spatial phase matching of the three waves, which is necessary for their sustained mutual interaction over extended durations of time and regions of space.

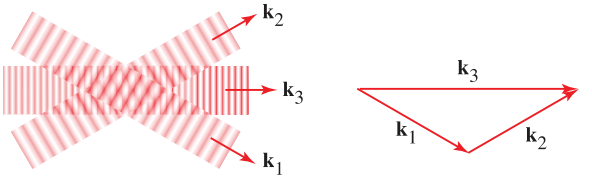


Figure 22.2-7 The phase-matching condition.

Three-Wave Mixing Modalities

When two optical waves of angular frequencies ω_1 and ω_2 travel through a second-order nonlinear optical medium they mix and produce a polarization density with components at a number of frequencies. We assume that only the component at the sum frequency $\omega_3=\omega_1+\omega_2$ satisfies the phase-matching condition. Other frequencies cannot be sustained by the medium since they are assumed not to satisfy the phase-matching condition.

Once wave 3 is generated, it interacts with wave 1 and generates a wave at the difference frequency $\omega_2=\omega_3-\omega_1$. Clearly, the phase-matching condition for this interaction is also satisfied. Waves 3 and 2 similarly combine and radiate at ω_1. The three waves therefore undergo mutual coupling in which each pair of waves interacts and contributes to the third wave. The process is called **three-wave mixing**.

Two-wave mixing is not, in general, possible. Two waves of arbitrary frequencies ω_1 and ω_2 cannot be coupled by the medium without the help of a third wave. Two-wave mixing can occur only in the degenerate case, $\omega_2=2\omega_1$, in which the second-harmonic of wave 1 contributes to wave 2; and the subharmonic $\omega_2/2$ of wave 2, which is at the frequency difference $\omega_2-\omega_1$, contributes to wave 1.

Three-wave mixing is known as a **parametric interaction** process. It takes a variety of forms, depending on which of the three waves is provided as an input, and which are extracted as outputs, as illustrated in the following examples (see Fig. 22.2-8):

- **Optical Frequency Conversion (OFC).** Waves 1 and 2 are mixed in an **up-converter**, generating a wave at the sum frequency $\omega_3 = \omega_1 + \omega_2$. This process, also called **sum-frequency generation (SFG)**, has already been illustrated in Fig. 22.2-6. Second-harmonic generation (SHG) is a degenerate special case of SFG. The opposite process of **downconversion** or **difference-frequency generation (DFG)** is realized by an interaction between waves 3 and 1 to generate wave 2, at the difference frequency $\omega_2 = \omega_3 - \omega_1$. Up- and down-converters are used to generate coherent light at wavelengths where no adequate lasers are available, and as optical mixers in optical communication systems.
- **Optical Parametric Amplifier (OPA).** Waves 1 and 3 interact so that wave 1 grows, and in the process an auxiliary wave 2 is created. The device operates as a coherent amplifier at frequency ω_1 and is known as an OPA. Wave 3, called the **pump**, provides the required energy, whereas wave 2 is known as the **idler** wave. The amplified wave is called the **signal**. Clearly, the gain of the amplifier depends on the power of the pump. OPAs are used for the detection of weak light at wavelengths for which sensitive detectors are not available.
- **Optical Parametric Oscillator (OPO)** . With proper feedback, the parametric amplifier can operate as a parametric oscillator, in which only a pump wave is supplied. OPOs are used for the generation of coherent light and mode-locked pulse trains over a continuous range of frequencies, usually in frequency bands where there is a paucity of tunable laser sources.
- **Spontaneous Parametric Downconversion (SPDC)** . Here, the only input to the nonlinear crystal is the pump wave 3, and downconversion to the lower-frequency waves 1 and 2 is spontaneous. The frequency- and phase-matching conditions (22.2-14) and (22.2-15) lead to multiple solutions, each forming a pair of waves 1 and 2 with specific frequencies and directions. The downconverted light takes the form of a cone of multispectral light, as illustrated in Fig. 22.2-8.

Further details pertaining to these parametric devices are provided in Sec. 22.4.

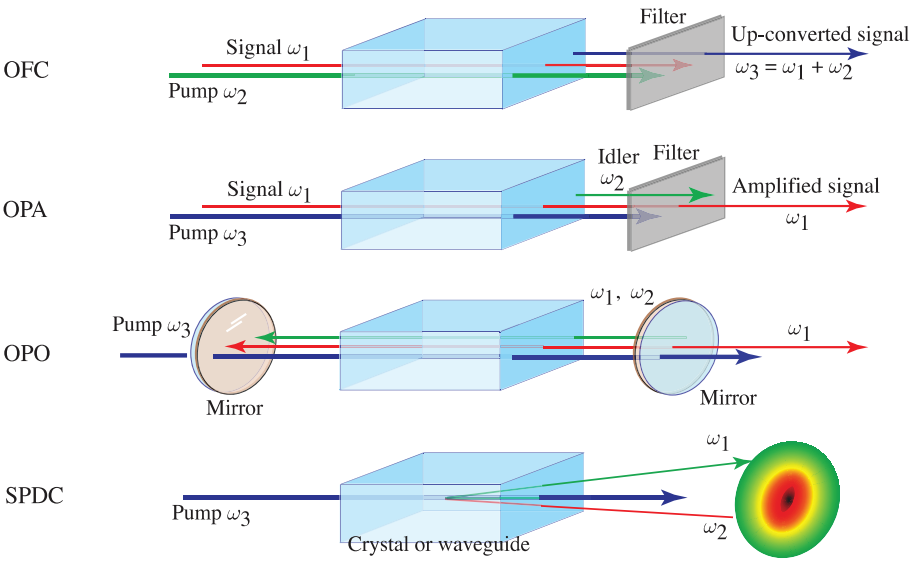


Figure 22.2-8 Optical parametric devices in bulk crystals or integrated waveguides: optical frequency converter (OFC); optical parametric amplifier (OPA); optical parametric oscillator (OPO); spontaneous parametric downconverter (SPDC). Fiber lasers, laser diodes, quantum cascade lasers, and diode-pumped solid-state lasers often serve as pumps for optical parametric devices.

Wave Mixing as a Photon Interaction Process

The three-wave-mixing process can be viewed from a photon-optics perspective as a process of three-photon interaction in which two photons of lower frequency, ω_1 and ω_2, are annihilated, and a photon of higher frequency ω_3 is created, as illustrated in Fig. 22.2-9(a). Alternatively, the annihilation of a photon of high frequency ω_3 is accompanied by the creation of two low-frequency photons, of frequencies ω_1 and ω_2, as illustrated in Fig. 22.2-9(b). Since $\hbar\omega$ and $\hbar\mathbf{k}$ are the energy and momentum of a photon of frequency ω and wavevector $\mathbf{k}$ (see Sec. 13.1), conservation of energy and momentum, in either case, requires that

$$\hbar\omega_1 + \hbar\omega_2 = \hbar\omega_3 \tag{22.2-16}$$

$$\hbar\mathbf{k}_1 + \hbar\mathbf{k}_2 = \hbar\mathbf{k}_3, \tag{22.2-17}$$

where $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$ are the wavevectors of the three photons. The frequency- and phase-matching conditions presented in (22.2-14) and (22.2-15) are thus reproduced.

The energy diagram for the three-photon-mixing process displayed in Fig. 22.2-9(b) bears some similarity to that for an optically pumped three-level laser, illustrated in Fig. 22.2-9(c) (see Sec. 15.2B). There are significant distinctions between the two processes, however:

- One of the three transitions involved in the laser process is non-radiative.
- An exchange of energy between the field and medium takes place in the laser process.
- The energy levels associated with the laser process are relatively sharp and are established by the atomic or molecular system, whereas the energy levels of the parametric process are dictated by photon energy and phase-matching conditions and are tunable over wide spectral regions.

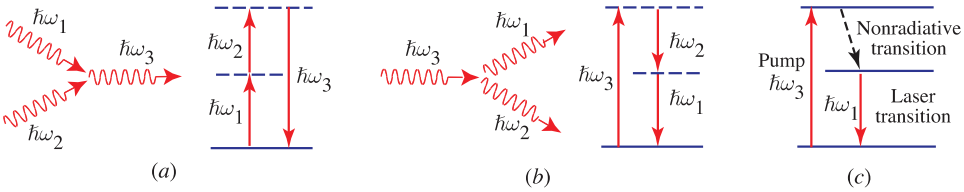


Figure 22.2-9 Comparison of parametric processes in a second-order nonlinear medium and laser action. (a) Annihilation of two low-frequency photons and creation of one high-frequency photon. Dashed lines indicate virtual states. (b) Annihilation of one high-frequency photon and creation of two low-frequency photons. (c) Optically pumped 3-level laser, a nonparametric process in which the medium participates in energy transfer.

The process of wave mixing involves an energy exchange among the interacting waves. Clearly, energy must be conserved, as is assured by the frequency-matching condition, $\omega_1 + \omega_2 = \omega_3$. Photon numbers must also be conserved, consistent with the photon interaction. Consider the photon-splitting process represented in Fig. 22.2-9(b). If $\Delta\Phi_1$, $\Delta\Phi_2$, and $\Delta\Phi_3$ are the net changes in the photon fluxes (photons per second) in the course of the interaction (the flux of photons leaving minus the flux of photons entering) at frequencies ω_1, ω_2, and ω_3, then $\Delta\Phi_1 = \Delta\Phi_2 = -\Delta\Phi_3$, so that for each of the ω_3 photons lost, one each of the ω_1 and ω_2 photons is gained.

If the three waves travel in the same direction, the z direction for example, then by taking a cylinder of unit area and incremental length $\Delta z \to 0$ as the interaction volume, we conclude that the photon-flux densities ϕ_1, ϕ_2, ϕ_3 (photons/s-m^2) of the three waves must satisfy

$$\frac{d\phi_1}{dz} = \frac{d\phi_2}{dz} = -\frac{d\phi_3}{dz}. \tag{22.2-18}$$

(22.2-18) Photon-Number Conservation

Since the wave intensities (W/m^2) are $I_1 = \hbar\omega_1\phi_1$, $I_2 = \hbar\omega_2\phi_2$, and $I_3 = \hbar\omega_3\phi_3$, (22.2-18) gives

$$\frac{d}{dz}\left(\frac{I_1}{\omega_1}\right) = \frac{d}{dz}\left(\frac{I_2}{\omega_2}\right) = -\frac{d}{dz}\left(\frac{I_3}{\omega_3}\right). \tag{22.2-19}$$

(22.2-19) Manley–Rowe Relations

Equations (22.2-19) are known as the Manley–Rowe relations. It was first derived in the context of wave interactions in nonlinear electronic systems. The Manley–Rowe relations can be derived using wave optics, without invoking the concept of the photon (Exercise 22.4-2).

D. Phase Matching and Tuning Curves

Phase Matching in Collinear Three-Wave Mixing

If the mixed three waves are collinear, i.e., they travel in the same direction, and if the medium is nondispersive, then the phase-matching condition (22.2-15) yields the scalar equation $n\omega_1/c_o + n\omega_2/c_o = n\omega_3/c_o$, which is automatically satisfied if the frequency matching condition $\omega_1 + \omega_2 = \omega_3$ is met. However, since all materials are in actuality dispersive, the three waves actually travel at different velocities corresponding to different refractive indices, n_1, n_2, and n_3, and the frequency- and phase-matching conditions are independent:

$$\omega_1 + \omega_2 = \omega_3, \qquad \omega_1 n_1 + \omega_2 n_2 = \omega_3 n_3, \tag{22.2-20}$$

(22.2-20) Matching Conditions

and must be simultaneously satisfied. Since this is usually not possible, birefringence, which is a feature of anisotropic media, is often used to compensate dispersion.

For an anisotropic medium, the three refractive indices n_1, n_2, and n_3 are generally dependent on the polarizations of the waves and their directions relative to the principal axes (see Sec. 6.3C). This offers other degrees of freedom to satisfy the matching conditions. Precise control of the refractive indices at the three frequencies is often achieved by appropriate selection of polarization, orientation of the crystal, and in some cases by temperature control.

For an optical waveguide, the phase-matching condition (22.2-15) should be replaced with $\beta_1 + \beta_2 = \beta_3$, which relates the propagation constants of the waveguide modes at the wavelengths of the three mixed waves. These propagation constants depend on the refractive indices of the waveguide material, the polarization (TE or TM), and the waveguide geometry and dimensions (see Sec. 9.2A). These additional degrees of freedom offer more flexibility in satisfying the phase-matching condition.

In practice, the medium is often a uniaxial crystal characterized by its optic axis and frequency-dependent ordinary and extraordinary refractive indices $n_o(\omega)$ and $n_e(\omega)$. Each of the three waves can be ordinary (o) or extraordinary (e) and the process is labeled accordingly. For example, the label e-o-o indicates that waves 1, 2, and 3 are e, o, and o waves, respectively. For an o wave, $n(\omega) = n_o(\omega)$; for an e wave, $n(\omega) = n(\theta, \omega)$ depends on the angle θ between the direction of the wave and the optic axis of the crystal, in accordance with the relation

$$\frac{1}{n^2(\theta,\omega)} = \frac{\cos^2\theta}{n_o^2(\omega)} + \frac{\sin^2\theta}{n_e^2(\omega)}, \tag{22.2-21}$$

which is represented graphically by an ellipse [see (6.3-15) and Fig. 6.3-7]. If the polarizations of the signal and idler waves are the same, the wave mixing is said to be **Type-I**; if they are orthogonal, it is said to be **Type-II**.

EXAMPLE 22.2-1. ***Collinear Type-I Second-Harmonic Generation (SHG).*** For SHG, waves 1 and 2 have the same frequency ($\omega_1 = \omega_2 = \omega$) and $\omega_3 = 2\omega$. For Type-I mixing, waves 1 and 2 have identical polarizations so that $n_1 = n_2$. Therefore, from (22.2-20), the phase-matching condition is $n_3 = n_1$, i.e., the fundamental wave has the same refractive index as the second-harmonic wave. Because of dispersion, this condition cannot usually be satisfied unless the polarizations of these two waves are different. For a uniaxial crystal, the process is either o-o-e or e-e-o. In either case, the direction at which the wave enters the crystal is adjusted in such a way that $n_3 = n_1$, i.e., such that birefringence compensates exactly for dispersion.

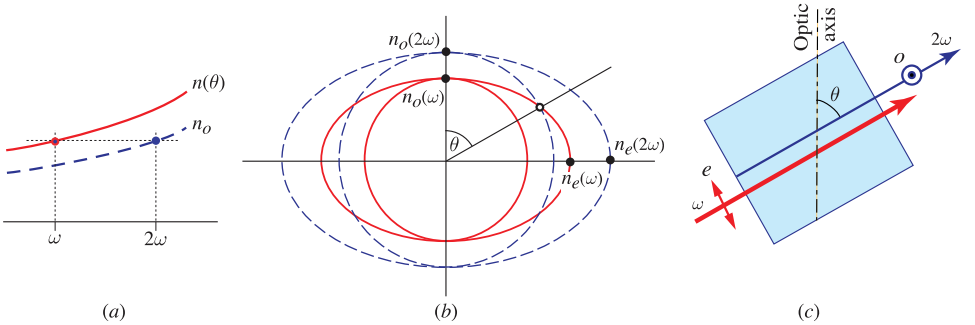


Figure 22.2-10 Phase matching in e-e-o SHG. (a) Matching the index of the e wave at ω with that of the o wave at 2ω. (b) Index surfaces at ω (solid curves) and 2ω (dashed curves) for a uniaxial crystal. (c) The wave is chosen to travel at an angle θ with respect to the crystal optic axis, such that the extraordinary refractive index $n_e(\theta,\omega)$ of the ω wave equals the ordinary refractive index $n_o(2\omega)$ of the 2ω wave.

For an e-e-o process such as that illustrated in Fig. 22.2-10, the fundamental wave is extraordinary and the second-harmonic wave is ordinary, $n_1 = n(\theta,\omega)$ and $n_3 = n_o(2\omega)$, so that the matching condition is: $n(\theta,\omega) = n_o(2\omega)$. This is achieved by selecting an angle θ for which

$$\boxed{n(\theta,\omega) = n_o(2\omega),} \tag{22.2-22}$$

SHG Type-I e-e-o

where $n(\theta,\omega)$ is given by (22.2-21). This is illustrated graphically in Fig. 22.2-10, which displays the ordinary and extraordinary refractive indices (a circle and an ellipse) at ω (solid curves) and at 2ω (dashed curves). The angle at which phase matching is satisfied is that at which the circle at 2ω intersects the ellipse at ω.

As an example, for KDP at a fundamental wavelength λ = 694 nm, $n_o(\omega)$ = 1.506, $n_e(\omega)$ = 1.466; and at $\lambda/2$ = 347 nm, $n_o(2\omega)$ = 1.534, $n_e(2\omega)$ = 1.490. In this case, (22.2-22) and (22.2-21) gives $\theta = 52°$. This is called the cut angle of the crystal. Similar equations may be written for SHG in the o-o-e configuration. In this case, for KDP at a fundamental wavelength $\lambda = 1.06$ μm, $\theta = 41°$.

EXAMPLE 22.2-2. ***Collinear Optical Parametric Oscillator (OPO).*** The oscillation frequencies of an OPO are determined from the frequency and phase-matching conditions. For a Type-I o-o-e mixing configuration,

$$\boxed{\omega_1 + \omega_2 = \omega_3, \qquad \omega_1 n_o(\omega_1) + \omega_2 n_o(\omega_2) = \omega_3 n(\theta,\omega_3).} \tag{22.2-23}$$

OPO Type-I o-o-e

For Type-II e-o-e mixing,

$$\omega_1 + \omega_2 = \omega_3, \qquad \omega_1 n(\theta, \omega_1) + \omega_2 n_o(\omega_2) = \omega_3 n(\theta, \omega_3). \tag{22.2-24}$$

OPO Type-II e-o-e

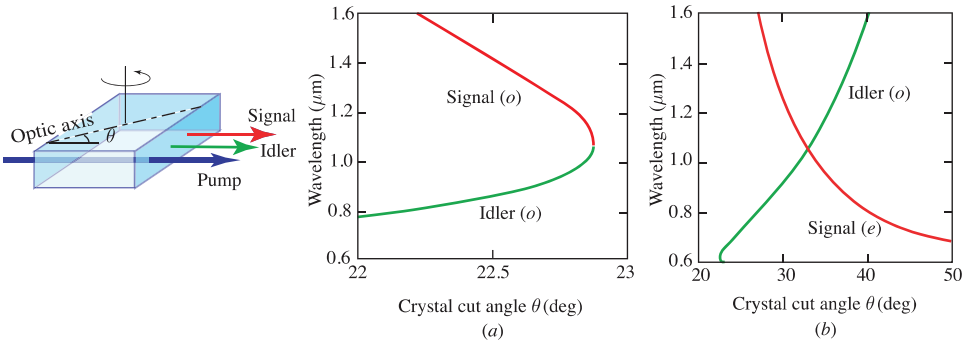


Figure 22.2-11 Tuning curves for a collinear OPO using a BBO crystal and a 532-nm pump, which is readily obtained from a frequency doubled Nd:YAG laser (a) Type-I, and (b) Type-II.

The functions $n_o(\omega)$ and $n_e(\omega)$ are determined from the Sellmeier equation (5.5-28), and the extraordinary index $n(\theta, \omega)$ is determined as a function of the angle θ between the optic axis of the crystal and the direction of the waves by use of (22.2-21). For a given pump frequency ω_3, the solutions of (22.2-23) and (22.2-24), ω_1 and ω_2, are often plotted versus the angle θ, a plot known as the tuning curve. Examples are illustrated in Fig. 22.2-11.

Phase Matching in Non-Collinear Three-Wave Mixing

In the non-collinear case, the phase-matching condition $\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3$ is equivalent to $\omega_1 n_1 \hat{u}_1 + \omega_2 n_2 \hat{u}_2 = \omega_3 n_3 \hat{u}_3$, where $\hat{u}_1$, $\hat{u}_2$, and $\hat{u}_3$ are unit vectors in the directions of propagation of the waves. The refractive indices n_1, n_2, and n_3 depend on the directions of the waves relative to the crystal axes, as well as the polarizations and frequencies. This vector equation is equivalent to two scalar equations so that the matching conditions become

$$\omega_1 + \omega_2 = \omega_3, \quad \omega_1 n_1 \sin\theta_1 = \omega_2 n_2 \sin\theta_2, \quad \omega_1 n_1 \cos\theta_1 + \omega_2 n_2 \cos\theta_2 = \omega_3 n_3, \tag{22.2-25}$$

where θ_1 and θ_2 are the angles waves 1 and 2 make with wave 3. The design of a 3-wave mixing device centers about the selection of directions and polarizations to satisfy these equations, as demonstrated by the following exercise and example.

EXERCISE 22.2-1

Non-Collinear Type-II Second-Harmonic Generation (SHG). Figure 22.2-12 illustrates Type-II o-e-e non-collinear SHG. An ordinary wave and an extraordinary wave, both at the fundamental frequency ω, create an extraordinary second-harmonic wave at the frequency 2ω. It is assumed here that the directions of propagation of the three waves and the optic axis are coplanar and the two fundamental waves and the optic axis make angles θ_1, θ_2, and θ with the direction of the second-harmonic wave. The refractive indices that appear in the phase-matching equations (22.2-25) are $n_1 = n_o(\omega)$, $n_2 = n(\theta + \theta_2, \omega)$, and $n_3 = n(\theta, 2\omega)$, i.e.,

$$n_o(\omega)\sin\theta_1 = n(\theta+\theta_2,\omega)\sin\theta_2, \qquad n_o(\omega)\cos\theta_1 + n(\theta+\theta_2,\omega)\cos\theta_2 = 2n(\theta,2\omega). \tag{22.2-26}$$

SHG Type-II o-e-e

For a KDP crystal and a fundamental wave of wavelength 1.06 μm (Nd^{3+}:YAG laser), determine the crystal orientation and the angles θ_1 and θ_2 for efficient second-harmonic generation.

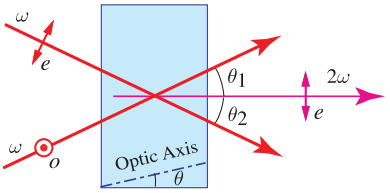


Figure 22.2-12 Non-collinear Type-II second-harmonic generation.

EXAMPLE 22.2-3. ***Spontaneous Parametric Downconversion (SPDC).*** In SPDC, a pump wave of frequency ω_3 creates pairs of waves 1 and 2, at frequencies ω_1 and ω_2, and angles θ_1 and θ_2, all satisfying the frequency- and phase-matching conditions (22.2-25). For example, in the Type-I o-o-e case, $n_1 = n_o(\omega_1)$, $n_2 = n_o(\omega_2)$ and $n_3 = n(\theta, \omega_3)$. These relations together with the Sellmeier equations for $n_o(\omega)$ and $n_e(\omega)$ yield a continuum of solutions (ω_1, θ_1), (ω_2, θ_2) for the signal and idler waves, as illustrated by the example in Fig. 22.2-13.

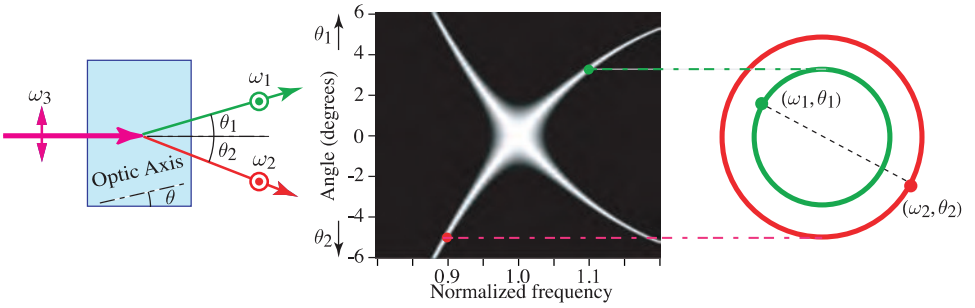


Figure 22.2-13 Tuning curves for non-collinear Type-I o-o-e spontaneous parametric downconversion in a BBO crystal at an angle $\theta = 33.53°$ for a 351.5-nm pump (from an Ar^+-ion laser). Each point in the bright area of the middle picture represents the frequency ω_1 and angle θ_1 of a possible downconverted wave, and has a matching point at a complementary frequency $\omega_2 = \omega_3 - \omega_1$ with angle θ_2. Frequencies are normalized to the degenerate frequency $\omega_o = \omega_3/2$. For example, the two dots shown represent a pair of downconverted waves at frequencies $0.9\omega_o$ and $1.1\omega_o$. Because of circular symmetry, each point is actually a ring of points all of the same frequency, but each point on a ring matches only one diametrically opposite point on the corresponding ring, as illustrated in the right graph.

Tolerable Phase Mismatch and Coherence Length

A slight phase mismatch $\Delta\mathbf{k} = \mathbf{k}_3 - \mathbf{k}_1 - \mathbf{k}_2 \neq 0$ may result in a significant reduction in the wave-mixing efficiency. If waves 1 and 2 are plane waves with wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$, so that $E(\omega_1) = A_1 \exp(-j\mathbf{k}_1 \cdot \mathbf{r})$ and $E(\omega_2) = A_2 \exp(-j\mathbf{k}_2 \cdot \mathbf{r})$, then in accordance with (22.2-13d), $P_{\text{NL}}(\omega_3) = 2\mathrm{d}E(\omega_1)E(\omega_2) = 2\mathrm{d}A_1A_2 \exp[-j(\mathbf{k}_1 + \mathbf{k}_2) \cdot \mathbf{r}] = 2\mathrm{d}A_1A_2 \exp(j\Delta\mathbf{k} \cdot \mathbf{r}) \exp(-j\mathbf{k}_3 \cdot \mathbf{r})$. By virtue of (22.1-7) this creates a source with angular frequency ω_3, wavevector $\mathbf{k}_3$, and complex amplitude $2\omega_3^2\mu_o\mathrm{d}A_1A_2 \exp(j\Delta\mathbf{k} \cdot \mathbf{r})$. It can be shown (Prob. 22.2-6) that the intensity of the generated wave is proportional to the squared integral of the source amplitude over the interaction volume V,

$$I_3 \propto \left| \int_V \mathrm{d}A_1A_2 \exp(j\Delta\mathbf{k} \cdot \mathbf{r}) \mathrm{d}\mathbf{r} \right|^2. \tag{22.2-27}$$

Because the contributions of different points within the interaction volume are added as phasors, the position-dependent phase $\Delta\mathbf{k} \cdot \mathbf{r}$ in the phase mismatched case results in a reduction of the total intensity below the value obtained in the matched case. Consider

the special case of a one-dimensional interaction volume of length L in the z direction: $I_3 \propto |\int_0^L \exp(j\Delta k\, z)dz|^2 = L^2\,\text{sinc}^2(\Delta k\, L/2\pi)$, where Δk is the z component of $\Delta\mathbf{k}$ and $\text{sinc}(x) = \sin(\pi x)/(\pi x)$. It follows that in the presence of a wavevector mismatch Δk, I_3 is reduced by the factor $\text{sinc}^2(\Delta k L/2\pi)$, which is unity for $\Delta k = 0$ and drops as Δk increases, reaching a value of $(2/\pi)^2 \approx 0.4$ when $|\Delta k| = \pi/L$, and vanishing when $|\Delta k| = 2\pi/L$ (Fig. 22.2-14). For a given value of L, the mismatch Δk corresponding to a prescribed efficiency reduction factor is inversely proportional to L, so that the phase-matching requirement becomes more stringent as L increases. For a given mismatch Δk, the length

$$L_c = 2\pi/|\Delta k| \tag{22.2-28}$$

Coherence Length

is a measure of the maximum length within which the parametric interaction process is efficient; L_c is often called the **wave-mixing coherence length**.

For example, for a second-harmonic generation $|\Delta k| = 2(2\pi/\lambda_o)|n_3 - n_1|$, where λ_o is the free-space wavelength of the fundamental wave and n_1 and n_3 are the refractive indices of the fundamental and the second-harmonic waves. In this case, $L_c = \lambda_o/2|n_3 - n_1|$ is inversely proportional to $|n_3 - n_1|$, which is governed by the material dispersion. For example, for $|\ n_3 - n_1\ | = 10^{-2}$, $L_c = 50\lambda$.

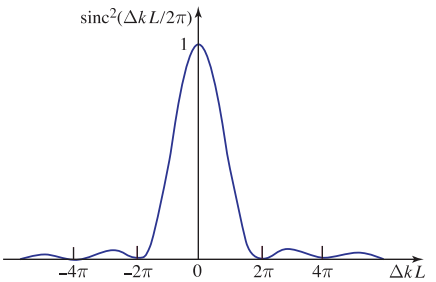


Figure 22.2-14 The factor by which the efficiency of three-wave mixing is reduced as a result of a phase mismatch $\Delta k\, L$ between waves interacting within a distance L.

The tolerance of the interaction process to the phase mismatch can be regarded as a result of the wavevector uncertainty $\Delta k \propto 1/L$ associated with confinement of the waves within a distance L [see (A.2-6) in Appendix A]. The corresponding momentum uncertainty $\Delta p = \hbar\Delta k \propto 1/L$ explains the apparent violation of the law of conservation of momentum in the wave-mixing process.

Phase-Matching Bandwidth

As previously noted, for a finite interaction length L, a phase mismatch $|\Delta k| \leq 2\pi/L$ is tolerated. If exact phase matching is achieved at a set of nominal frequencies of the mixed waves, then small frequency deviations from those values may be tolerated, as long as the condition $\omega_1 + \omega_2 = \omega_3$ is perfectly satisfied. The spectral bands associated with such tolerance are established by the condition $|\Delta k| \leq 2\pi/L$.

As an example, in SHG we have two waves with frequencies $\omega_1 = \omega$ and $\omega_3 = 2\omega$. The mismatch Δk is a function $\Delta k(\omega)$ of the fundamental frequency ω. The device is designed for exact phase matching at a nominal fundamental frequency ω_0, i.e., $\Delta k(\omega_0) = 0$. The bandwidth $\Delta\omega$ is then established by the condition $|\Delta k(\omega_0 + \Delta\omega)| = 2\pi/L$. If $\Delta\omega$ is sufficiently small, we may write $\Delta k(\omega_0 + \Delta\omega) = \Delta k'\Delta\omega$, where $\Delta k' = (d/d\omega)\Delta k$ at ω_0. Therefore, $\Delta\omega = 2\pi/|\Delta k'|L$, from which the spectral width in Hz is

$$\Delta\nu = 1/|\Delta k'|L. \tag{22.2-29}$$

Phase-Matching Bandwidth

Since $\Delta k(\omega) = k_3(2\omega) - 2k_1(\omega)$, the derivative $\Delta k' = dk_3(2\omega)/d\omega - 2dk_1(\omega)/d\omega = 2[dk_3(2\omega)/d(2\omega) - dk_1(\omega)/d\omega] = 2[1/v_3 - 1/v_1]$, where v_1 and v_3 are the group velocities of waves 1 and 3 at frequencies ω and 2ω, respectively (see Sec. 5.7). The spectral width is therefore related to the length L and the group velocity mismatch by

$$\Delta\nu = \frac{1}{2}\left|\frac{L}{v_3} - \frac{L}{v_1}\right|^{-1} = \frac{c_o}{2L}\frac{1}{|N_3 - N_1|}, \tag{22.2-30}$$

Phase-Matching Bandwidth

where N_1 and N_3 are the group indices of the material at the fundamental and second-harmonic frequencies.

It is apparent that second-harmonic generation of a broadband wave, or an ultranarrow pulse (see Sec. 23.5A), can be accomplished by use of a thin crystal (at a cost of lower conversion efficiency), and by the use of an additional design constraint, group velocity matching, $v_3 \approx v_1$ or $N_3 \approx N_1$. Phase-matching tolerance in SPDC is revealed in Fig. 22.2-13 by the thickness of the curves.

E. Quasi-Phase Matching

In the presence of a wavevector mismatch $\Delta\mathbf{k}$, points within the interaction volume radiate with position-dependent phases $\Delta\mathbf{k}\cdot\mathbf{r}$, so that the magnitude of the generated parametric wave is significantly reduced. Since phase matching can be difficult to achieve, or can severely constrain the choice of the nonlinear optical coefficient or the crystal configuration that maximizes the efficiency of wave conversion, one approach is to allow a phase mismatch, but to compensate it by using a medium with position-dependent periodic nonlinearity. Such periodicity introduces an opposite phase that brings back the phases of the distributed radiation elements into better alignment. The technique is called **quasi-phase matching (QPM)**.

If the medium has a position-dependent nonlinear optical coefficient $\mathrm{d}(\mathbf{r})$, then (22.2-27) becomes

$$I_3 \propto \left|\int_V \mathrm{d}(\mathbf{r})\exp(j\Delta\mathbf{k}\cdot\mathbf{r})\,d\mathbf{r}\right|^2. \tag{22.2-31}$$

If $\mathrm{d}(\mathbf{r})$ is a harmonic function $\mathrm{d}(\mathbf{r}) = \mathrm{d}_o\exp(-j\mathbf{G}\cdot\mathbf{r})$, with $\mathbf{G} = \Delta\mathbf{k}$, then the phase mismatch is fully eliminated. Accordingly, the phase-matching condition (22.2-15) is replaced with

$$\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{G} = \mathbf{k}_3. \tag{22.2-32}$$

In effect, the nonlinear medium serves as a phase grating (or longitudinal Bragg grating) with a wavevector $\mathbf{G}$.

It is generally difficult to fabricate a medium with a continuously varying harmonic nonlinear optical coefficient, $\mathrm{d}(\mathbf{r}) = \mathrm{d}_o\exp(-j\mathbf{G}\cdot\mathbf{r})$, but it is possible to fabricate simpler periodic structures, e.g., media with nonlinear optical coefficients of constant magnitude but periodically reversed sign. Since any periodic function can be decomposed into a superposition of harmonic functions via Fourier series, one such function can serve to correct the phase mismatch, with the others playing no role in the wave-mixing process because they introduce greater phase mismatch.

QPM in Collinear Wave Mixing

For collinear waves traveling in the z direction and having a phase mismatch Δk, the required phase grating is of the form $\exp(-jGz)$, where $G = \Delta k$. Such a grating

may be obtained by use of a periodic nonlinear optical coefficient $\mathrm{d}(z)$ described by the Fourier series $\mathrm{d}(z) = \sum_{m=-\infty}^{\infty} \mathrm{d}_m \exp(-j2\pi mz/\Lambda)$, where Λ is the period and $\{\mathrm{d}_m\}$ are the Fourier coefficients. Any of these components may be used for phase matching. For example, for the mth harmonic, $G = m2\pi/\Lambda = \Delta k$, so that

$$\Lambda = m2\pi/\Delta k = mL_c, \tag{22.2-33}$$

QPM Condition

i.e., the grating period Λ equals an integer multiple of the coherence length $L_c = 2\pi/\Delta k$.

Equation (22.2-32) together with the frequency matching condition yield

$$\omega_1 + \omega_2 = \omega_3, \quad \omega_1 n_1 + \omega_2 n_2 + m2\pi c/\Lambda = \omega_3 n_3. \tag{22.2-34}$$

QPM Tuning Curves

These equations are used in lieu of (22.2-20) to determine the tuning curves and the crystal angles in the design of parametric devices. It is evident that QPM offers some flexibility in the design of desired tuning curves.

QPM in a Medium with Periodically Reversed Nonlinear Coefficient

The simplest periodic pattern of the nonlinear optical coefficient $\mathrm{d}(z)$ alternates between two constant values, $+\mathrm{d}_o$ and $-\mathrm{d}_o$, at distances $\Lambda/2$, as shown in Fig. 22.2-15.

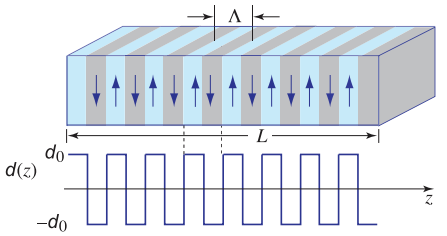


Figure 22.2-15 A nonlinear crystal with periodically varying nonlinear optical coefficient $\mathrm{d}(z)$ of period Λ.

The physical mechanism by which the periodic reversal of the sign of nonlinearity compensates the position-dependent phase of the radiation is illustrated in Fig. 22.2-16 for $m = 1$, i.e., when the grating period Λ equals the coherence length $L_c = 2\pi/\Delta k$.

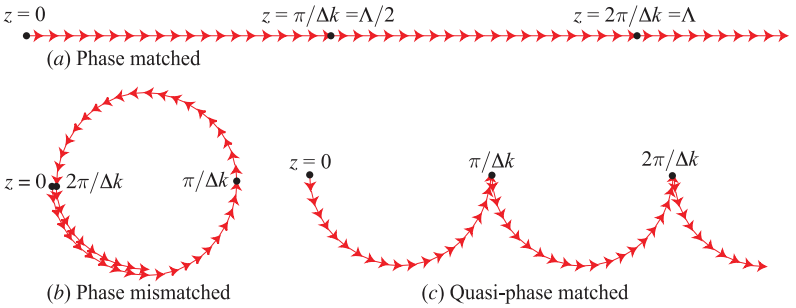


Figure 22.2-16 Phasors of the waves radiated by incremental elements at different positions z in the nonlinear medium. (a) In the phase-matched case ($\Delta k = 0$) the phasors are all aligned and maximum conversion efficiency is attained. (b) In the presence of a phase mismatch Δk, the phasors are misaligned and the efficiency is significantly reduced. (c) In the quasi-phase matched case, the misaligned phasors are periodically reversed by reversing the sign of the nonlinear optical coefficient at intervals of $\Lambda/2$. The conversion efficiency is partially restored.

The improvement of the conversion efficiency afforded by QPM may be determined quantitatively as follows. In accordance with Fourier-series theory, $\mathrm{d}_m = (2/m\pi)\mathrm{d}_o$, for odd m, and zero, otherwise. If phase matching is accomplished via the mth harmonic, i.e., $\Lambda = mL_c$, then the parametric conversion efficiency is proportional to $\mathrm{d}_m^2 = (2/\pi m)^2\mathrm{d}_o^2$. By contrast, a homogeneous medium with nonlinear optical coefficient d_o and the same length L, but with wavevector mismatch Δk, has a conversion efficiency $\mathrm{d}_o^2\,\mathrm{sinc}^2(\Delta kL/2\pi) = \mathrm{d}_o^2\,\mathrm{sinc}^2(L/L_c)$, which falls as $(\mathrm{d}_o^2/\pi^2)(L_c/L)^2$ when $L \gg L_c$. Since $L_c = \Lambda/m$, the improvement of conversion efficiency is by a factor of $4(L/\Lambda)^2$; it is proportional to the square of the number of periods of the periodic structure. Clearly, the use of a periodic medium can offer a significant improvement in conversion efficiency.

The most challenging aspect of quasi-phase matching is the fabrication of the periodic nonlinear structure. A uniform nonlinear crystal may be altered periodically by reversing the principal axis direction in alternating layers, thereby creating a d coefficient with alternating sign. This may be implemented by lithographically exposing the crystal to a periodic electric field that reverses the direction of the crystal's permanent electric polarization, a technique called **poling**. This approach has been applied to ferroelectric crystals such as $LiTaO_3$, KTP, and $LiNbO_3$; indeed, the latter has spawned a technology known as **periodically poled lithium niobate** (**PPLN**). Semiconductor crystals such as GaAs also have been used for the same purpose.

Periodic poling has also been implemented in integrated nonlinear optical waveguides. For example, PPLN waveguides may be fabricated by Ti-indiffusion at high temperatures or by annealed proton exchange. Ridge waveguides may be cut in PPLN by reactive ion etching. Single-mode waveguides are typically used to enhance the efficiency of wave mixing. An illustration of second-harmonic generation in a periodically poled, nonlinear waveguide is provided in Fig. 22.2-17. For example, a z-cut $LiNbO_3$ ridge waveguide of 1-μm height and 5-μm width can operate in a single TM mode at a wavelengths of 1550 nm or shorter. Poling periods of the order of 10–15 μm achieve quasi-phase matching for second-harmonic generation at this wavelength.

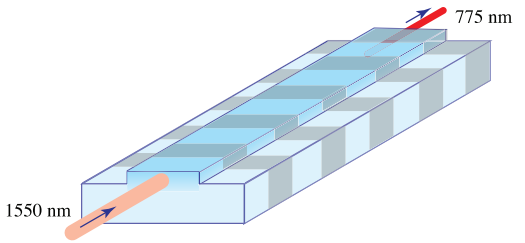


Figure 22.2-17 Schematic of second-harmonic generation in a periodically poled, integrated, nonlinear ridge waveguide.

22.3 THIRD-ORDER NONLINEAR OPTICS

In media possessing centrosymmetry, the second-order nonlinear term in (22.1-5) is absent since the polarization must reverse when the electric field reverses. The dominant nonlinearity is then of third order,

$$\mathcal{P}_{\mathrm{NL}} = 4\chi^{(3)}\mathcal{E}^3 \tag{22.3-1}$$

(Fig. 22.3-1) and the material is called a **third-order nonlinear medium** or a **Kerr medium**. Kerr media respond to optical fields by generating third harmonics and sums and differences of triplets of frequencies.

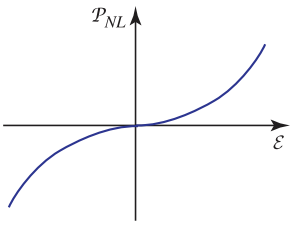


Figure 22.3-1 Third-order nonlinearity in a Kerr medium.

EXERCISE 22.3-1

Third-Order Nonlinear optical Media Exhibit the Kerr Electro-Optic Effect. Consider a monochromatic optical field $E(\omega)$ incident on a third-order nonlinear medium in the presence of a steady electric field $E(0)$. The optical field is taken to be much smaller than the electric field, so that $|E(\omega)|^2 \ll |E(0)|^2$. Use (22.3-1) to show that the component of $\mathcal{P}_{\rm NL}$ at frequency ω is approximately given by $P_{\rm NL}(\omega) \approx 12\chi^{(3)}E^2(0)E(\omega)$. Demonstrate that this component of the polarization density is equivalent to a refractive-index change $\Delta n = -\frac{1}{2}{\rm s}n^3E^2(0)$, where

$$\mathrm{s} = -\frac{12}{\epsilon_o n^4}\chi^{(3)}. \tag{22.3-2}$$

The proportionality between the refractive-index change and the squared electric field is the Kerr (quadratic) electro-optic effect described in Sec. 21.1A, where s is the Kerr coefficient [see (21.1-5)].

A. Third-Harmonic Generation (THG) and Optical Kerr Effect

Third-Harmonic Generation (THG)

In accordance with (22.3-1), the response of a third-order nonlinear medium to a monochromatic optical field $\mathcal{E}(t) = {\rm Re}\{E(\omega)\exp(j\omega t)\}$ is a nonlinear polarization density $\mathcal{P}_{\rm NL}(t)$ containing a component at frequency ω and another at frequency 3ω:

$$P_{\rm NL}(\omega) = 3\chi^{(3)}|E(\omega)|^2E(\omega) \tag{22.3-3a}$$

$$P_{\rm NL}(3\omega) = \chi^{(3)}E^3(\omega). \tag{22.3-3b}$$

The presence of a component of polarization density at the frequency 3ω in (22.3-3b) indicates that third-harmonic light is generated. In most cases, however, the energy conversion efficiency is low. Indeed, THG is often achieved via second-harmonic generation followed by sum-frequency generation of the fundamental and second-harmonic waves (an example is provided in Sec. 15.3B).

Optical Kerr Effect

The polarization-density component at frequency ω in (22.3-3a) corresponds to an incremental change of the susceptibility $\Delta\chi$ at frequency ω given by

$$\epsilon_o\Delta\chi = \frac{P_{\rm NL}(\omega)}{E(\omega)} = 3\chi^{(3)}|E(\omega)|^2 = 6\chi^{(3)}\eta I, \tag{22.3-4}$$

where $I = |E(\omega)|^2/2\eta$ is the optical intensity of the initial wave [see (5.4-8)]. Since $n^2 = 1+\chi$, we have $2n\Delta n = \Delta\chi$ so the incremental susceptibility change is equivalent to an incremental refractive-index change $\Delta n = \Delta\chi/2n$:

$$\Delta n = \frac{3\eta}{\epsilon_o n}\chi^{(3)}I \equiv n_2 I, \tag{22.3-5}$$

where

$$n_2 = \frac{3\eta_o}{n^2\epsilon_o}\chi^{(3)}. \tag{22.3-6}$$

Optical Kerr Coefficient

Thus, the change in the refractive index Δn is proportional to the optical intensity I. The overall refractive index is therefore a linear function[†] of the optical intensity,

$$n(I) = n + n_2 I. \tag{22.3-7}$$

Optical Kerr Effect

The proportionality of Δn to I is known as the **optical Kerr effect** because of its similarity to the electro-optic Kerr effect discussed in Exercise 22.3-1 and Sec. 21.1A, wherein Δn is proportional to the square of the steady electric field. The optical Kerr effect is a self-induced effect in which the phase velocity of the wave depends on the wave's own intensity. It is an example of **nonlinear refraction**.

The order of magnitude of the coefficient n_2 (in units of cm^2/W) is 10^{-16} to 10^{-14} in glasses, 10^{-14} to 10^{-7} in doped glasses, 10^{-10} to 10^{-8} in organic materials, and 10^{-10} to 10^{-2} in semiconductors. It is sensitive to the operating wavelength (see Sec. 22.7) and depends on the polarization.

B. Self-Phase Modulation (SPM), Self-Focusing, and Spatial Solitons

Self-Phase Modulation (SPM)

As a result of the optical Kerr effect, an optical wave traveling in a third-order nonlinear medium undergoes **self-phase modulation** (**SPM**). The phase shift incurred by an optical beam of power P and cross-sectional area A, traveling a distance L in the medium, is $\varphi = -n(I)k_oL = -2\pi n(I)L/\lambda_o = -2\pi(n + n_2P/A)L/\lambda_o$. Thus, the change in phase arising from the optical Kerr effect is

$$\Delta\varphi = -2\pi n_2 \frac{L}{\lambda_o A} P, \tag{22.3-8}$$

which is proportional to the optical power P. Self-phase modulation is useful in applications in which light controls light.

To maximize the effect, L should be large and A small. These requirements are well served by the use of optical waveguides. The optical power at which $\Delta\varphi = -\pi$ is attained is $P_\pi = \lambda_o A/2Ln_2$. A doped glass fiber of length $L = 1$ m, cross-sectional area $A = 10^{-2}$ mm^2, and $n_2 = 10^{-10}$ cm^2/W, operating at $\lambda_o = 1$ μm, for example, switches the phase by a factor of π at an optical power $P_\pi = 0.5$ W. Materials with larger values of n_2 can be used in centimeter-long channel waveguides to achieve a phase shift of π at powers of a few mW.

Phase modulation may thence be converted into intensity modulation by employing one of the schemes used in conjunction with electro-optic modulators (see Sec. 21.1B): (1) using an interferometer (Mach–Zehnder, for example); (2) using the difference between the modulated phases of the two polarization components (birefringence) as a wave retarder placed between crossed polarizers; or (3) using an integrated-photonic directional coupler (Sec. 9.4B). The result is an all-optical modulator in which a weak optical beam may be controlled by an intense optical beam. All-optical switches are discussed in Sec. 24.3C.

[†] Equation (22.3-7) is sometimes written in the alternative form, $n(I) = n + n_2|E|^2/2$, where n_2 differs from (22.3-6) by the factor η.

Self-Focusing

Another important effect associated with self-phase modulation is **self-focusing**. If an intense optical beam is transmitted through a thin sheet of nonlinear material exhibiting the optical Kerr effect, as illustrated in Fig. 22.3-2, the refractive-index change mimics the intensity pattern in the transverse plane. If the beam has its highest intensity at the center, for example, the maximum change of the refractive index is also at the center. The sheet then acts as a graded-index medium that imparts to the wave a nonuniform phase shift, thereby causing wavefront curvature. Under certain conditions the medium can act as a lens with a power-dependent focal length, as shown in Exercise 22.3-2. Kerr-lens focusing is useful for laser mode locking, as discussed in Sec. 16.4D.

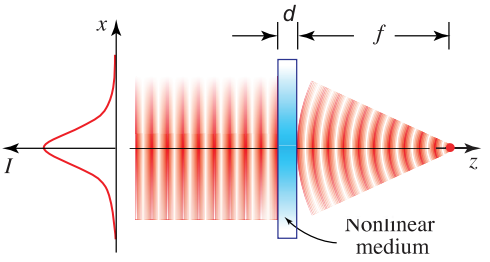


Figure 22.3-2 A third-order nonlinear medium acts as a lens whose focusing power depends on the intensity of the incident beam.

EXERCISE 22.3-2

Optical Kerr Lens. An optical beam traveling in the z direction is transmitted through a thin sheet of nonlinear optical material exhibiting the optical Kerr effect, $n(I) = n + n_2 I$. The sheet lies in the x–y plane and has a small thickness d so that its complex amplitude transmittance is $\exp(-jnk_o d)$, as shown in (2.4-3). The beam has an approximately planar wavefront and an intensity distribution $I \approx I_0[1-(x^2+y^2)/W^2]$ at points near the beam axis ($x,\ y \ll W$), where I_0 is the peak intensity and W is the beam width. Show that the medium acts as a thin lens with a focal length that is inversely proportional to I_0. *Hint*: A lens of focal length f has a complex amplitude transmittance proportional to $\exp[jk_o(x^2+y^2)/2f]$, as shown in (2.4-9) [see also Exercise 2.4-6].

Spatial Solitons

When an intense optical beam travels through a substantial thickness of nonlinear homogeneous medium, rather than a thin sheet, the refractive index is altered nonuniformly so that the medium can act as a graded-index waveguide. Thus, the beam can create its own waveguide. If the intensity of the beam has the same spatial distribution in the transverse plane as one of the modes of the waveguide that the beam itself creates, the beam propagates self-consistently without changing its spatial distribution. Under these conditions, *diffraction* is compensated by *self-phase modulation*, and the beam is confined to its self-created waveguide. Such self-guided beams are called **spatial solitons**. Analogous behavior occurs in the time domain when group velocity dispersion is compensated by self-phase modulation. As discussed in Sec. 23.5B, this leads to the formation of temporal solitons, which travel without changing shape.

The self-guiding of light in an optical Kerr medium is described mathematically by the Helmholtz equation, $\nabla^2 E + n^2(I)k_o^2 E = 0$, where $n(I) = n + n_2 I$, $k_o = \omega/c_o$, and $I = |E|^2/2\eta$. This is a nonlinear differential equation in E, which is simplified by writing $E = A\exp(-jkz)$, where $k = nk_o$, and assuming that the envelope $A = A(x,z)$ varies slowly in the z direction (in comparison with the wavelength $\lambda = 2\pi/k$), and does not vary in the y direction (see Sec. 2.2C). Using the approximation $(\partial^2/\partial z^2)[A\exp(-jkz)] \approx (-2jk\,\partial A/\partial z - k^2 A)\exp(-jkz)$, the Helmholtz

equation becomes

$$\frac{\partial^2 A}{\partial x^2} - j2k\frac{\partial A}{\partial z} + k_o^2[n^2(I) - n^2]A = 0. \tag{22.3-9}$$

Since the nonlinear effect is small ($n_2 I \ll n$), we write

$$[n^2(I) - n^2] = [n(I) - n][n(I) + n] \approx [n_2 I][2n] = \frac{2n_2 n|A|^2}{2\eta} = \frac{n^2 n_2}{\eta_o}|A|^2, \tag{22.3-10}$$

so that (22.3-9) becomes

$$\frac{\partial^2 A}{\partial x^2} + \frac{n_2}{\eta_o}k^2|A|^2 A = j2k\frac{\partial A}{\partial z}. \tag{22.3-11}$$

Equation (22.3-11) is known as the nonlinear Schrödinger equation. One of its solutions is

$$\boxed{A(x,z) = A_0 \operatorname{sech}\left(\frac{x}{W_0}\right)\exp\left(-j\frac{z}{4z_0}\right),} \tag{22.3-12}$$

Spatial Soliton

where W_0 is a constant; sech$(\cdot)$ indicates the hyperbolic-secant function; A_0 satisfies $n_2(A_0^2/2\eta_o) = 1/k^2W_0^2$, so that A_0W_0 is a constant; and $z_0 = \frac{1}{2}kW_0^2 = \pi W_0^2/\lambda$ [this is the same for the Rayleigh range of a Gaussian beam, as shown in (3.1-22)]. The intensity distribution

$$I(x,z) = \frac{|A(x,z)|^2}{2\eta} = \frac{A_0^2}{2\eta}\operatorname{sech}^2\left(\frac{x}{W_0}\right) \tag{22.3-13}$$

is independent of z and has a width W_0, as illustrated in Fig. 22.3-3. The distribution in (22.3-12) is the mode of a graded-index waveguide with a refractive index $n + n_2 I = n[1 + (1/k^2W_0^2)\text{sech}^2(x/W_0)]$, so that self-consistency is assured. Since $E = A\exp(-jkz)$, the wave travels with a propagation constant $k + 1/4z_0 = k(1 + \lambda^2/8\pi^2W_0^2)$ and phase velocity $c/(1 + \lambda^2/8\pi^2W_0^2)$. The velocity is smaller than c for localized beams (small W_0) but approaches c for large W_0.

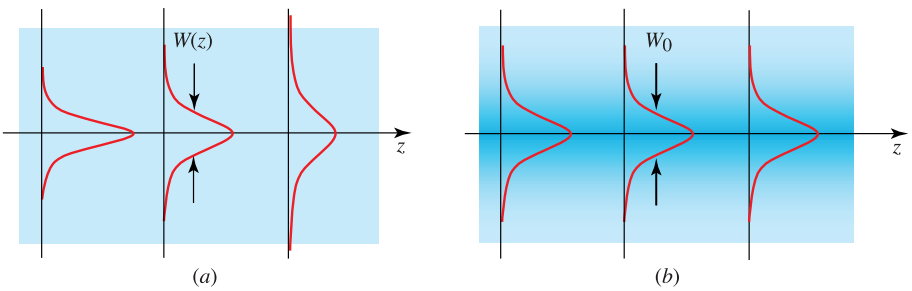


Figure 22.3-3 Comparison of (a) a Gaussian beam traveling in a linear medium, and (b) a spatial soliton (self-guided optical beam) traveling in a nonlinear medium.

Raman Gain

Much as the complex susceptibility $\chi = \chi' + j\chi''$ was constructed to accommodate loss and gain in linear optics (Sec. 5.5), we follow suit with the third-order nonlinear susceptibility $\chi^{(3)}$, setting $\chi^{(3)} = \chi^{(3)}_{\mathrm{R}} + j\chi^{(3)}_{\mathrm{I}}$. The self-phase modulation in (22.3-8), which is then also complex, may thus be written as

$$\Delta\varphi = -2\pi n_2 \frac{L}{\lambda_o A} P = -\frac{6\pi\eta_o}{\epsilon_o} \frac{\chi^{(3)}}{n^2} \frac{L}{\lambda_o A} P. \tag{22.3-14}$$

Consequently, the propagation phase factor $e^{j\varphi}$ is a combination of phase shift, $\Delta\varphi = -(6\pi\eta_o/\epsilon_o)(\chi^{(3)}_{\mathrm{R}}/n^2)(L/\lambda_o A)P$, and gain $G_R = \exp(\frac{1}{2}\gamma_R L)$, with a gain coefficient given by[†]

$$\boxed{\gamma_R = \frac{12\pi\eta_o}{\epsilon_o} \frac{\chi^{(3)}_{\mathrm{I}}}{n^2} \frac{1}{\lambda_o} \frac{P}{A},}$$

(22.3-15)
Raman Gain Coefficient

which is proportional to P/A.

The presence of such **Raman gain** arises from the pump field coherently driving the nonlinear polarization density of the Raman medium. It permits power to be transferred from the pump beam to the signal beam via an interaction between the light and the vibrational modes of the Raman medium. This is in contrast to the gain arising from population inversion in a stimulated-emission device, as discussed in Sec. 15.1. Raman gain underlies the operation of a distributed Raman fiber amplifier (RFA), where the pump and signal are both sent through the same fiber (Sec. 15.3D). When the gain exceeds the loss, and appropriate feedback is provided, the Raman amplifier becomes a Raman laser (Sec. 16.3C).

C. Cross-Phase Modulation (XPM)

We now consider the response of a third-order nonlinear medium to an optical field comprising two monochromatic waves of angular frequencies ω_1 and ω_2, $\mathcal{E}(t) = \mathrm{Re}\{E(\omega_1)\exp(j\omega_1 t)\} + \mathrm{Re}\{E(\omega_2)\exp(j\omega_2 t)\}$. On substitution in (22.3-1), the component $P_{\mathrm{NL}}(\omega_1)$ of the polarization density at frequency ω_1 turns out to be

$$P_{\mathrm{NL}}(\omega_1) = \chi^{(3)} \left[3|E(\omega_1)|^2 + 6|E(\omega_2)|^2\right] E(\omega_1)\,. \tag{22.3-16}$$

Assuming that the two waves have the same refractive index n, this relation may be cast in the form $P_{\mathrm{NL}}(\omega_1) = 2\epsilon_o n \Delta n E(\omega_1)$, where

$$\boxed{\Delta n = n_2(I_1 + 2I_2),}$$

(22.3-17)
Cross-Phase Modulation

with $n_2 = 3\eta_o\chi^{(3)}/\epsilon_o n^2$. The quantities $I_1 = |E(\omega_1)|^2/2\eta$ and $I_2 = |E(\omega_2)|^2/2\eta$ are the intensities of waves 1 and 2, respectively. Therefore, wave 1 travels with an effective refractive index $n + \Delta n$ controlled by its own intensity as well as that of wave 2. Wave 2 encounters a similar effect, so that the waves are coupled.

[†] The Raman gain coefficient is sometimes expressed in the form $\gamma_R/(P/A)$, which has units of cm·W^{-1}.

Since the phase shift encountered by wave 1 is modulated by the intensity of wave 2, this phenomenon is known as **cross-phase modulation** (**XPM**). It can result in the contamination of information between optical communication channels at neighboring frequencies, as in wavelength-division-multiplexing systems (WDM) (see Sec. 25.3C).

As we have seen in Sec. 22.2C, two-wave mixing is not possible in a second-order nonlinear medium (except in the degenerate case). Note, however, that two-wave mixing can occur in photorefractive media, as illustrated in Fig. 21.4-3.

EXERCISE 22.3-3

Optical Kerr Effect in the Presence of Three Waves. Three monochromatic waves with frequencies ω_1, ω_2, and ω_3 travel in a third-order nonlinear medium. Determine the complex amplitude of the component of $\mathcal{P}_{\text{NL}}(t)$ in (22.3-1) at frequency ω_1. Show that this wave travels with a velocity $c_o/(n+\Delta n)$, where

$$\Delta n = n_2(I_1 + 2I_2 + 2I_3), \tag{22.3-18}$$

and $n_2 = 3\eta_o\chi^{(3)}/\epsilon_o n^2$, with $I_q = |E(\omega_q)|^2/2\eta,\ q = 1, 2, 3.$

D. Four-Wave Mixing (FWM)

We now examine the case of **four-wave mixing** (**FWM**) in a third-order nonlinear medium. We begin by determining the response of the medium to a superposition of three waves of angular frequencies ω_1, ω_2, and ω_3, with field

$$\mathcal{E}(t) = \text{Re}\{E(\omega_1)\ \exp(j\omega_1 t)\} + \text{Re}\{E(\omega_2)\ \exp(j\omega_2 t)\} + \text{Re}\{E(\omega_3)\ \exp(j\omega_3 t)\}. \tag{22.3-19}$$

It is convenient to write $\mathcal{E}(t)$ as a sum of six terms

$$\mathcal{E}(t) = \sum_{q=\pm1,\pm2,\pm3} \tfrac{1}{2}E(\omega_q)\ \exp(j\omega_q t), \tag{22.3-20}$$

where $\omega_{-q} = -\omega_q$ and $E(-\omega_q) = E^*(\omega_q)$. Substituting (22.3-20) into (22.3-1), we write $\mathcal{P}_{\text{NL}}$ as a sum of $6^3 = 216$ terms,

$$\mathcal{P}_{\text{NL}}(t) = \tfrac{1}{8}\chi^{(3)} \sum_{q,r,l=\pm1,\pm2,\pm3} E(\omega_q)E(\omega_r)E(\omega_l)\ \exp[j(\omega_q + \omega_r + \omega_l)t]. \tag{22.3-21}$$

Thus, $\mathcal{P}_{\text{NL}}$ is the sum of harmonic components of frequencies $\omega_1, \ldots, 3\omega_1, \ldots, 2\omega_1 \pm \omega_2, \ldots, \pm\omega_1 \pm \omega_2 \pm \omega_3$. The amplitude $P_{\text{NL}}(\omega_q + \omega_r + \omega_l)$ of the component of frequency $\omega_q + \omega_r + \omega_l$ can be determined by adding appropriate permutations of q, r, and l in (22.3-21). For example, $P_{\text{NL}}(\omega_1 + \omega_2 - \omega_3)$ involves six permutations,

$$P_{\text{NL}}(\omega_1 + \omega_2 - \omega_3) = 6\chi^{(3)}E(\omega_1)E(\omega_2)E^*(\omega_3). \tag{22.3-22}$$

Equation (22.3-22) indicates that four waves of frequencies ω_1, ω_2, ω_3, and ω_4 are mixed by the medium if $\omega_4 = \omega_1 + \omega_2 - \omega_3$, or

$$\boxed{\omega_1 + \omega_2 = \omega_3 + \omega_4.} \tag{22.3-23}$$

Frequency-Matching Condition

This equation constitutes the frequency-matching condition for FWM.

Assuming that waves 1, 2, and 3 are plane waves of wavevectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$, so that $E(\omega_q) \propto \exp(-j\mathbf{k}_q \cdot \mathbf{r}), q = 1, 2, 3$, then (22.3-22) gives

$$P_{\text{NL}}(\omega_4) \propto \exp(-j\mathbf{k}_1 \cdot \mathbf{r}) \exp(-j\mathbf{k}_2 \cdot \mathbf{r}) \exp(j\mathbf{k}_3 \cdot \mathbf{r}) = \exp[-j(\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3) \cdot \mathbf{r}], \tag{22.3-24}$$

so that wave 4 is also a plane wave with wavevector $\mathbf{k}_4 = \mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3$, from which

$$\boxed{\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3 + \mathbf{k}_4.} \tag{22.3-25}$$

Phase-Matching Condition

Equation (22.3-25) is the phase-matching condition for FWM.

Several FWM processes occur simultaneously, all satisfying the frequency and phase-matching conditions. As shown before, waves 1, 2, and 3 interact and generate wave 4, in accordance with (22.3-22). Similarly, waves 3, 4, and 1 interact and generate wave 2, in accordance with

$$P_{\text{NL}}(\omega_2) = 6\chi^{(3)} E(\omega_3) E(\omega_4) E^*(\omega_1), \tag{22.3-26}$$

and so on.

The FWM process may also be interpreted as an interaction between four photons. A photon of frequency ω_3 and another of of frequency ω_4 are annihilated to create a photon of frequency ω_1 and another of frequency w_2, as illustrated in Fig. 22.3-4. Equations (22.3-23) and (22.3-25) represent conservation of energy and momentum, respectively.

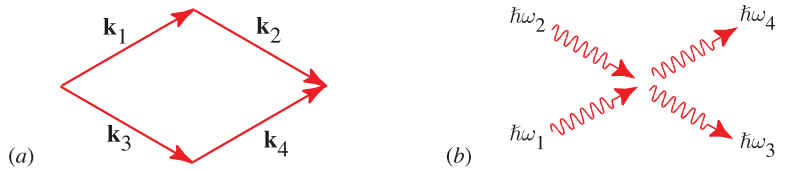


Figure 22.3-4 Four-wave mixing (FWM): (*a*) phase-matching condition; (*b*) interaction of four photons.

Three-Wave Mixing

In the partially degenerate case for which two of the four waves have the same frequency, $\omega_3 = \omega_4 \equiv \omega_0$, we have three waves with frequencies related by

$$\omega_1 + \omega_2 = 2\omega_0, \tag{22.3-27}$$

so that the frequencies ω_1 and ω_2 are symmetrically located with respect to the central frequency ω_0, much like the sidebands of an amplitude modulated sine wave, or the Stokes and anti-Stokes frequencies in Raman scattering. The components of the nonlinear polarization density at ω_1, ω_2, and ω_3 include terms of the form

$$P_{\text{NL}}(\omega_1) = 3\chi^{(3)} E^2(\omega_3) E^*(\omega_2), \tag{22.3-28a}$$

$$P_{\text{NL}}(\omega_2) = 3\chi^{(3)} E^2(\omega_3) E^*(\omega_1), \tag{22.3-28b}$$

$$P_{\text{NL}}(\omega_3) = 6\chi^{(3)} E(\omega_1) E(\omega_2) E^*(\omega_3). \tag{22.3-28c}$$

These terms are responsible for three-wave mixing, i.e., radiation at the frequency of each wave generated by mixing of the other waves. These mixing processes may be used for optical frequency conversion (OFC), optical parametric amplification (OPA) and oscillation (OPO), and spontaneous parametric downconversion (SPDC), much like three-wave mixing in second-order nonlinear media; the waves at ω_1, ω_2, and ω_3 may be regarded as the signal, idler, and pump of the parametric process. Note, however, that this *three-wave mixing* process involves *four* photons. For example, the annihilation of two photons at ω_0 and the creation of two photons at ω_1 and ω_2. An example of OPA in a $\chi^{(3)}$ medium, such as a silica-glass optical fiber, is illustrated in Fig. 22.3-5.

Figure 22.3-5 Three-wave, four-photon optical fiber parametric amplifier (OPA).

E. Optical Phase Conjugation (OPC)

The frequency-matching condition (22.3-23) is satisfied when all four waves are of the same frequency:

$$\omega_1 = \omega_2 = \omega_3 = \omega_4 = \omega. \tag{22.3-29}$$

The process is then called **degenerate four-wave mixing**.

Assuming further that two of the waves (waves 3 and 4) are uniform plane waves traveling in opposite directions,

$$E_3(\mathbf{r}) = A_3 \exp(-j\mathbf{k}_3 \cdot \mathbf{r}), \qquad E_4(\mathbf{r}) = A_4 \exp(-j\mathbf{k}_4 \cdot \mathbf{r}), \tag{22.3-30}$$

with

$$\mathbf{k}_4 = -\mathbf{k}_3, \tag{22.3-31}$$

and substituting (22.3-30) and (22.3-31) into (22.3-26), we see that the polarization density of wave 2 is $6\chi^{(3)} A_3 A_4 E_1^*(\mathbf{r})$. This term corresponds to a source emitting an optical wave (wave 2) of complex amplitude

$$\boxed{E_2(\mathbf{r}) \propto A_3 A_4 E_1^*(\mathbf{r}).} \tag{22.3-32}$$

Phase Conjugation

Since A_3 and A_4 are constants, wave 2 is proportional to a conjugated version of wave 1. The device serves as a **phase conjugator**. Waves 3 and 4 are called the **pump** waves and waves 1 and 2 are called the **probe** and **conjugate** waves, respectively. As will be demonstrated shortly, the conjugate wave is identical to the probe wave except that it travels in the opposite direction. The phase conjugator is a special mirror that reflects the wave back onto itself without altering its wavefronts.

To understand the phase conjugation process consider two simple examples:

EXAMPLE 22.3-1. ***Conjugate of a Plane Wave.*** If wave 1 is a uniform plane wave, $E_1(\mathbf{r}) = A_1 \exp(-j\mathbf{k}_1 \cdot \mathbf{r})$, traveling in the direction $\mathbf{k}_1$, then $E_2(\mathbf{r}) = A_1^* \exp(j\mathbf{k}_1 \cdot \mathbf{r})$ is a uniform plane wave traveling in the opposite direction $\mathbf{k}_2 = -\mathbf{k}_1$, as illustrated in Fig. 22.3-6(b). Thus, the phase-matching condition (22.3-25) is satisfied. The medium acts as a special "mirror" that reflects the incident plane wave back onto itself, no matter what the angle of incidence.

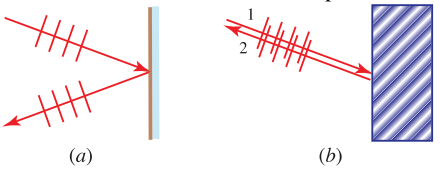


Figure 22.3-6 Reflection of a plane wave from (a) an ordinary mirror and (b) a phase conjugate mirror.

EXAMPLE 22.3-2. ***Conjugate of a Spherical Wave.*** If wave 1 is a spherical wave centered about the origin $\mathbf{r} = \mathbf{0}$, $E_1(\mathbf{r}) \propto (1/r)\exp(-jkr)$, then wave 2 has complex amplitude $E_2(\mathbf{r}) \propto (1/r)\exp(+jkr)$. This is a spherical wave traveling backward and converging toward the origin, as illustrated in Fig. 22.3-7(b).

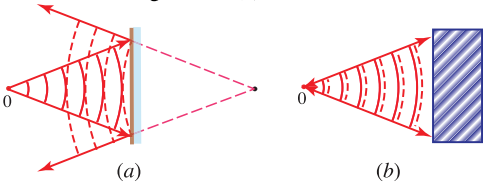


Figure 22.3-7 Reflection of a spherical wave from (a) an ordinary mirror and (b) a phase conjugate mirror.

Since an arbitrary probe wave may be regarded as a superposition of plane waves (see Chapter 4), each of which is reflected onto itself by the conjugator, the conjugate wave is identical to the incident wave everywhere, except for a reversed direction of propagation. The conjugate wave retraces the original wave by propagating backward, maintaining the same wavefronts.

Phase conjugation is analogous to time reversal. This may be understood by examining the field of the conjugate wave $\mathcal{E}_2(\mathbf{r},t) = \text{Re}\{E_2(\mathbf{r})\exp(j\omega t)\} \propto \text{Re}\{E_1^*(\mathbf{r})\exp(j\omega t)\}$. Since the real part of a complex number equals the real part of its complex conjugate, $\mathcal{E}_2(\mathbf{r},t) \propto \text{Re}\{E_1(\mathbf{r})\exp(-j\omega t)\}$. Comparing this to the field of the probe wave $\mathcal{E}_1(\mathbf{r},t) = \text{Re}\{E_1(\mathbf{r})\exp(j\omega t)\}$, we readily see that one is obtained from the other by the transformation $t \to -t$, so that the conjugate wave appears as a time-reversed version of the probe wave.

The conjugate wave may carry more power than the probe wave. This can be seen by observing that the intensity of the conjugate wave (wave 2) is proportional to the product of the intensities of the pump waves 3 and 4 [see (22.3-32)]. When the powers of the pump waves are increased so that the conjugate wave (wave 2) carries more power than the probe wave (wave 1), the medium acts as an "amplifying mirror." An example of an optical arrangement that provides phase conjugation is shown in Fig. 22.3-8.

Degenerate Four-Wave Mixing as a Form of Real-Time Holography

The degenerate four-wave-mixing process is analogous to volume holography (see Sec. 4.5). Holography is a two-step process in which the interference pattern formed by the superposition of an object wave E_1 and a reference wave E_3 is recorded in a photographic emulsion. Another reference wave E_4 is subsequently transmitted through or reflected from the emulsion, creating the conjugate of the object wave $E_2 \propto E_4 E_3 E_1^*$, or its replica $E_2 \propto E_4 E_1 E_3^*$, depending on the geometry [see Fig 4.5-10(a) and (b)]. The nonlinear medium permits a real-time simultaneous holographic recording and reconstruction process. This process occurs in both the Kerr medium and the

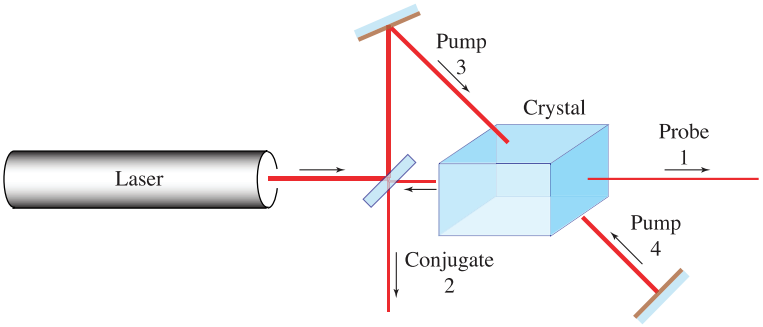


Figure 22.3-8 An optical system for degenerate four-wave mixing using a nonlinear crystal. The pump waves 3 and 4 and the probe wave 1 are obtained from a laser using a beamsplitter and two mirrors. The conjugate wave 2 is created within the crystal.

photorefractive medium (see Sec. 21.4).

When four waves are mixed in a nonlinear medium, each pair of waves interferes and creates a grating, from which a third wave is reflected to produce the fourth wave. The roles of reference and object are exchanged among the four waves, so that there are two types of gratings as illustrated in Fig. 22.3-9. Consider first the process illustrated in Fig. 22.3-9(a) [see also Fig. 4.5-10(a)]. Assume that the two reference waves (denoted as waves 3 and 4) are counterpropagating plane waves. The two steps of holography are:

1. The object wave 1 is added to the reference wave 3 and the intensity of their sum is recorded in the medium in the form of a volume grating (hologram).
2. The reconstruction reference wave 4 is Bragg reflected from the grating to create the conjugate wave (wave 2).

This grating is called the transmission grating.

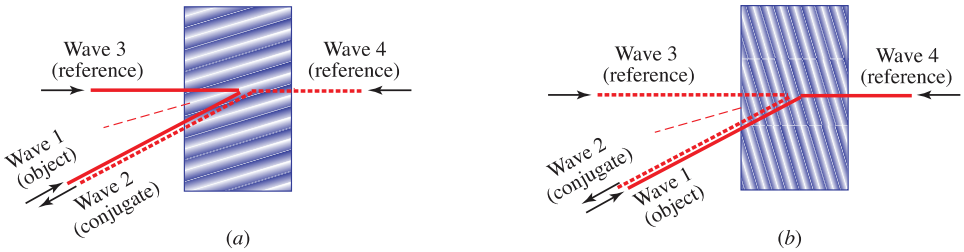


Figure 22.3-9 Four-wave mixing in a nonlinear medium. A reference and object wave interfere and create a grating from which the second reference wave reflects and produces a conjugate wave. There are two possibilities corresponding to (a) transmission and (b) reflection gratings.

The second possibility, illustrated in Fig. 22.3-9(b), is for the reference wave 4 to interfere with the object wave 1 and create a grating, called the reflection grating, from which the second reference wave 3 is reflected to create the conjugate wave 2. These two gratings can exist together but they usually have different efficiencies.

In summary, four-wave mixing can provide a means for real-time holography and phase conjugation, which have a number of applications in optical signal processing.

Use of Phase Conjugators in Wave Restoration

The ability to reflect a wave onto itself so that it retraces its path in the opposite direction suggests a number of useful applications, including the removal of wavefront

aberrations. The idea is based on the principle of reciprocity, illustrated in Fig. 22.3-10. Rays traveling through a linear optical medium from left to right follow the same path if they reverse and travel back in the opposite direction. The same principle applies to waves.

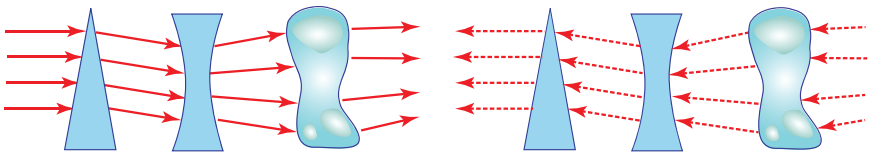


Figure 22.3-10 Optical reciprocity.

If the wavefront of an optical beam is distorted by an aberrating medium, the original wave can be restored by use of a conjugator that reflects the beam onto itself and transmits it once more through the same medium, as illustrated in Fig. 22.3-11.

One important application is in optical resonators (see Chapter 11). If the resonator contains an aberrating medium, replacing one of the mirrors with a conjugate mirror ensures that the distortion is removed in each round trip, so that the resonator modes have undistorted wavefronts transmitted through the ordinary mirror, as illustrated in Fig. 22.3-12.

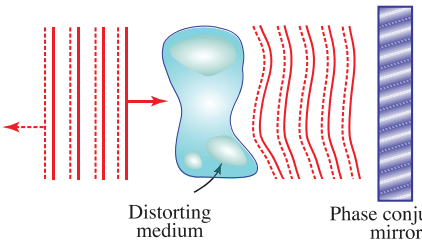


Figure 22.3-11 A phase conjugate mirror reflects a distorted wave onto itself, so that when it retraces its path, the distortion is compensated.

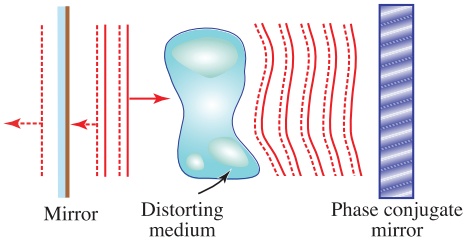


Figure 22.3-12 An optical resonator with an ordinary mirror and a phase conjugate mirror.

*22.4 SECOND-ORDER NONLINEAR OPTICS: COUPLED WAVES

A quantitative analysis of the process of three-wave mixing in a second-order nonlinear optical medium is provided in this section using a coupled-wave theory. Unlike the treatment provided in Sec. 22.2, all three waves are treated on equal footing. To simplify the analysis, consideration of anisotropic and dispersive effects is deferred to Secs. 22.6 and 22.7, respectively.

Coupled-Wave Equations

In accordance with (22.1-6) and (22.1-7), wave propagation in a second-order nonlinear medium is governed by the basic wave equation

$$\nabla^2 \mathcal{E} - \frac{1}{c^2}\frac{\partial^2 \mathcal{E}}{\partial t^2} = -\mathcal{S}, \tag{22.4-1}$$

where

$$\mathcal{S} = -\mu_o \frac{\partial^2 \mathcal{P}_{\text{NL}}}{\partial t^2} \tag{22.4-2}$$

is regarded as a radiation source, and

$$\mathcal{P}_{\text{NL}} = 2\text{d}\mathcal{E}^2 \tag{22.4-3}$$

is the nonlinear component of the polarization density.

In three-wave mixing, the field $\mathcal{E}(t)$ is taken as a superposition of three waves of angular frequencies ω_1, ω_2, and ω_3 with complex amplitudes E_1, E_2, and E_3:

$$\mathcal{E}(t) = \sum_{q=1,2,3} \text{Re}\left\{E_q \exp(j\omega_q t)\right\} = \sum_{q=1,2,3} \tfrac{1}{2}\left[E_q \exp(j\omega_q t) + E_q^* \exp(-j\omega_q t)\right] \tag{22.4-4}$$

[compare with (22.2-12) in the context of the first Born approximation]. It is convenient to rewrite (22.4-4) in the compact form

$$\mathcal{E}(t) = \sum_{q=\pm1,\pm2,\pm3} \tfrac{1}{2} E_q \exp(j\omega_q t), \tag{22.4-5}$$

where $\omega_{-q} = -\omega_q$ and $E_{-q} = E_q^*$. The corresponding polarization density obtained by substituting (22.4-5) into (22.4-3) is a sum of $6 \times 6 = 36$ terms,

$$\mathcal{P}_{\text{NL}}(t) = 2\text{d} \cdot \tfrac{1}{4} \sum_{q,r=\pm1,\pm2,\pm3} E_q E_r \exp\left[j(\omega_q + \omega_r)t\right]. \tag{22.4-6}$$

Thus, the corresponding radiation source is

$$\mathcal{S} = \tfrac{1}{2}\mu_o \text{d} \sum_{q,r=\pm1,\pm2,\pm3} (\omega_q + \omega_r)^2 E_q E_r \exp\left[j(\omega_q + \omega_r)t\right], \tag{22.4-7}$$

which generates a sum of harmonic components whose frequencies are sums and differences of the original frequencies ω_1, ω_2, and ω_3.

Substituting (22.4-5) and (22.4-7) into the wave equation (22.4-1) leads to a single differential equation with many terms, each of which is a harmonic function of some frequency. If the frequencies ω_1, ω_2, and ω_3 are distinct, we can separate this equation into three time-independent differential equations by equating terms on both sides of (22.4-1) at each of the frequencies ω_1, ω_2, and ω_3, separately. The result is cast in the form of three Helmholtz equations with associated sources,

$$(\nabla^2 + k_1^2)E_1 = -S_1 \tag{22.4-8a}$$

$$(\nabla^2 + k_2^2)E_2 = -S_2 \tag{22.4-8b}$$

$$(\nabla^2 + k_3^2)E_3 = -S_3, \tag{22.4-8c}$$

where S_q is the complex amplitude of the component of $\mathcal{S}$ with frequency ω_q and $k_q = n\omega_q/c_o$, $q = 1, 2, 3$. Each of the complex amplitudes of the three waves satisfies the Helmholtz equation with a source equal to the component of $\mathcal{S}$ at its frequency. Under certain conditions, the source for one wave depends on the electric fields of the other two waves, so that the three waves are coupled.

In the absence of nonlinearity, $\mathrm{d} = 0$ whereupon the source term $\mathcal{S}$ vanishes and each of the three waves satisfies the Helmholtz equation independently of the other two, as expected in linear optics.

If the frequencies ω_1, ω_2, and ω_3 are not commensurate (one frequency is not the sum or difference of the other two, and one frequency is not twice another), then the source term $\mathcal{S}$ does not contain any components of frequencies ω_1, ω_2, or ω_3. The components S_1, S_2, and S_3 then vanish and the three waves do not interact.

For the three waves to be coupled by the medium, their frequencies must be commensurate. Assume, for example, that one frequency is the sum of the other two,

$$\omega_1 + \omega_2 = \omega_3. \tag{22.4-9}$$

The source $\mathcal{S}$ then contains components at the frequencies ω_1, ω_2, and ω_3. Examining the 36 terms of (22.4-7) yields

$$S_1 = 2\mu_o\omega_1^2 \mathrm{d}\, E_3 E_2^* \tag{22.4-10}$$

$$S_2 = 2\mu_o\omega_2^2 \mathrm{d}\, E_3 E_1^* \tag{22.4-11}$$

$$S_3 = 2\mu_o\omega_3^2 \mathrm{d}\, E_1 E_2. \tag{22.4-12}$$

The source for wave 1 is proportional to $E_3E_2^*$ (since $\omega_1 = \omega_3 - \omega_2$), so that waves 2 and 3 together contribute to the growth of wave 1. Similarly, the source for wave 3 is proportional to E_1E_2 (since $\omega_3 = \omega_1 + \omega_2$), so that waves 1 and 2 combine to amplify wave 3, and so on. The three waves are thus coupled or "mixed" by the medium in a process described by three coupled differential equations in E_1, E_2, and E_3,

$$(\nabla^2 + k_1^2)E_1 = -2\mu_o\omega_1^2 \mathrm{d}\, E_3 E_2^* \tag{22.4-13a}$$

$$(\nabla^2 + k_2^2)E_2 = -2\mu_o\omega_2^2 \mathrm{d}\, E_3 E_1^* \tag{22.4-13b}$$

$$(\nabla^2 + k_3^2)E_3 = -2\mu_o\omega_3^2 \mathrm{d}\, E_1 E_2. \tag{22.4-13c}$$

3-Wave-Mixing Coupled Equations

EXERCISE 22.4-1

SHG as Degenerate Three-Wave Mixing. Equations (22.4-13) are valid only when the frequencies ω_1, ω_2, and ω_3 are distinct. Consider now the degenerate case for which $\omega_1 = \omega_2 = \omega$ and $\omega_3 = 2\omega$, so that there are two instead of three waves, with amplitudes E_1 and E_3. This corresponds to second-harmonic generation (SHG). Show that these waves satisfy the Helmholtz equation with sources

$$S_1 = 2\mu_o\omega_1^2 \mathrm{d}\, E_3 E_1^* \tag{22.4-14}$$

$$S_3 = \mu_o\omega_3^2 \mathrm{d}\, E_1 E_1, \tag{22.4-15}$$

so that the coupled wave equations are

$$(\nabla^2 + k_1^2)E_1 = -2\mu_o\omega_1^2 \mathrm{d}\, E_3 E_1^*, \tag{22.4-16a}$$

$$(\nabla^2 + k_3^2)E_3 = \ -\mu_o\omega_3^2 \mathrm{d}\, E_1 E_1. \tag{22.4-16b}$$

SHG Coupled Equations

Note that these equations are not obtained from the three-wave-mixing equations (22.4-13) by substituting $E_1 = E_2$ [the factor of 2 is absent in (22.4-16b)].

Mixing of Three Collinear Uniform Plane Waves

Assume that the three waves are plane waves traveling in the z direction with complex amplitudes $E_q = A_q \exp(-jk_q z)$, complex envelopes A_q, and wavenumbers $k_q = \omega_q/c$, $q = 1, 2, 3$. It is convenient to normalize the complex envelopes by defining the variables $a_q = A_q/(2\eta\hbar\omega_q)^{1/2}$, where $\eta = \eta_o/n$ is the impedance of the medium, $\eta_o = (\mu_o/\epsilon_o)^{1/2}$ is the impedance of free space, and $\hbar\omega_q$ is the energy of a photon of angular frequency ω_q. Thus,

$$E_q = \sqrt{2\eta\hbar\omega_q}\,\mathrm{a}_q \exp(-jk_q z), \qquad q = 1, 2, 3, \tag{22.4-17}$$

and the intensities of the three waves are $I_q = |E_q|^2/2\eta = \hbar\omega_q|\mathrm{a}_q|^2$. The photon-flux densities (photons/s-m^2) associated with these waves are

$$\phi_q = \frac{I_q}{\hbar\omega_q} = |\mathrm{a}_q|^2. \tag{22.4-18}$$

The variable a_q therefore represents the complex envelope of wave q, scaled such that $|\mathrm{a}_q|^2$ is the photon-flux density. This scaling is convenient since the process of wave mixing must be governed by photon-number conservation (see Sec. 22.2C).

As a result of the interaction between the three waves, the complex envelopes a_q vary with z so that $\mathrm{a}_q = \mathrm{a}_q(z)$. If the interaction is weak, the $\mathrm{a}_q(z)$ vary slowly with z, so that they can be assumed approximately constant within a distance of a wavelength. This makes it possible to use the slowly varying envelope approximation wherein $d^2\mathrm{a}_q/dz^2$ is neglected relative to $k_q d\mathrm{a}_q/dz = (2\pi/\lambda_q)da_q/dz$ and

$$(\nabla^2 + k_q^2)[\mathrm{a}_q \exp(-jk_q z)] \approx -j2k_q \frac{d\mathrm{a}_q}{dz} \exp(-jk_q z) \tag{22.4-19}$$

(see Sec. 2.2C). With this approximation (22.4-13) reduce to simpler equations that are akin to the paraxial Helmholtz equations, in which the mismatch in phase is considered:

$$\frac{d\mathrm{a}_1}{dz} = -j\mathrm{g}\mathrm{a}_3\mathrm{a}_2^* \exp(-j\Delta k\, z) \tag{22.4-20a}$$

$$\frac{d\mathrm{a}_2}{dz} = -j\mathrm{g}\mathrm{a}_3\mathrm{a}_1^* \exp(-j\Delta k\, z) \tag{22.4-20b}$$

$$\frac{d\mathrm{a}_3}{dz} = -j\mathrm{g}\mathrm{a}_1\mathrm{a}_2 \exp(j\Delta k\, z) \tag{22.4-20c}$$

3-Wave-Mixing Coupled Equations

where

$$\mathrm{g}^2 = 2\hbar\omega_1\omega_2\omega_3\eta^3\mathrm{d}^2 \tag{22.4-21}$$

and

$$\Delta k = k_3 - k_2 - k_1 \tag{22.4-22}$$

represents the error in the phase-matching condition. The variations of a_1, a_2, and a_3 with z are therefore governed by three coupled first-order differential equations (22.4-20), which we proceed to solve under the different boundary conditions corresponding to various applications. It is useful, however, first to derive some invariants of the

wave-mixing process. These are functions of a_1, a_2, and a_3 that are independent of z. Invariants are useful since they can be used to reduce the number of independent variables. Exercises 22.4-3 and 22.4-2 develop invariants based on conservation of energy and conservation of photons.

EXERCISE 22.4-2

Photon-Number Conservation: The Manley–Rowe Relations. Using (22.4-20), show that

$$\frac{d}{dz}|\mathrm{a}_1|^2 = \frac{d}{dz}|\mathrm{a}_2|^2 = -\frac{d}{dz}|\mathrm{a}_3|^2, \tag{22.4-23}$$

from which the Manley–Rowe relations (22.2-19), derived using photon-number conservation, follow. Equation (22.4-23) implies that $|\mathrm{a}_1|^2 + |\mathrm{a}_3|^2$ and $|\mathrm{a}_2|^2 + |\mathrm{a}_3|^2$ are also invariants of the wave-mixing process.

EXERCISE 22.4-3

Energy Conservation. Show that the sum of the intensities $I_q = \hbar\omega_q|\mathrm{a}_q|^2$, $q = 1, 2, 3$, of the three waves governed by (22.4-20) is invariant to z, so that

$$\frac{d}{dz}(I_1 + I_2 + I_3) = 0. \tag{22.4-24}$$

A. Second-Harmonic Generation (SHG)

Second-harmonic generation (SHG) is a degenerate case of three-wave mixing in which

$$\omega_1 = \omega_2 = \omega \qquad \text{and} \qquad \omega_3 = 2\omega. \tag{22.4-25}$$

Two forms of interaction occur: in SHG, two photons of frequency ω combine to form a single photon of frequency 2ω, as illustrated in Fig. 22.4-1(a); in degenerate parametric downconversion, one photon of frequency 2ω splits into two photons of frequency ω.

The interaction of the two waves is described by the paraxial Helmholtz equations with sources. Conservation of momentum requires that

$$2\mathbf{k}_1 = \mathbf{k}_3. \tag{22.4-26}$$

EXERCISE 22.4-4

Coupled-Wave Equations for SHG. Apply the slowly varying envelope approximation (22.4-19) to the Helmholtz equations (22.4-16), which describe two collinear waves in the degenerate case, to show that

$$\frac{d\mathrm{a}_1}{dz} = -j\mathrm{g}\mathrm{a}_3\mathrm{a}_1^* \exp(-j\Delta kz) \tag{22.4-27a}$$

$$\frac{d\mathrm{a}_3}{dz} = -j\frac{\mathrm{g}}{2}\mathrm{a}_1\mathrm{a}_1 \exp(j\Delta kz), \tag{22.4-27b}$$

where $\Delta k = k_3 - 2k_1$ and

$$\mathrm{g}^2 = 4\hbar\omega^3\eta^3\mathrm{d}^2. \tag{22.4-28}$$

Assuming two collinear waves with perfect phase matching ($\Delta k = 0$), equations (22.4-27) reduce to

$$\frac{d\mathrm{a}_1}{dz} = -j\mathrm{g}\mathrm{a}_3\mathrm{a}_1^* \tag{22.4-29a}$$

$$\frac{d\mathrm{a}_3}{dz} = -j\frac{\mathrm{g}}{2}\mathrm{a}_1\mathrm{a}_1. \tag{22.4-29b}$$

SHG Coupled Equations

At the input to the device ($z = 0$) the amplitude of the second-harmonic wave is assumed to be zero, $\mathrm{a}_3(0) = 0$, and that of the fundamental wave, $\mathrm{a}_1(0)$, is assumed to be real. We seek a solution for which $\mathrm{a}_1(z)$ is real everywhere. Using the energy conservation relation $\mathrm{a}_1^2(z) + 2|\mathrm{a}_3(z)|^2 = \mathrm{a}_1^2(0)$, (22.4-29b) gives a differential equation in $\mathrm{a}_3(z)$,

$$d\mathrm{a}_3/dz = -j(g/2)[\mathrm{a}_1^2(0) - 2|\mathrm{a}_3(z)|^2], \tag{22.4-30}$$

whose solution may be substituted in (22.4-29a) to obtain the overall solution:

$$\mathrm{a}_1(z) = \mathrm{a}_1(0)\,\mathrm{sech}\left(\tfrac{1}{\sqrt{2}}\mathrm{g}\mathrm{a}_1(0)z\right) \tag{22.4-31a}$$

$$\mathrm{a}_3(z) = -\tfrac{j}{\sqrt{2}}\,\mathrm{a}_1(0)\tanh\left(\tfrac{1}{\sqrt{2}}\mathrm{g}\mathrm{a}_1(0)z\right). \tag{22.4-31b}$$

Consequently, the photon-flux densities $\phi_1(z) = |\mathrm{a}_1(z)|^2$ and $\phi_3(z) = |\mathrm{a}_3(z)|^2$ evolve in accordance with

$$\phi_1(z) = \phi_1(0)\,\mathrm{sech}^2\left(\frac{\gamma z}{2}\right) \tag{22.4-32a}$$

$$\phi_3(z) = \tfrac{1}{2}\phi_1(0)\,\tanh^2\left(\frac{\gamma z}{2}\right), \tag{22.4-32b}$$

where $\gamma/2 = \mathrm{g}\mathrm{a}_1(0)/\sqrt{2}$, i.e.,

$$\gamma^2 = 2\mathrm{g}^2\mathrm{a}_1^2(0) = 2\mathrm{g}^2\phi_1(0) = 8\mathrm{d}^2\eta^3\hbar\omega^3\phi_1(0) = 8\mathrm{d}^2\eta^3\omega^2 I_1(0). \tag{22.4-33}$$

Since $\mathrm{sech}^2(\cdot) + \tanh^2(\cdot) = 1$, $\phi_1(z) + 2\phi_3(z) = \phi_1(0)$ is constant, indicating that at each position z, photons of wave 1 are converted to half as many photons of wave 3. The fall of $\phi_1(z)$ and the rise of $\phi_3(z)$ with z are shown in Fig. 22.4-1(*b*).

Note that there would be no inception of the interaction characterized by (22.4-29) under the initial conditions $\mathrm{a}_1(0) = 0$ and $\mathrm{a}_3(0) > 0$, so that the inverse process of spontaneous parametric downconversion (see Fig. 22.2-8) is not permitted within the confines of these classical equations.

Efficiency of SHG

The efficiency of second-harmonic generation for an interaction region of length L is

$$\eta_{\mathrm{SHG}} = \frac{I_3(L)}{I_1(0)} = \frac{\hbar\omega_3\phi_3(L)}{\hbar\omega_1\phi_1(0)} = \frac{2\phi_3(L)}{\phi_1(0)} = \tanh^2\left(\frac{\gamma L}{2}\right). \tag{22.4-34}$$

For large γL (long cell, large input intensity, or large nonlinear parameter), the efficiency approaches one. This signifies that all the input power (at frequency ω) has been transformed into power at frequency 2ω; all input photons of frequency ω are converted into half as many photons of frequency 2ω.

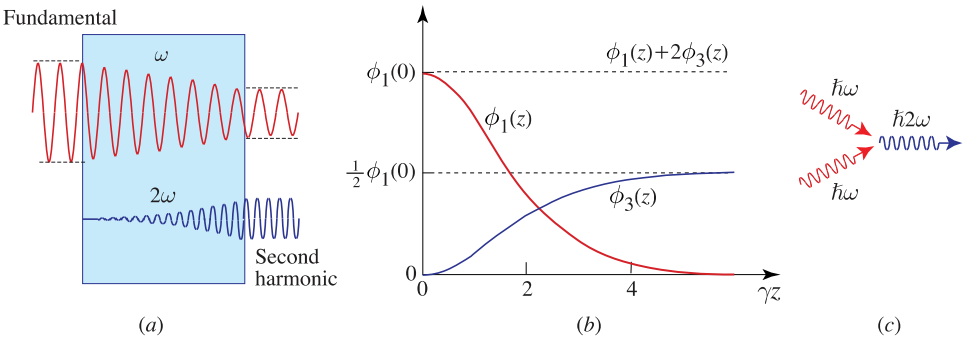


Figure 22.4-1 Second-harmonic generation. (a) A wave of frequency ω incident on a nonlinear crystal generates a wave of frequency 2ω. (b) As the photon-flux density $\phi_1(z)$ of the fundamental wave decreases, the photon-flux density $\phi_3(z)$ of the second-harmonic wave increases. Since photon numbers are conserved, the sum $\phi_1(z)+2\phi_3(z)=\phi_1(0)$ is a constant. (c) Two photons of frequency ω combine to make one photon of frequency 2ω.

For small γL [small device length L, small nonlinear parameter d, or small input photon-flux density $\phi_1(0)$], the argument of the tanh function is small and therefore the approximation $\tanh x \approx x$ may be used. The efficiency of second-harmonic generation is then

$$\eta_{\mathrm{SHG}} = \frac{I_3(L)}{I_1(0)} \approx \tfrac{1}{4}\gamma^2 L^2 = \tfrac{1}{2}\mathrm{g}^2 L^2 \phi_1(0) = 2\mathrm{d}^2\eta^3\hbar\omega^3 L^2\phi_1(0) = 2\mathrm{d}^2\eta^3\omega^2 L^2 I_1(0), \tag{22.4-35}$$

so that

$$\boxed{\eta_{\mathrm{SHG}} = C^2 \frac{L^2}{A} P, \qquad C^2 = 2\omega^2 \eta_o^3 \frac{\mathrm{d}^2}{n^3},} \tag{22.4-36}$$

SHG Efficiency

where $P = I_1(0)A$ is the incident optical power at the fundamental frequency and A is the cross-sectional area. This reproduces (22.2-6) and shows that the constant C^2 is proportional to the material parameter d^2/n^3, which is a figure of merit used for comparing different nonlinear materials.

EXAMPLE 22.4-1. ***Efficiency of SHG.*** For a material with $\mathrm{d}^2/n^3 = 10^{-46}$ C/V^2 (see Table 22.6-3 for typical values of d) and a fundamental wave of wavelength 1 μm, $C^2 = 38\times10^{-9}$ W^{-1} $= 0.038$ (MW)$^{-1}$. In this case, the SHG efficiency is 10% if $PL^2/A = 2.63$ MW. If the aspect ratio of the interaction volume is 1000, i.e., $L^2/A = 10^6$, the required power is 2.63 W. This may be realized using L = 1 cm and A = 100 μm^2, corresponding to a power density $P/A = 2.63\times10^6$ W/cm^2. The SHG efficiency may be improved by using higher power density, longer interaction length, or material with a larger value of d^2/n^3.

Phase Mismatch in SHG

To study the effect of phase (or momentum) mismatch, the general equations (22.4-27) are used with $\Delta k \neq 0$. For simplicity, we limit ourselves to the weak-coupling case for which $\gamma L \ll 1$. In this case, the amplitude of the fundamental wave $\mathrm{a}_1(z)$ varies only slightly with z [see Fig. 22.4-1(a)], and may be assumed approximately constant.

Substituting $\mathcal{a}_1(z) \approx \mathcal{a}_1(0)$ in (22.4-27b), and integrating, we obtain

$$\mathcal{a}_3(L) = -j\frac{\mathcal{g}}{2}\mathcal{a}_1^2(0)\int_0^L \exp(j\,\Delta kz)\,dz = -\left(\frac{\mathcal{g}}{2\,\Delta k}\right)\mathcal{a}_1^2(0)[\exp(j\Delta kL) - 1], \tag{22.4-37}$$

from which $\phi_3(L) = |\mathcal{a}_3(L)|^2 = (\mathcal{g}/\Delta k)^2\phi_1^2(0)\sin^2(\Delta kL/2)$, where $\mathcal{a}_1(0)$ is assumed to be real. The efficiency of second-harmonic generation is therefore

$$\eta_{\text{SHG}} = \frac{I_3(L)}{I_1(0)} = \frac{2\phi_3(L)}{\phi_1(0)} = C^2\frac{L^2}{A}P\,\text{sinc}^2(\Delta kL/2\pi), \tag{22.4-38}$$

where $\text{sinc}(x) = \sin(\pi x)/(\pi x)$.

The effect of phase mismatch is therefore to reduce the efficiency of second-harmonic generation by the factor $\text{sinc}^2(\Delta kL/2\pi)$. This confirms the previous results displayed in Fig. 22.2-14. For a given mismatch Δk, the process of SHG is efficient for lengths smaller than the coherence length $L_c = 2\pi/|\Delta k|$.

B. Optical Frequency Conversion (OFC)

A frequency up-converter [Fig. 22.4-2(a)] converts a wave of frequency ω_1 into a wave of higher frequency ω_3 by use of an auxiliary wave at frequency ω_2, called the **pump**. A photon $\hbar\omega_2$ from the pump is added to a photon $\hbar\omega_1$ from the **signal** to form a photon $\hbar\omega_3$ of the **up-converted signal** at an up-converted frequency $\omega_3 = \omega_1 + \omega_2$.

The conversion process is governed by the three coupled equations (22.4-20). For simplicity, assume that the three waves are phase matched ($\Delta k = 0$) and that the pump is sufficiently strong so that its amplitude does not change appreciably within the interaction distance of interest; i.e., $\mathcal{a}_2(z) \approx \mathcal{a}_2(0)$ for all z between 0 and L. The three equations (22.4-20) then reduce to two,

$$\boxed{\begin{aligned}\frac{d\mathcal{a}_1}{dz} &= -j\frac{\gamma}{2}\mathcal{a}_3 \qquad (22.4\text{-}39a)\\ \frac{d\mathcal{a}_3}{dz} &= -j\frac{\gamma}{2}\mathcal{a}_1, \qquad (22.4\text{-}39b)\end{aligned}}$$

where $\gamma = 2\mathcal{g}\mathcal{a}_2(0)$ and $\mathcal{a}_2(0)$ is assumed real. These are simple differential equations with harmonic solutions

$$\mathcal{a}_1(z) = \mathcal{a}_1(0)\cos\left(\frac{\gamma z}{2}\right) \tag{22.4-40a}$$

$$\mathcal{a}_3(z) = -j\mathcal{a}_1(0)\sin\left(\frac{\gamma z}{2}\right). \tag{22.4-40b}$$

The corresponding photon-flux densities are

$$\boxed{\begin{aligned}\phi_1(z) &= \phi_1(0)\cos^2\left(\frac{\gamma z}{2}\right) \qquad (22.4\text{-}41a)\\ \phi_3(z) &= \phi_1(0)\sin^2\left(\frac{\gamma z}{2}\right). \qquad (22.4\text{-}41b)\end{aligned}}$$

The dependencies of the photon-flux densities ϕ_1 and ϕ_3 on z are sketched in Fig. 22.4-2(b). Photons are exchanged periodically between the two waves. In the region between $z = 0$ and $z = \pi/\gamma$, the input ω_1 photons combine with the pump ω_2 photons and generate the up-converted ω_3 photons. Wave 1 is therefore attenuated, whereas wave 3 is amplified. In the region $z = \pi/\gamma$ to $z = 2\pi/\gamma$, the ω_3 photons are more abundant; they disintegrate into ω_1 and ω_2 photons, so that wave 3 is attenuated and wave 1 amplified. The process is repeated periodically as the waves travel through the medium.

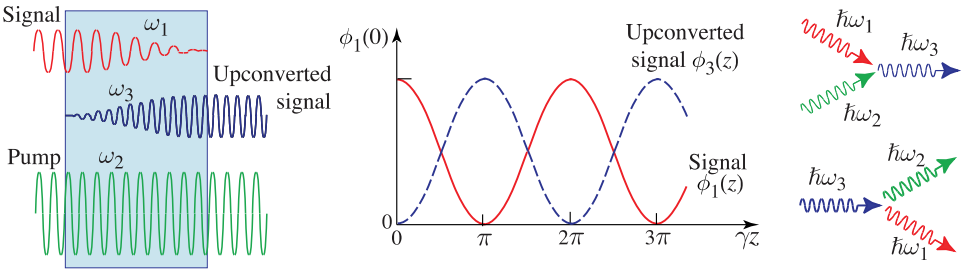


Figure 22.4-2 The frequency up-converter; (a) wave mixing; (b) evolution of the photon-flux densities of the input ω_1-wave and the up-converted ω_3-wave. The pump ω_2-wave is assumed constant; (c) photon interactions.

The efficiency of up-conversion for a device of length L is

$$\eta_{\text{OFC}} = \frac{I_3(L)}{I_1(0)} = \frac{\omega_3}{\omega_1}\sin^2\left(\frac{\gamma L}{2}\right). \tag{22.4-42}$$

For $\gamma L \ll 1$, and using (22.4-21), this is approximated by $I_3(L)/I_1(0) \approx (\omega_3/\omega_1)(\gamma L/2)^2 = (\omega_3/\omega_1)\mathrm{g}^2L^2\phi_2(0) = 2\omega_3^2L^2\mathrm{d}^2\eta^3I_2(0)$ from which

$$\boxed{\eta_{\text{OFC}} = C^2\frac{L^2}{A}P_2, \qquad C^2 = 2\omega_3^2\,\eta_o^3\frac{\mathrm{d}^2}{n^3},} \tag{22.4-43}$$

OFC Efficiency

where A is the cross-sectional area and $P_2 = I_2(0)A$ is the pump power. This expression is similar to (22.4-36) for the efficiency of second-harmonic generation.

EXERCISE 22.4-5

Infrared Up-Conversion. An up-converter uses a proustite crystal ($\mathrm{d} = 1.5 \times 10^{-22}$ C/V^2, $n = 2.6$, $\mathrm{d}^2/n^3 = 1.3 \times 10^{-45}$ C^2/V^4). The input wave is obtained from a CO_2 laser of wavelength $\lambda_o = 10.6\,\mu$m, and the pump from a 1-W Nd^{3+}:YAG laser of wavelength $\lambda_o = 1.06\,\mu$m focused to a cross-sectional area 10^{-2} mm^2 (see Fig. 22.2-6). Determine the wavelength of the up-converted wave and the efficiency of up-conversion if the waves are collinear and the interaction length is 1 cm.

C. Optical Parametric Amplification (OPA) and Oscillation (OPO)

Optical Parametric Amplifier (OPA)

The OPA uses three-wave mixing in a nonlinear crystal to provide optical gain [Fig. 22.4-3(a)]. The process is governed by the same three coupled equations (22.4-20) with the waves identified as follows. Wave 1 is the **signal** to be amplified; it is incident on the crystal with a small intensity $I_1(0)$. Wave 3, the **pump**, is an intense wave that provides power to the amplifier. Wave 2, called the **idler**, is an auxiliary wave created by the interaction process.

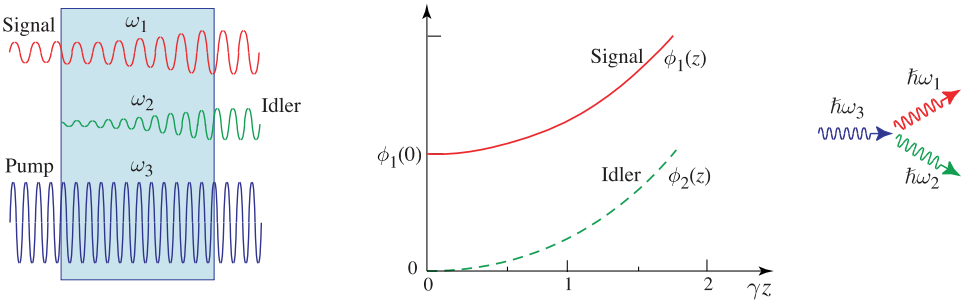


Figure 22.4-3 The optical parametric amplifier: (a) wave mixing; (b) photon-flux densities of the signal and the idler (the pump photon-flux density is assumed constant); (c) photon mixing.

Assuming perfect phase matching ($\Delta k = 0$), and an undepleted pump, $\mathrm{a}_3(z) \approx \mathrm{a}_3(0)$, the coupled-wave equations (22.4-20) provide

$$\frac{d\mathrm{a}_1}{dz} = -j\frac{\gamma}{2}\mathrm{a}_2^* \tag{22.4-44a}$$

$$\frac{d\mathrm{a}_2}{dz} = -j\frac{\gamma}{2}\mathrm{a}_1^*, \tag{22.4-44b}$$

where $\gamma = 2g\mathrm{a}_3(0)$. If $\mathrm{a}_3(0)$ is real, γ is also real, and the differential equations have the solution

$$\mathrm{a}_1(z) = \mathrm{a}_1(0)\cosh\left(\frac{\gamma z}{2}\right) - j\mathrm{a}_2^*(0)\sinh\left(\frac{\gamma z}{2}\right) \tag{22.4-45a}$$

$$\mathrm{a}_2(z) = -j\mathrm{a}_1^*(0)\sinh\left(\frac{\gamma z}{2}\right) + \mathrm{a}_2(0)\cosh\left(\frac{\gamma z}{2}\right). \tag{22.4-45b}$$

If $\mathrm{a}_2(0) = 0$, i.e., the initial idler field is zero, then the corresponding photon-flux densities are

$$\phi_1(z) = \phi_1(0)\cosh^2\left(\frac{\gamma z}{2}\right) \tag{22.4-46a}$$

$$\phi_2(z) = \phi_1(0)\sinh^2\left(\frac{\gamma z}{2}\right). \tag{22.4-46b}$$

Both $\phi_1(z)$ and $\phi_2(z)$ grow monotonically with z, as illustrated in Fig. 22.4-3(b). This growth saturates when sufficient energy is drawn from the pump so that the assumption of an undepleted pump no longer holds.

The overall gain of an amplifier of length L is $G = \phi_1(L)/\phi_1(0) = \cosh^2(\gamma L/2)$. In the limit $\gamma L \gg 1, G = (e^{\gamma L/2} + e^{-\gamma L/2})^2/4 \approx e^{\gamma L}/4$, so that the gain increases exponentially with γL. The gain coefficient $\gamma = 2\mathrm{g}\mathfrak{a}_3(0) = 2\mathrm{d}\sqrt{2\hbar\omega_1\omega_2\omega_3\eta^3}\,\mathfrak{a}_3(0)$, from which

$$\gamma = 2C\sqrt{I_3(0)} = 2C\sqrt{P_3/A}, \qquad C^2 = 2\omega_1\omega_2\,\eta_o^3\frac{\mathrm{d}^2}{n^3}, \tag{22.4-47}$$

OPA Gain Coefficient

where $P_3 = I_3(0)A$ is the pump power and A is the cross-sectional area, and C^2 is a parameter similar to that describing SHG and OFC.

The interaction is tantamount to a pump photon $\hbar\omega_3$ splitting into a photon $\hbar\omega_1$ that amplifies the signal, and a photon $\hbar\omega_2$ that creates the idler [Fig. 22.4-3(c)].

EXERCISE 22.4-6

Gain of an OPA. An OPA amplifies light at $\lambda_o = 2.5\,\mu$m by using a 2-cm long KTP crystal pumped by a Nd:YAG laser with $\lambda_o = 1.064\,\mu$m. Determine the wavelength of the idler wave and the C coefficient in (22.4-47). Determine the appropriate laser power and beam cross-sectional area such that the total amplifier gain is 3 dB. Assume that $n = 1.75$ and $\mathrm{d} = 2.3 \times 10^{-23}$ C/V^2 for KTP.

Optical Parametric Oscillator (OPO)

A parametric oscillator is constructed by providing feedback at either or both the signal and the idler frequencies of a parametric amplifier, as illustrated in Fig. 22.4-4. In the former case, the oscillator is called a **singly resonant oscillator** (**SRO**); in the latter, it is called a **doubly resonant oscillator** (**DRO**).

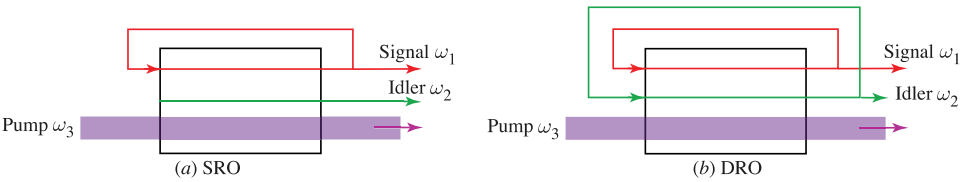


Figure 22.4-4 The parametric oscillator generates light at frequencies ω_1 and ω_2. A pump of frequency $\omega_3 = \omega_1 + \omega_2$ serves as the source of energy. (a) Singly resonant oscillator (SRO). (b) Doubly resonant oscillator (DRO).

The oscillation frequencies ω_1 and ω_2 of the parametric oscillator are determined by the frequency- and phase-matching conditions, $\omega_1 + \omega_2 = \omega_3$ and $n_1\omega_1 + n_2\omega_2 = n_3\omega_3$, in the collinear case. The solution of these two equations yields ω_1 and ω_2, as described in Sec. 22.2D. In addition, these frequencies must also coincide with the resonance frequencies of the resonator modes, much the same as for conventional lasers (see Sec. 16.1B). The system therefore tends to be over-constrained, particularly in the DRO case for which both the signal and idler frequencies must coincide with resonator modes.

Another condition for oscillation is that the gain of the amplifier must exceed the loss introduced by the mirrors for one round trip of propagation within the resonator. By equating the gain and the loss, expressions for the threshold amplifier gain and the corresponding threshold pump intensity may be determined, as shown below for the SRO and DRO configurations.

SRO. At the threshold of oscillation, the signal's amplified and doubly reflected amplitude $\mathcal{a}_1(L)\,\mathrm{r}_1^2$ equals the initial amplitude $\mathcal{a}_1(0)$, where L is the length of the nonlinear medium and r_1 is the magnitude of the amplitude reflectance of a mirror (the two mirrors are assumed identical and the phase associated with a round trip is not included since it is a multiple of 2π). Using (22.4-45a), together with the boundary condition $\mathcal{a}_2(0) = 0$, we obtain $\mathrm{r}_1^2 \cosh(\gamma L/2) = 1$, from which

$$\cosh^2\left(\frac{\gamma L}{2}\right) = \frac{1}{\mathcal{R}_1^2}\,. \tag{22.4-48}$$

Here, $\mathcal{R}_1 = \mathrm{r}_1^2$ is the mirror power reflectance at the signal frequency. Since $\mathcal{R}_1$ is typically slightly smaller than unity, $\cosh^2(\gamma L/2)$ is slightly greater than unity, i.e., $\gamma L/2 \ll 1$ and the approximation $\cosh^2(x) \approx 1 + x^2$ may be used. It follows that at threshold $(\gamma L/2)^2 \approx (1 - \mathcal{R}_1^2)/\mathcal{R}_1^2$. Using (22.4-47), we obtain the threshold intensity, from which the threshold power of the pump is obtained,

$$\boxed{P_3|_{\text{threshold}}(0) \approx \frac{1}{C^2}\,\frac{A}{L^2}\,\frac{1-\mathcal{R}_1^2}{\mathcal{R}_1^2},} \tag{22.4-49}$$

SRO Threshold Pump Power

where $C^2 = 2\omega_1\omega_2\,\eta_o^3 \mathrm{d}^2/n^3$ and A is the cross-sectional area. For example, if $L^2/A = 10^6$, $C^2 = 10^{-7}$ W^{-1}, and $\mathcal{R}_1 = 0.9$, then $P_3|_{\text{threshold}}(0) \approx 2.3$ W.

DRO. At threshold, two conditions must be satisfied: $\mathcal{a}_1(L)\,\mathrm{r}_1^2 = \mathcal{a}_1(0)$ and $\mathcal{a}_2(L)\,\mathrm{r}_2^2 = \mathcal{a}_2(0)$, where r_1 and r_2 are the magnitudes of the amplitude reflectances of the mirrors at the signal and idler frequencies, respectively. Substituting for $\mathcal{a}_1(L)$ from (22.4-45a), and substituting for $\mathcal{a}_2(L)$ from (22.4-45b) and forming the conjugate, we obtain

$$\left(1 - \mathcal{R}_1 \cosh\frac{\gamma L}{2}\right)\mathcal{a}_1(0) + \left(j\mathcal{R}_1 \sinh\frac{\gamma L}{2}\right)\mathcal{a}_2^*(0) = 0 \tag{22.4-50a}$$

$$\left(-j\mathcal{R}_2 \sinh\frac{\gamma L}{2}\right)\mathcal{a}_1(0) + \left(1 - \mathcal{R}_2 \cosh\frac{\gamma L}{2}\right)\mathcal{a}_2^*(0) = 0, \tag{22.4-50b}$$

where $\mathcal{R}_1 = \mathrm{r}_1^2$ and $\mathcal{R}_2 = \mathrm{r}_2^2$ are the power reflectances of the mirrors at the signal and idler frequencies, respectively. Equating the values of the ratio $\mathcal{a}_1(0)/\mathcal{a}_2^*(0)$ obtained from (22.4-50a) and (22.4-50b) then leads to

$$\cosh\left(\frac{\gamma L}{2}\right) = \frac{1 + \mathcal{R}_1\mathcal{R}_2}{\mathcal{R}_1 + \mathcal{R}_2}. \tag{22.4-51}$$

If $\gamma L/2 \ll 1$, we can again use the approximation $\cosh^2 x \approx 1 + x^2$ and write $(\gamma L/2)^2 \approx (1 - \mathcal{R}_1^2)(1 - \mathcal{R}_2^2)/(\mathcal{R}_1 + \mathcal{R}_2)^2$, from which we obtain the threshold pump

power:

$$P_3|_{\text{threshold}}(0) \approx \frac{1}{C^2}\frac{A}{L^2}\frac{(1-\mathcal{R}_1^2)(1-\mathcal{R}_2^2)}{(\mathcal{R}_1+\mathcal{R}_2)^2}. \tag{22.4-52}$$

DRO Threshold Pump Power

The ratio of the threshold pump power for the DRO configuration, to that for the SRO configuration, as calculated from (22.4-49) and (22.4-52), is then determined to be $\mathcal{R}_1^2(1-\mathcal{R}_2^2)/(\mathcal{R}_1+\mathcal{R}_2)^2$. Since $\mathcal{R}_1 \approx 1$ and $\mathcal{R}_2 \approx 1$, this is approximately equal to $(1-\mathcal{R}_2)/2$, which is small. Thus, the threshold power for the DRO is substantially lower than that for the SRO. Unfortunately, DROs are more sensitive to fluctuations of the resonator length because of the requirement that the oscillation frequencies of *both* the signal and the idler match resonator modes. DROs therefore often have poor stability and spiky spectra.

*22.5 THIRD-ORDER NONLINEAR OPTICS: COUPLED WAVES

A. Four-Wave Mixing (FWM)

We now derive the coupled differential equations that describe FWM in a third-order nonlinear medium, using an approach similar to that employed in the three-wave mixing case in Sec. 22.4.

Coupled-Wave Equations

Consider four waves constituting a total field

$$\mathcal{E}(t) = \sum_{q=1,2,3,4} \text{Re}[E_q \exp(j\omega_q t)] = \sum_{q=\pm1,\pm2,\pm3,\pm4} \tfrac{1}{2} E_q \exp(j\omega_q t) \tag{22.5-1}$$

traveling in a medium characterized by a nonlinear polarization density

$$\mathcal{P}_{\text{NL}} = 4\chi^{(3)}\mathcal{E}^3. \tag{22.5-2}$$

The corresponding source of radiation, $\mathcal{S} = -\mu_o \partial^2 \mathcal{P}_{\text{NL}}/\partial t^2$, is therefore a sum of $8^3 = 512$ terms,

$$\mathcal{S} = \tfrac{1}{2}\mu_o \chi^{(3)} \sum_{q,p,r=\pm1,\pm2,\pm3,\pm4} (\omega_q+\omega_p+\omega_r)^2 E_q E_p E_r \exp[j(\omega_q+\omega_p+\omega_r)t]. \tag{22.5-3}$$

Substituting (22.5-1) and (22.5-3) into the wave equation (22.4-1), and equating terms at each of the four frequencies ω_1, ω_2, ω_3, and ω_4, leads to four Helmholtz equations with their associated sources,

$$(\nabla^2 + k_q^2)E_q = -S_q, \qquad q = 1,2,3,4, \tag{22.5-4}$$

where S_q is the complex amplitude of the component of $\mathcal{S}$ at frequency ω_q.

For the four waves to be coupled, their frequencies must be commensurate. Consider, for example, the case for which the sum of two frequencies equals the sum of the other two frequencies,

$$\omega_1 + \omega_2 = \omega_3 + \omega_4, \tag{22.5-5}$$

and assume that these frequencies are distinct. Three waves can then combine and create a source at the fourth frequency. Using (22.5-5), terms in (22.5-3) at each of the four frequencies are

$$S_1 = \mu_o \omega_1^2 \chi^{(3)} \{6E_3 E_4 E_2^* + 3E_1[|E_1|^2 + 2|E_2|^2 + 2|E_3|^2 + 2|E_4|^2]\} \tag{22.5-6a}$$

$$S_2 = \mu_o \omega_2^2 \chi^{(3)} \{6E_3 E_4 E_1^* + 3E_2[|E_2|^2 + 2|E_1|^2 + 2|E_3|^2 + 2|E_4|^2]\} \tag{22.5-6b}$$

$$S_3 = \mu_o \omega_3^2 \chi^{(3)} \{6E_1 E_2 E_4^* + 3E_3[|E_3|^2 + 2|E_2|^2 + 2|E_1|^2 + 2|E_4|^2]\} \tag{22.5-6c}$$

$$S_4 = \mu_o \omega_4^2 \chi^{(3)} \{6E_1 E_2 E_3^* + 3E_4[|E_4|^2 + 2|E_1|^2 + 2|E_2|^2 + 2|E_3|^2]\}. \tag{22.5-6d}$$

Each wave is therefore driven by a source with two components. The first component is a result of mixing of the other three waves. The first term in S_1, for example, is proportional to $E_3 E_4 E_2^*$ and therefore represents the mixing of waves 2, 3, and 4 to create a source for wave 1. The second component is proportional to the complex amplitude of the wave itself. The second term of S_1, for example, is proportional to E_1, so that it plays the role of refractive-index modulation, and therefore represents the optical Kerr effect (Exercise 22.3-3).

It is therefore convenient to separate the two contributions to these sources by defining

$$S_q = \overline{S}_q + (\omega_q / c_o)^2 \Delta\chi_q E_q, \qquad q = 1, 2, 3, 4 \tag{22.5-7}$$

where

$$\overline{S}_1 = 6\mu_o \omega_1^2 \chi^{(3)} E_3 E_4 E_2^* \tag{22.5-8a}$$

$$\overline{S}_2 = 6\mu_o \omega_2^2 \chi^{(3)} E_3 E_4 E_1^* \tag{22.5-8b}$$

$$\overline{S}_3 = 6\mu_o \omega_3^2 \chi^{(3)} E_1 E_2 E_4^* \tag{22.5-8c}$$

$$\overline{S}_4 = 6\mu_o \omega_4^2 \chi^{(3)} E_1 E_2 E_3^*, \tag{22.5-8d}$$

and

$$\Delta\chi_q = 6\frac{\eta}{\epsilon_o}\chi^{(3)}(2I - I_q), \qquad q = 1, 2, 3, 4. \tag{22.5-9}$$

Here $I_q = |E_q|^2/2\eta$ are the intensities of the waves, $I = I_1 + I_2 + I_3 + I_4$ is the total intensity, which is constant in view of conservation of energy, and η is the impedance of the medium. This enables us to rewrite the Helmholtz equations (22.5-4) as

$$(\nabla^2 + \overline{k}_q^2) E_q = -\overline{S}_q, \qquad q = 1, 2, 3, 4\,, \tag{22.5-10}$$

where

$$\overline{k}_q = \bar{n}_q \frac{\omega_q}{c_o}\,, \tag{22.5-11}$$

$$\bar{n}_q^2 = n^2 + 2nn_2(2I - I_q)\,, \tag{22.5-12}$$

and

$$n_2 = \frac{3\eta_o}{\epsilon_o n^2}\chi^{(3)}, \tag{22.5-13}$$

which matches (22.3-6). If the second term of (22.5-12) is much smaller than the first, then

$$\boxed{\bar{n}_q \approx n + n_2(2I - I_q)\,.} \tag{22.5-14}$$

Optical Kerr Effect

The Helmholtz equation for each wave is therefore modified in two ways:

1. A source representing the combined effects of the other three waves is present. This may lead to the amplification of an existing wave, or the generation of a new wave at that frequency.
2. The refractive index for each wave is altered, becoming a function of the intensities of the four waves.

These equations are used to generate four coupled nonlinear differential equations that may be solved for the fields, or their complex envelopes, under the appropriate boundary conditions. This was the approach followed for second-order nonlinear processes, and will now be applied to several special cases in third-order nonlinear processes.

B. Three-Wave Mixing and Third-Harmonic Generation (THG)

We now consider degenerate cases for which two or three of the four waves have the same frequency.

Three-Wave Mixing

In the degenerate case for which two of the four waves have the same frequency $\omega_3 = \omega_4 \equiv \omega_0$, we have three waves with frequencies related by $\omega_1 + \omega_2 = 2\omega_0$. A coupled-wave theory of this three-wave mixing process can be formulated by identifying the radiation sources generated at the three frequencies:

$$S_1 = \mu_o\omega_1^2\chi^{(3)}\left\{3E_0^2E_2^* + 3E_1\left[|E_1|^2 + 2|E_2|^2 + 2|E_0|^2\right]\right\} \tag{22.5-15a}$$

$$S_2 = \mu_o\omega_2^2\chi^{(3)}\left\{3E_0^2E_1^* + 3E_2\left[|E_2|^2 + 2|E_1|^2 + 2|E_0|^2\right]\right\} \tag{22.5-15b}$$

$$S_0 = \mu_o\omega_0^2\chi^{(3)}\left\{6E_1E_2E_0^* + 3E_0\left[|E_0|^2 + 2|E_1|^2 + 2|E_2|^2\right]\right\}. \tag{22.5-15c}$$

When substituted in the Helmholtz equations $(\nabla^2 + k_q^2)E_q = -S_q,\ q = 0, 1, 2$, the result is a set of coupled equations that can, in principle, be solved under appropriate initial conditions.

Collinear waves traveling in the z direction are written $E_q(\mathbf{r}) = A_q\exp(-jk_qz)$. As with second-order nonlinear processes, we use the slowly varying envelope approximation, $(\nabla^2 + k_q^2)[A_q\exp(-jk_qz)] \approx -j2k_q(dA_q/dz)\exp(-jk_qz)$, and write the complex amplitudes $A_q = \sqrt{2\eta\hbar\omega_q}\,\mathrm{a}_q$, in terms of the variables a_q, which are normalized such that $\phi_q = |\mathrm{a}_q|^2$ are photon-flux densities. The analysis is simplified by assuming that $\omega_1 \approx \omega_2 \approx \omega_0$ when calculating the coupling coefficients. The result is the following set of coupled equations:

$$\frac{d\mathrm{a}_1}{dz} = -j\mathrm{g}\left[\mathrm{a}_0^2\mathrm{a}_2^*\exp(-j\Delta k\,z) + \mathrm{a}_1\left(|\mathrm{a}_1|^2 + 2|\mathrm{a}_2|^2 + 2|\mathrm{a}_0|^2\right)\right] \tag{22.5-16a}$$

$$\frac{d\mathrm{a}_2}{dz} = -j\mathrm{g}\left[\mathrm{a}_0^2\mathrm{a}_1^* \exp(-j\Delta k\, z) + \mathrm{a}_2\left(|\mathrm{a}_2|^2 + 2|\mathrm{a}_1|^2 + 2|\mathrm{a}_0|^2\right)\right] \tag{22.5-16b}$$

$$\frac{d\mathrm{a}_0}{dz} = -j\mathrm{g}\left[2\mathrm{a}_1\mathrm{a}_2\mathrm{a}_0^* \exp(j\Delta k\, z) + \mathrm{a}_0\left(|\mathrm{a}_0|^2 + 2|\mathrm{a}_1|^2 + 2|\mathrm{a}_2|^2\right)\right], \tag{22.5-16c}$$

where

$$\mathrm{g} = \hbar\omega_0(\omega_0/c_o)n_2, \tag{22.5-17}$$

and

$$\Delta k = 2k_0 - k_1 - k_2 \tag{22.5-18}$$

represents the phase-matching error.

This set of nonlinear equations can be readily solved in the undepleted pump approximation ($|\mathrm{a}_1|, |\mathrm{a}_2| \ll |\mathrm{a}_0|$) since in this case $\mathrm{a}_0(z)$ is approximately constant. In the phase matched case ($\Delta k = 0$), (22.5-16) are approximated by two linear differential equations

$$\frac{d\mathrm{a}_1}{dz} = -j\gamma(\mathrm{a}_2^* + 2\mathrm{a}_1) \tag{22.5-19a}$$

$$\frac{d\mathrm{a}_2}{dz} = -j\gamma(\mathrm{a}_1^* + 2\mathrm{a}_2), \tag{22.5-19b}$$

where $\gamma = \mathrm{g}\mathrm{a}_0^2$ is a constant proportional to the constant pump intensity. The solution to these equations is written in terms of the initial values of the two waves:

$$\mathrm{a}_1(z) = [(1 - j\gamma z)\mathrm{a}_1(0) - j\gamma z\mathrm{a}_2^*(0)] \exp(-j\gamma z) \tag{22.5-20a}$$

$$\mathrm{a}_2(z) = [-j\gamma z\mathrm{a}_1^*(0) + (1 - j\gamma z)\mathrm{a}_2(0)] \exp(-j\gamma z). \tag{22.5-20b}$$

If the initial idler amplitude is $\mathrm{a}_2(0) = 0$, then the photon-flux density $\phi_1(z) = |a_1(z)|^2$ of the signal grows as $\phi_1(z) = (1 + \gamma^2 z^2)\phi_1(0)$. The rate of growth is sensitive to the magnitude and phase of the initial idler wave. For example, if $\mathrm{a}_2(0) = \mathrm{r}e^{j\varphi}\mathrm{a}_1(0)$, then

$$\phi_1(z) = \left[1 + (2\mathrm{r}\sin\varphi)\gamma z + (1 + \mathrm{r}^2 + 2\mathrm{r}\cos\varphi)\gamma^2 z^2\right]\phi_1(0), \tag{22.5-21}$$

which is a function of the phase difference φ that reaches its maximum value when $\tan\varphi = 2/\gamma z$. At small z, maximum growth occurs when $\varphi = \pi/2$. Clearly, the amplifier is a **phase-sensitive amplifier**.

To examine the effect of pump depletion and phase mismatch, the full set of equations (22.5-16) must be solved. One step in this direction is taken by writing the complex amplitudes $\mathrm{a}_q = \mathrm{b}_q \exp(j\varphi_q)$ in terms of their magnitudes b_q and phases φ_q. Substituting into (22.5-16) and equating the real and imaginary parts of each equation leads to the following set of nonlinear equations in real variables:

$$\frac{d\mathrm{b}_1}{dz} = \mathrm{g}\mathrm{b}_0^2\mathrm{b}_2 \sin\varphi \tag{22.5-22a}$$

$$\frac{d\mathrm{b}_2}{dz} = \mathrm{g}\mathrm{b}_0^2\mathrm{b}_1 \sin\varphi \tag{22.5-22b}$$

$$\frac{d\mathrm{b}_0}{dz} = -\mathrm{g}\mathrm{b}_0\mathrm{b}_1\mathrm{b}_2 \sin\varphi \tag{22.5-22c}$$

$$\frac{d\varphi}{dz} = \Delta k + \mathrm{g}\left[2\mathrm{b}_0^2 - \mathrm{b}_1^2 - \mathrm{b}_2^2\right] + \mathrm{g}\left[\mathrm{b}_0^2\mathrm{b}_1/\mathrm{b}_2 + \mathrm{b}_0^2\mathrm{b}_2/\mathrm{b}_1 - 4\mathrm{b}_1\mathrm{b}_2\right]\cos\varphi, \tag{22.5-22d}$$

where $\varphi = \Delta k\, z + \varphi_1 + \varphi_2 - 2\varphi_0$. Two invariants can be easily identified. Consistent with conservation of optical intensity, the sum $\mathrm{b}_1^2 + \mathrm{b}_2^2 + \mathrm{b}_0^2$ must be constant. Also, consistent with conservation of photons, the difference $\mathrm{b}_1^2 - \mathrm{b}_2^2$ must be constant (this is a version of the Manley–Rowe relations). Other invariants involving the phase φ may also be identified[†] and used to study the role of phase mismatch and initial amplitudes and phase difference between the signal and idler. For example, it can be readily seen from (22.5-22a) that the initial rate of growth of the signal occurs when $\sin\varphi = 0$, i.e., when $\varphi = \pi/2$.

Third-Harmonic Generation (THG)

Another degenerate special case of four-wave mixing is third-harmonic generation. Here, three of the four waves have identical frequencies, $\omega_1 = \omega_2 = \omega_4 = \omega$, and the fourth has the sum frequency $\omega_3 = \omega_1 + \omega_2 + \omega_4 = 3\omega$. In effect, we have two waves, 1 and 3, whose amplitudes are coupled by the third-order nonlinear medium. A coupled-wave theory can be formulated using the approach followed in the four- and three-wave mixing cases. This leads to two Helmholtz equations $(\nabla^2 + k_q^2)E_q = -S_q$, where

$$S_1 = \mu_o \omega_1^2 \chi^{(3)} \left\{ 3E_3 E_1^* E_1^* + 3E_1 \left[|E_1|^2 + 2|E_3|^2 \right] \right\} \tag{22.5-23a}$$

$$S_3 = \mu_o \omega_3^2 \chi^{(3)} \left\{ E_1^3 + 3E_3 \left[|E_3|^2 + 2|E_1|^2 \right] \right\}. \tag{22.5-23b}$$

These equations may be used to derive coupled equations for E_1 and E_3, as was done in previous cases.

EXERCISE 22.5-1

THG in the Undepleted-Pump Approximation. Assume that the fundamental and third-harmonic waves are plane waves traveling in the z direction with complex envelopes A_q, $q = 1, 3$. Use the slowly varying envelope approximation to write coupled differential equations for A_1 and A_3. Show that in the undepleted pump approximation [$A_3 \ll A_1$ and $A_1(z) \approx A_1(0)$],

$$\frac{d\mathrm{a}_3}{dz} = -j\mathrm{g}\mathrm{a}_1^3 \exp(-j\Delta k z), \tag{22.5-24}$$

where $A_q = \sqrt{2\eta\hbar\omega_q}\,\mathrm{a}_q$ and $\Delta k = 3k_1 - k_3$. Derive an expression for g.

C. Optical Phase Conjugation (OPC)

We now develop and solve the coupled-wave equations in the fully degenerate case for which all four waves have the same frequency $\omega_1 = \omega_2 = \omega_3 = \omega_4 = \omega$. As was assumed in Sec. 22.3E, two of the waves (waves 3 and 4), called the pump waves, are plane waves propagating in opposite directions, with complex amplitudes $E_3(\mathbf{r}) = A_3 \exp(-j\mathbf{k}_3\cdot\mathbf{r})$ and $E_4(\mathbf{r}) = A_4 \exp(-j\mathbf{k}_4\cdot\mathbf{r})$ and wavevectors related by $\mathbf{k}_4 = -\mathbf{k}_3$. Their intensities are assumed to be much greater than those of waves 1 and 2, so that they are approximately undepleted by the interaction process, allowing us to assume that their complex envelopes A_3 and A_4 are constant. The total intensity of the four waves I is then also approximately constant, $I \approx [|A_3|^2 + |A_4|^2]/2\eta$. The terms $2I - I_1$ and $2I - I_2$, which govern the effective refractive index $\bar{n}$ for waves 1 and 2 in (22.5-14), are approximately equal to $2I$, and are therefore also constant, so that the optical Kerr effect amounts to a constant change of the refractive index. Its effect will therefore be ignored.

[†] See, e.g., G. Cappellini and S. Trillo, Third-Order Three-Wave Mixing in Single-Mode Fibers: Exact Solutions and Spatial Instability Effects, *Journal of the Optical Society of America B*, vol. 8, pp. 824–838, 1991.

With these assumptions the problem is reduced to a problem of two coupled waves, 1 and 2. Equations (22.5-10) and (22.5-8) give

$$(\nabla^2 + k^2)E_1 = -\xi E_2^* \tag{22.5-25a}$$
$$(\nabla^2 + k^2)E_2 = -\xi E_1^*, \tag{22.5-25b}$$

where

$$\xi = 6\mu_o\omega^2\chi^{(3)}E_3E_4 = 6\mu_o\omega^2\chi^{(3)}A_3A_4 \tag{22.5-26}$$

and $k = \bar{n}\omega/c_o$, where $\bar{n} \approx n + 2n_2I$ is a constant.

The four nonlinear coupled differential equations have thus been reduced to two *linear* coupled equations, each of which takes the form of the Helmholtz equation with a source term. The source for wave 1 is proportional to the conjugate of the complex amplitude of wave 2, and similarly for wave 2.

Phase Conjugation

Assume that waves 1 and 2 are also plane waves propagating in opposite directions along the z axis, as illustrated in Fig. 22.5-1,

$$E_1 = A_1\exp(-jkz), \qquad E_2 = A_2\exp(jkz). \tag{22.5-27}$$

This assumption is consistent with the phase-matching condition since $k_1 + k_2 = k_3 + k_4$.

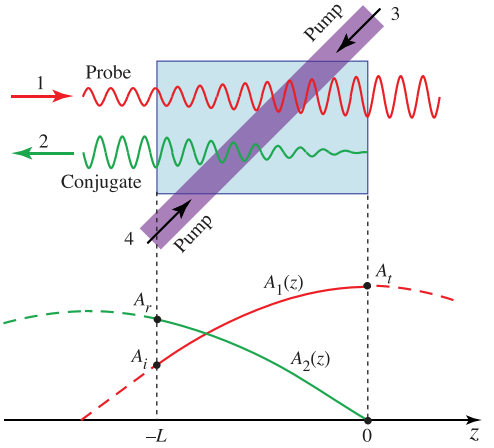


Figure 22.5-1 Degenerate four-wave mixing. Waves 3 and 4 are intense pump waves traveling in opposite directions. Wave 1, the probe wave, and wave 2, the conjugate wave, also travel in opposite directions and have increasing amplitudes.

Substituting (22.5-27) in (22.5-25) and using the slowly varying envelope approximation, (22.4-19), we reduce equations (22.5-25) to two first-order differential equations,

$$\frac{dA_1}{dz} = -j\gamma A_2^* \tag{22.5-28a}$$
$$\frac{dA_2}{dz} = j\gamma A_1^*, \tag{22.5-28b}$$

where

$$\gamma = \frac{\xi}{2k} = 3\omega\eta_o \frac{\chi^{(3)}}{\bar{n}} A_3 A_4 \tag{22.5-29}$$

is a coupling coefficient whose magnitude may be written in the form

$$|\gamma| = 2C\sqrt{I_3 I_4}\,, \qquad C = 3\omega\eta_o^2 \frac{\chi^{(3)}}{\bar{n}^2}\,. \tag{22.5-30}$$

Coupling Coefficient

Here $I_3 = |A_3|^2/2\eta$ and $I_4 = |A_4|^2/2\eta$ are the intensities of the two waves and $\eta = \eta_0/\bar{n}$.

For simplicity, assume that $A_3 A_4$ is real, so that γ is real. The solution of (22.5-28) is then two harmonic functions, $A_1(z)$ and $A_2(z)$, with a 90° phase shift between them. If the nonlinear medium extends over a distance between the planes $z = -L$ to $z = 0$, as illustrated in Fig. 22.5-1, wave 1 has amplitude $A_1(-L) = A_i$, at the entrance plane, and wave 2 has zero amplitude at the exit plane, $A_2(0) = 0$. Under these boundary conditions the solution of (22.5-28) is

$$A_1(z) = \frac{A_i}{\cos\gamma L}\cos(\gamma z) \tag{22.5-31}$$

$$A_2(z) = j\frac{A_i^*}{\cos\gamma L}\sin(\gamma z). \tag{22.5-32}$$

The amplitude of the reflected wave at the entrance plane, $A_r = A_2(-L)$, is

$$A_r = -jA_i^* \tan(\gamma L), \tag{22.5-33}$$

Reflected Wave Amplitude

whereas the amplitude of the transmitted wave, $A_t = A_1(0)$, is

$$A_t = \frac{A_i}{\cos(\gamma L)}. \tag{22.5-34}$$

Transmitted Wave Amplitude

Equations (22.5-33) and (22.5-34) suggest a number of applications:

- The reflected wave is a conjugated version of the incident wave. The device acts as a **phase conjugator** (see Sec. 22.3E).
- The power reflectance, $|A_r|^2/|A_i|^2 = \tan^2(\gamma L)$, may be smaller or greater than 1, corresponding to attenuation or gain, respectively. The medium can therefore act as a **reflection amplifier** (an "amplifying mirror").
- The transmittance $|A_t|^2/|A_i|^2 = 1/\cos^2(\gamma L)$ is always greater than 1, so that the medium always acts as a **transmission amplifier**.
- When $\gamma L = \pi/2$, or odd multiples thereof, the reflectance and transmittance are infinite, indicating instability. The device may then be used as an **oscillator**.

*22.6 ANISOTROPIC NONLINEAR MEDIA

In an anisotropic medium, each of the three components of the polarization-density vector $\mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3)$ is generally a function of the three components of the electric-field vector $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3)$. These functions are linear for small magnitudes of $\mathcal{E}$ (see Sec. 6.3) but deviate slightly from linearity as $\mathcal{E}$ increases. They may therefore be expanded in a Taylor series in terms of the three components of $\mathcal{E}$, just as in the scalar analysis presented in Sec. 22.1:

$$\mathcal{P}_i = \epsilon_o \sum_j \chi_{ij}\mathcal{E}_j + 2\sum_{jk} \mathrm{d}_{ijk}\mathcal{E}_j\mathcal{E}_k + 4\sum_{jkl} \chi^{(3)}_{ijkl}\mathcal{E}_j\mathcal{E}_k\mathcal{E}_l \qquad i,j,k,l = 1,2,3. \tag{22.6-1}$$

The coefficients χ_{ij}, d_{ijk}, and $\chi^{(3)}_{ijkl}$ are tensor components that correspond to the scalar coefficients χ, d, and $\chi^{(3)}$, respectively, and (22.6-1) is a tensor generalization of (22.1-2). Because the component d_{ijk} of the (third-rank) **second-order optical nonlinearity tensor** is proportional to $\partial^2\mathcal{P}_i/\partial\mathcal{E}_j\partial\mathcal{E}_k$, it is invariant to exchange of j and k. Similarly, the component $\chi^{(3)}_{ijkl}$ of the (fourth-rank) **third-order optical nonlinearity tensor** is invariant to permutations of j, k, and l. For lossless nondispersive media, there are additional intrinsic symmetries: $\chi_{ij} = \chi_{ji}$, as shown in Sec. 6.3A; also d_{ijk} and $\chi^{(3)}_{ijkl}$ are invariant to permutations of their indices. This full-permutation symmetry does not generally hold for dispersive nonlinear media.

Exploiting the symmetry condition $\mathrm{d}_{ijk} = \mathrm{d}_{ikj}$, components of the tensor d_{ijk} are usually listed as a 3×6 array d_{iJ}, where the six independent combinations $(j,k) = 11, 22, 33, 23, 31, 12$ are represented by a single index $J = 1, 2, 3, 4, 5, 6$, in that order (see Table 21.2-1). For example, d_{25} denotes the coefficients $\mathrm{d}_{231} = \mathrm{d}_{213}$.

The third-order coefficients $\chi^{(3)}_{ijkl}$ are similarly described by a 6×6 array $\chi^{(3)}_{IK}$, where the pair (i,j) is contracted into a single index $I = 1, 2, \cdots, 6$, and the pair (k,l) is contracted into $K = 1, 2, \cdots, 6$.

The structural symmetry of the crystal places additional constraints on the tensor components d_{ijk} and $\chi^{(3)}_{ijkl}$. When the coordinate system (1,2,3) coincides with the principal axes of the crystal, which are determined from the tensor χ_{ij}, some entries in the arrays d_{iJ} and $\chi^{(3)}_{IK}$ are zero, while others are equal or are related by some simple rule. Representative examples are provided in Tables 22.6-1 and 22.6-2. Values for the d_{iJ} coefficients for a number of representative nonlinear crystals are provided in Table 22.6-3. Though cubic crystals have isotropic linear optical properties, their well-defined crystal axes (as determined by their structural symmetry) endow them with anisotropic nonlinear optical properties.

Table 22.6-1 Second-order nonlinear coefficients d_{iJ} for some representative crystal groups.

$\begin{bmatrix} 0 & 0 & 0 & \mathrm{d}_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{d}_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{d}_{14} \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 & \mathrm{d}_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{d}_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{d}_{36} \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 & 0 & \mathrm{d}_{15} & -\mathrm{d}_{22} \\ -\mathrm{d}_{22} & \mathrm{d}_{22} & 0 & \mathrm{d}_{15} & 0 & 0 \\ \mathrm{d}_{31} & \mathrm{d}_{31} & \mathrm{d}_{33} & 0 & 0 & 0 \end{bmatrix}$
Cubic $\bar{4}3m$	Tetragonal $\bar{4}2m$	Trigonal $3m$
(e.g., GaAs, CdTe, InAs)	(e.g., KDP, ADP)	(e.g., BBO, $LiNbO_3$, $LiTaO_3$)

The tensors d_{ijk} and $\chi^{(3)}_{ijkl}$ are closely related to the Pockels and Kerr tensors r_{ijk} and s_{ijkl}, respectively, as demonstrated in Prob. 22.6-2. They also have the same symmetries, as can be seen by comparing Tables 22.6-1 and 22.6-2, which list d_{iJ} and $\chi^{(3)}_{IK}$, with Tables 21.2-2 and 21.2-3, which provide r_{Ik} and s_{IK}, for a number of crystal groups. Note, however, that d_{iJ} is analogous to the transpose of r_{Ik}.

Table 22.6-2 Third-order nonlinear coefficients $\chi^{(3)}_{IK}$ for an isotropic medium.

$$\begin{bmatrix} \chi^{(3)}_{11} & \chi^{(3)}_{12} & \chi^{(3)}_{12} & 0 & 0 & 0 \\ \chi^{(3)}_{12} & \chi^{(3)}_{11} & \chi^{(3)}_{12} & 0 & 0 & 0 \\ \chi^{(3)}_{12} & \chi^{(3)}_{12} & \chi^{(3)}_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & \chi^{(3)}_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & \chi^{(3)}_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \chi^{(3)}_{44} \end{bmatrix}, \quad \chi^{(3)}_{44} = \tfrac{1}{2}\left(\chi^{(3)}_{11} - \chi^{(3)}_{12}\right).$$

Table 22.6-3 Representative magnitudes of second-order nonlinear optical coefficients for selected materials.[a]

Crystal	d_{iJ} (C/V^2)	$\mathrm{d}_{iJ}/\epsilon_o$ (pm/V)[b]
β-BaB_2O_4 (BBO)	$\mathrm{d}_{22} = 2.0 \times 10^{-23}$	2.2
	$\mathrm{d}_{31} = 3.5 \times 10^{-25}$	0.04
LiB_3O_5 (LBO)	$\mathrm{d}_{31} = 5.9 \times 10^{-24}$	0.67
	$\mathrm{d}_{32} = 7.5 \times 10^{-24}$	0.85
	$\mathrm{d}_{33} = 3.5 \times 10^{-25}$	0.04
$LiIO_3$	$\mathrm{d}_{31} = 3.9 \times 10^{-23}$	4.4
	$\mathrm{d}_{33} = 4.1 \times 10^{-23}$	4.6
$LiNbO_3$	$\mathrm{d}_{22} = 1.9 \times 10^{-23}$	2.1
	$\mathrm{d}_{31} = 4.1 \times 10^{-23}$	4.6
	$\mathrm{d}_{33} = 2.2 \times 10^{-22}$	25.2
$KNbO_3$	$\mathrm{d}_{31} = 1.1 \times 10^{-22}$	11.9
	$\mathrm{d}_{32} = 1.2 \times 10^{-22}$	13.7
$KTiOPO_4$ (KTP)	$\mathrm{d}_{31} = 2.0 \times 10^{-23}$	2.2
	$\mathrm{d}_{32} = 3.3 \times 10^{-23}$	3.7
	$\mathrm{d}_{33} = 1.3 \times 10^{-22}$	14.6
KH_2PO_4 (KDP)	$\mathrm{d}_{36} = 3.1 \times 10^{-24}$	0.38
$NH_4H_2PO_4$ (ADP)	$\mathrm{d}_{36} = 4.2 \times 10^{-24}$	0.47
α-SiO_2 (quartz)	$\mathrm{d}_{11} = 2.7 \times 10^{-24}$	0.30
$KBe_2BO_3F_2$ (KBBF)	$\mathrm{d}_{11} = 4.3 \times 10^{-24}$	0.49
GaAs	$\mathrm{d}_{14} = 1.5 \times 10^{-21}$	170.
Te	$\mathrm{d}_{11} = 5.8 \times 10^{-21}$	650.

[a] Most of the coefficients are as reported by D. N. Nikogosyan, *Nonlinear Optical Crystals: A Complete Survey*, Springer-Verlag, 2005. Values are provided at a wavelength $\lambda_o = 1.06$ μm except for Te, which is provided at $\lambda_o = 10.6$ μm.

[b] The coefficients d/ϵ_o, specified in units of pm/V, are often used in practice. The nonlinear optical coefficients in C/V^2 (MKS units) are readily converted to pm/V by dividing d by $10^{-12}\epsilon_o \approx 8.85 \times 10^{-24}$.

Three-Wave Mixing in Anisotropic Second-Order Nonlinear Media

When an optical field comprising two monochromatic linearly polarized waves of angular frequencies ω_1 and ω_2, and complex amplitudes $\mathbf{E}(\omega_1)$ and $\mathbf{E}(\omega_2)$, travel through a second-order nonlinear crystal, the induced nonlinear polarization-density vector $\mathbf{P}(\omega_3)$ at frequency $\omega_3 = \omega_1 + \omega_2$ has components

$$P_i(\omega_3) = 2\sum_{jk} \mathrm{d}_{ijk} E_j(\omega_1) E_k(\omega_2), \qquad i, j, k = 1, 2, 3, \tag{22.6-2}$$

where $E_j(\omega_1)$, $E_k(\omega_2)$, and $P_i(\omega_3)$ are the components of these vectors along the principal axes of the crystal. This equation is a generalization of (22.2-13d).

Using the contracted notation $(j,k) = J$, (22.6-2) may be conveniently written in matrix form as:

$$\begin{bmatrix} P_1(\omega_3) \\ P_2(\omega_3) \\ P_3(\omega_3) \end{bmatrix} = 2 \begin{bmatrix} \mathrm{d}_{11} & \mathrm{d}_{12} & \cdots & \mathrm{d}_{16} \\ \mathrm{d}_{21} & \mathrm{d}_{22} & \cdots & \mathrm{d}_{26} \\ \mathrm{d}_{31} & \mathrm{d}_{32} & \cdots & \mathrm{d}_{36} \end{bmatrix} \begin{bmatrix} E_1(\omega_1)E_1(\omega_2) \\ E_2(\omega_1)E_2(\omega_2) \\ E_3(\omega_1)E_3(\omega_2) \\ E_2(\omega_1)E_3(\omega_2) + E_3(\omega_1)E_2(\omega_2) \\ E_3(\omega_1)E_1(\omega_2) + E_1(\omega_1)E_3(\omega_2) \\ E_1(\omega_1)E_2(\omega_2) + E_2(\omega_1)E_1(\omega_2) \end{bmatrix}. \tag{22.6-3}$$

Effective value of d. If $E_j(\omega_1) = E(\omega_1)\cos\theta_{1j}$ and $E_k(\omega_2) = E(\omega_2)\cos\theta_{2k}$, where θ_{1j} and θ_{2k} are the angles that the vectors $\mathbf{E}(\omega_1)$ and $\mathbf{E}(\omega_2)$ make with the principal axes, then (22.6-2) may be written in the form

$$P_i(\omega_3) = 2\left[\sum_{jk} \mathrm{d}_{ijk}\cos\theta_{1j}\cos\theta_{2k}\right] E(\omega_1)E(\omega_2). \tag{22.6-4}$$

Since the polarization-density vector $\mathbf{P}(\omega_3)$ is the source for wave 3, only the component $\mathbf{P}_\perp(\omega_3)$ in the plane orthogonal to the wavevector $\mathbf{k}_3$ contributes; the component parallel to $\mathbf{k}_3$ cannot radiate a TEM wave. If $\mathbf{P}_\perp(\omega_3)$ makes angles $\overline{\theta}_{3i}$ with the principal axes, then its magnitude is

$$P_\perp(\omega_3) = \sum_i P_i(\omega_3)\cos\overline{\theta}_{3i}. \tag{22.6-5}$$

It follows from (22.6-4) and (22.6-5) that

$$P_\perp(\omega_3) = 2\mathrm{d}_{\mathrm{eff}}E(\omega_1)E(\omega_2), \tag{22.6-6}$$

where the effective second-order nonlinear optical coefficient is

$$\mathrm{d}_{\mathrm{eff}} = \sum_{ijk} \mathrm{d}_{ijk}\cos\overline{\theta}_{3i}\cos\theta_{1j}\cos\theta_{2k}. \tag{22.6-7}$$

Equation (22.6-6) takes the same form as that used in the scalar formulation provided in Secs. 22.2C and 22.4; the effective second-order nonlinear coefficient $\mathrm{d}_{\mathrm{eff}}$ plays the role of the nonlinear-optics coefficient d. Example 22.6-1 illustrates the computation of $\mathrm{d}_{\mathrm{eff}}$ for a three-wave mixing configuration in an anisotropic crystal.

EXAMPLE 22.6-1. ***Collinear Type-I Three-Wave Mixing in a KDP Crystal.*** In this example, we determine the effective nonlinear optical coefficient $\mathrm{d}_{\mathrm{eff}}$ for three collinear waves traveling in a KDP crystal at an arbitrary direction (θ,ϕ) defined in a spherical coordinate system with the crystal optic axis pointing in the z direction, as illustrated in Fig. 22.6-1. Waves 1 and 2 are ordinary waves at frequencies ω_1 and ω_2, and wave 3 is extraordinary with frequency $\omega_3 = \omega_1 + \omega_2$.

Using (22.6-2) and Table 22.6-1 for crystals of $\bar{4}2m$ symmetry, such as KDP, the nonlinear components of the polarization-density vector are given by

$$\begin{aligned} P_1(\omega_3) &= 2\mathrm{d}_{14}\left[E_2(\omega_1)E_3(\omega_2) + E_3(\omega_1)E_2(\omega_2)\right] \\ P_2(\omega_3) &= 2\mathrm{d}_{14}\left[E_3(\omega_1)E_1(\omega_2) + E_1(\omega_1)E_3(\omega_2)\right] \\ P_3(\omega_3) &= 2\mathrm{d}_{36}\left[E_1(\omega_1)E_2(\omega_2) + E_2(\omega_1)E_1(\omega_2)\right]. \end{aligned} \tag{22.6-8}$$

In this geometry, the electric field components of waves 1 and 2 are:

$$E_1(\omega_1) = E(\omega_1)\sin\phi, \quad E_2(\omega_1) = -E(\omega_1)\cos\phi, \quad E_3(\omega_1) = 0,$$
$$E_1(\omega_2) = E(\omega_2)\sin\phi, \quad E_2(\omega_2) = -E(\omega_2)\cos\phi, \quad E_3(\omega_2) = 0. \tag{22.6-9}$$

Therefore, based on (22.6-8), the components of the polarization-density vector for wave 3 are

$$P_1(\omega_3) = 0, \quad P_2(\omega_3) = 0, \quad P_3(\omega_3) = -4\mathrm{d}_{36}\sin\phi\cos\phi\, E(\omega_1)E(\omega_2). \tag{22.6-10}$$

In this case, the component $P_\perp(\omega_3) = -P_3(\omega_3)\sin\theta$, so that

$$\boxed{\mathrm{d}_{\mathrm{eff}} = -\mathrm{d}_{36}\sin\theta\sin 2\phi.} \tag{22.6-11}$$

This result can also be obtained by direct use of (22.6-7) with the appropriate angles and coefficients.

The effective nonlinear optical coefficient in (22.6-11) has its maximum magnitude d_{36} if the angles are $\theta = 90°$ and $\phi = 45°$, as illustrated in Fig. 22.6-1.

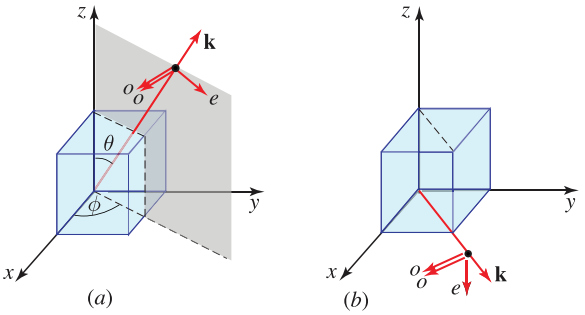


Figure 22.6-1 (a) Geometry for collinear Type-I o-o-e three-wave mixing in a uniaxial crystal whose optic axis is in the z direction. (b) Direction of propagation for achieving maximum $\mathrm{d}_{\mathrm{eff}}$.

*22.7 DISPERSIVE NONLINEAR MEDIA

This section provides a brief discussion of the origin of dispersion and its effect on nonlinear optical processes. Anisotropic dispersive media are briefly considered at the end of this section. A dispersive medium is a medium with memory (see Sec. 5.2); the polarization density $\mathcal{P}(t)$ resulting from an applied electric field $\mathcal{E}(t)$ does not appear instantaneously. Rather, the response $\mathcal{P}(t)$ at time t is a function of the applied electric field $\mathcal{E}(t')$ at times $t' \le t$. When the medium is also nonlinear, the functional relation between $\mathcal{P}(t)$ and $\{\mathcal{E}(t),\ t' \le t\}$ is nonlinear. There are two means for describing such nonlinear dynamical systems:

1. A phenomenological integral relation between $\mathcal{P}(t)$ and $\mathcal{E}(t)$ based on a Volterra-series expansion, which is similar to a Taylor-series expansion. The coefficients of the expansion characterize the medium phenomenologically.
2. A nonlinear differential equation for $\mathcal{P}(t)$, with $\mathcal{E}(t)$ as a driving force, obtained by developing a model for the physics of the polarization process, much as the Lorentz model was developed for linear media.

Integral-Transform Description of Dispersive Nonlinear Media

If the deviation from linearity is small, a Volterra-series expansion may be used to describe the relation between $\mathcal{P}(t)$ and $\mathcal{E}(t)$. The first term of the expansion is a linear

combination of $\mathcal{E}(t')$ for all $t' \leq t$,

$$\mathcal{P}(t) = \epsilon_o \int_{-\infty}^{\infty} \mathrm{x}(t - t')\mathcal{E}(t')\, dt'. \tag{22.7-1}$$

This describes a linear system with impulse response function $\epsilon_o \mathrm{x}(t)$ [see Sec. 5.2, particularly (5.2-23), and Appendix B].

The second term in the expansion is a superposition of the products $\mathcal{E}(t')\mathcal{E}(t'')$ at *pairs* of times $t' \leq t$ and $t'' \leq t$,

$$\mathcal{P}(t) = \epsilon_o \int\int_{-\infty}^{\infty} \mathrm{x}^{(2)}(t - t', t - t'')\, \mathcal{E}(t')\mathcal{E}(t'')\, dt'\, dt'', \tag{22.7-2}$$

where $\mathrm{x}^{(2)}(t', t'')$ is a function of two variables that characterizes the second-order dispersive nonlinearity. The third term represents a third-order nonlinearity that can be characterized by a function $\mathrm{x}^{(3)}(t', t'', t''')$ together with a similar triple-integral relation.

The *linear* dispersive contribution described by (22.7-1) can also be completely characterized by the response to monochromatic fields. If $\mathcal{E}(t) = \mathrm{Re}\{E(\omega)\exp(j\omega t)\}$, then $\mathcal{P}(t) = \mathrm{Re}\{P(\omega)\exp(j\omega t)\}$, where $P(\omega) = \epsilon_o \chi(\omega) E(\omega)$ and $\chi(\omega)$ is the Fourier transform of $\mathrm{x}(t)$ at $\nu = \omega/2\pi$. The medium is then characterized completely by the frequency-dependent susceptibility $\chi(\omega)$.

The *second-order nonlinear* contribution described by (22.7-2) is characterized by the response to a superposition of *two* monochromatic waves of angular frequencies ω_1 and ω_2. Substituting

$$\mathcal{E}(t) = \mathrm{Re}\{E(\omega_1)\exp(j\omega_1 t) + E(\omega_2)\exp(j\omega_2 t)\} \tag{22.7-3}$$

into (22.7-2), it can be shown that the polarization-density component of angular frequency $\omega_3 = \omega_1 + \omega_2$ has an amplitude

$$\boxed{P(\omega_3) = 2\mathrm{d}(\omega_3; \omega_1, \omega_2)\, E(\omega_1)E(\omega_2).} \tag{22.7-4}$$

The coefficient $\mathrm{d}(\omega_3; \omega_1, \omega_2)$ is a frequency-dependent version of the nonlinear optical coefficient d in (22.2-13d). The relation between this coefficient and the response function $\mathrm{x}^{(2)}(t', t'')$ is established by defining

$$\mathcal{X}^{(2)}(\omega_1, \omega_2) = \int\int_{-\infty}^{\infty} \mathrm{x}^{(2)}(t', t'') \exp[-j(\omega_1 t' + \omega_2 t'')]\, dt'\, dt'', \tag{22.7-5}$$

which is the two-dimensional Fourier transform of $\mathrm{x}^{(2)}(t', t'')$ evaluated at $\nu_1 = -\omega_1/2\pi$ and $\nu_2 = -\omega_2/2\pi$ [see (A.3-2) in Appendix A]. Substituting (22.7-3) into (22.7-2) and using (22.7-5), we obtain

$$\mathrm{d}(\omega_3; \omega_1, \omega_2) = \epsilon_o \mathcal{X}^{(2)}(\omega_1, \omega_2). \tag{22.7-6a}$$

Thus, the second-order nonlinear dispersive medium is completely characterized by either of the frequency-dependent functions, $\mathcal{X}^{(2)}(\omega_1, \omega_2)$ or $\mathrm{d}(\omega_3; \omega_1, \omega_2)$.

The degenerate case of second-harmonic generation in a second-order nonlinear medium is also readily described by substituting $\mathcal{E}(t) = \text{Re}\{E(\omega)\exp(j\omega t)\}$ into (22.7-2) and using (22.7-5). The resultant polarization density has a component at frequency 2ω with amplitude $P(2\omega) = \text{d}(2\omega;\omega,\omega)\,E(\omega)E(\omega)$, where

$$\text{d}(2\omega;\omega,\omega) = \tfrac{1}{2}\epsilon_o \mathcal{X}^{(2)}(\omega,\omega). \tag{22.7-6b}$$

Other d coefficients representing various wave-mixing processes may similarly be related to the two-dimensional function $\mathcal{X}^{(2)}(\omega_1,\omega_2)$. The electro-optic effect, for example, is a result of interaction between a steady electric field ($\omega_1 = 0$) and an optical wave ($\omega_2 = \omega$) to generate a polarization density at $\omega_3 = \omega$. The pertinent coefficient for this interaction is $\text{d}(\omega;0,\omega) = 2\epsilon_o \mathcal{X}^{(2)}(\omega,0)$; it determines the Pockels coefficient r in accordance with (22.2-11).

In a *third-order nonlinear* medium, an electric field comprising three harmonic functions of angular frequencies ω_1, ω_2, and ω_3 creates a sum-frequency polarization density with a component at angular frequency $\omega_4 = \omega_1 + \omega_2 + \omega_3$ of amplitude

$$\boxed{P(\omega_4) = 6\chi^{(3)}(\omega_4;\omega_1,\omega_2,\omega_3)\,E(\omega_1)E(\omega_2)E(\omega_3),} \tag{22.7-7}$$

where the function $\chi^{(3)}(\omega_4;\omega_1,\omega_2,\omega_3)$ replaces the nonlinear optical coefficient $\chi^{(3)}$ that describes the nondispersive case. The function $\chi^{(3)}(\omega_4;\omega_1,\omega_2,\omega_3)$ can be determined from $\text{x}^{(3)}(t',t'',t''')$ by relations similar to (22.7-6a).

In short, as a consequence of dispersion, the second- and third-order coefficients d and $\chi^{(3)}$ are dependent on the frequencies of the waves involved in the wave-mixing process.

Differential-Equation Description of Dispersive Nonlinear Media

An example of a nonlinear dynamic relation between $\mathcal{P}(t)$ and $\mathcal{E}(t)$ is provided by the differential equation

$$\frac{d^2\mathcal{P}}{dt^2} + \zeta\frac{d\mathcal{P}}{dt} + \omega_0^2\mathcal{P} + \omega_0^2\epsilon_o\chi_0 b\,\mathcal{P}^2 = \omega_0^2\epsilon_o\chi_0\mathcal{E}, \tag{22.7-8}$$

where ζ, ω_0, χ_0, and b are constants. In the absence of the nonlinear term, $\omega_0^2\epsilon_o\chi_0 b\,\mathcal{P}^2$, (22.7-8) reduces to (5.5-15), which is appropriate for a linear resonant dielectric medium described by the Lorentz oscillator model (see Sec. 5.5C). Each atom is then characterized by a harmonic oscillator in which an electron of mass m is subjected to an electric-field force $-e\mathcal{E}$, an elastic restoring force $-\kappa x$, and a frictional force $m\zeta\,dx/dt$, where x is the displacement of the electron from its equilibrium position and $\omega_0 = \sqrt{\kappa/m}$ is the resonance angular frequency. The medium is then linear and dispersive with a susceptibility given by [see (5.5-18)]

$$\boxed{\chi(\omega) = \chi_0\frac{\omega_0^2}{\omega_0^2 - \omega^2 + j\omega\zeta}.} \tag{22.7-9}$$

Linear Susceptibility (Harmonic-Oscillator)

When the restoring force is a nonlinear function of displacement, $-\kappa x - \kappa_2 x^2$, where κ and κ_2 are constants, the result is an anharmonic oscillator described by (22.7-8), where b is proportional to κ_2. The medium is then nonlinear.

EXERCISE 22.7-1

Polarization Density for an Anharmonic-Oscillator Medium. Show that for a medium containing N atoms per unit volume, each modeled as an anharmonic (nonlinear) oscillator with restraining force $-\kappa x - \kappa_2 x^2$, the relation between $\mathcal{P}(t)$ and $\mathcal{E}(t)$ is the nonlinear differential equation (22.7-8), where $\chi_0 = Ne^2/\epsilon_o m\omega_0^2$ and $b = \kappa_2/e^3N^2$.

Equation (22.7-8) cannot be solved exactly. However, if the nonlinear term is small, an iterative approach provides an approximate solution. Let (22.7-8) be written in the form

$$\mathcal{L}\{\mathcal{P}\} = \mathcal{E} - b\mathcal{P}^2, \tag{22.7-10}$$

where $\mathcal{L} = (\omega_0^2\epsilon_o\chi_0)^{-1}(d^2/dt^2 + \zeta d/dt + \omega_0^2)$ is a linear differential operator. The iterative solution of (22.7-10) is carried out via the following steps:

1. Find a first-order approximation $\mathcal{P}_1$ by neglecting the nonlinear term $b\mathcal{P}^2$ in (22.7-10), and solving the *linear* equation

$$\mathcal{L}\{\mathcal{P}_1\} \approx \mathcal{E}. \tag{22.7-11}$$

2. Use this approximate solution to determine the small nonlinear term $b\mathcal{P}_1^2$.
3. Obtain a second-order approximation by solving (22.7-10) with the term $b\mathcal{P}^2$ replaced by $b\mathcal{P}_1^2$. The solution of the resulting *linear* equation is denoted $\mathcal{P}_2$,

$$\mathcal{L}\{\mathcal{P}_2\} = \mathcal{E} - b\mathcal{P}_1^2. \tag{22.7-12}$$

4. Repeat the process to obtain a third-order approximation as illustrated by the block diagram of Fig. 22.7-1.

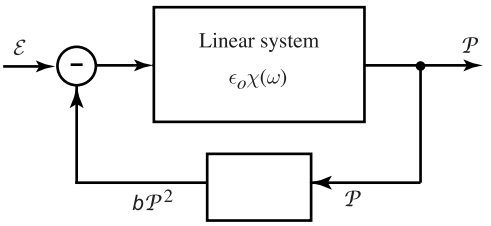


Figure 22.7-1 Block diagram representing the nonlinear differential equation (22.7-10). The linear system represented by the operator equation $\mathcal{L}\{\mathcal{P}\} = \mathcal{E}$ has a transfer function $\epsilon_o\chi(\omega)$.

We first examine the special case of monochromatic light, $\mathcal{E} = \text{Re}\{E(\omega)\exp(j\omega t)\}$. In the first iteration $\mathcal{P}_1 = \text{Re}\{P_1(\omega)\exp(j\omega t)\}$, where $P_1(\omega) = \epsilon_o\chi(\omega)E(\omega)$ and $\chi(\omega)$ is given by (22.7-9). In the second iteration, the linear system is driven by a force

$$\begin{aligned}\mathcal{E} - b\mathcal{P}_1^2 &= \text{Re}\{E(\omega)e^{j\omega t}\} - b\big[\text{Re}\{\epsilon_o\chi(\omega)E(\omega)e^{j\omega t}\}\big]^2 \\ &= \text{Re}\{E(\omega)e^{j\omega t}\} - \tfrac{1}{2}b\,\text{Re}\{[\epsilon_o\chi(\omega)E(\omega)]^2\,e^{j2\omega t}\} - \tfrac{1}{2}b\,|\epsilon_o\chi(\omega)E(\omega)\,|^2.\end{aligned}$$

Since these three terms have frequencies ω, 2ω, and 0, the linear system responds with susceptibilities $\chi(\omega)$, $\chi(2\omega)$, and $\chi(0)$, respectively. The component of $\mathcal{P}_2$ at

frequency 2ω has an amplitude $P_2(2\omega) = \epsilon_o\chi(2\omega)\{-\frac{1}{2}b\,[\epsilon_o\chi(\omega)E(\omega)]^2\}$. Since $P(2\omega) = \mathrm{d}(2\omega;\omega,\omega)E(\omega)E(\omega)$, we conclude that

$$\boxed{\mathrm{d}(2\omega;\omega,\omega) = -\tfrac{1}{2}b\,\epsilon_o^3[\chi(\omega)]^2\chi(2\omega).} \tag{22.7-13}$$

EXERCISE 22.7-2

Miller's Rule. For the nonlinear resonant medium described by (22.7-8), if the light comprises a superposition of two monochromatic waves of angular frequencies ω_1 and ω_2, show that the second-order approximation described by (22.7-11) and (22.7-12) yields a component of polarization density at frequency $\omega_3 = \omega_1 + \omega_2$ with amplitude $P_2(\omega_3) = 2\mathrm{d}(\omega_3;\omega_1,\omega_2)E(\omega_1)E(\omega_2)$, where

$$\boxed{\mathrm{d}(\omega_3;\omega_1,\omega_2) = -\tfrac{1}{2}b\,\epsilon_o^3\chi(\omega_1)\chi(\omega_2)\chi(\omega_3).} \tag{22.7-14}$$

Miller's Rule

Equation (22.7-14) is known as *Miller's rule*.

Miller's rule states that the coefficient of second-order nonlinearity for the generation of a wave of frequency $\omega_3 = \omega_1 + \omega_2$, from two waves of frequencies ω_1 and ω_2, is proportional to the product of the linear susceptibilities at the three frequencies, $\chi(\omega_1)\chi(\omega_2)\chi(\omega_3)$. The three frequencies must therefore lie within the optical transmission window of the medium (away from resonance). If these frequencies are much smaller than the resonance frequency ω_0, then (22.7-9) gives $\chi(\omega) = \chi_0$, and (22.7-14) then yields $\mathrm{d}(\omega_3;\omega_1,\omega_2) = -\frac{1}{2}b\,\epsilon_o^2\chi_0^3$, which is independent of frequency. The medium is then approximately nondispersive, and the results of the previous sections in which dispersion was neglected are applicable. Miller's rule also indicates that materials with large refractive indices (large χ_0) tend to have large d.

Anisotropic Dispersive Media

When both anisotropic and dispersive properties are considered, three-wave mixing in a second-order medium is described by the more general relation

$$\boxed{P_i(\omega_3) = 2\sum_{jk}\mathrm{d}_{ijk}(\omega_3;\omega_1,\omega_2)\,E_j(\omega_1)E_k(\omega_2),} \tag{22.7-15}$$

where $\omega_3 = \omega_1 + \omega_2$. The coefficients d_{ijk} are now dependent on the frequencies of the mixed waves. This relation is similar to the relation $P_i(\omega) = \sum_j \chi_{ij}(\omega)E_j(\omega)$, which describes linear media. Similarly, four-wave mixing in a third-order medium is described by

$$\boxed{P_i(\omega_4) = 6\sum_{jkl}\chi^{(3)}_{ijkl}(\omega_4;\omega_1,\omega_2,\omega_3)\,E_j(\omega_1)E_k(\omega_2)E_l(\omega_3),} \tag{22.7-16}$$

where $\omega_4 = \omega_1 + \omega_2 + \omega_3$.

The frequency-dependent tensor components d_{ijk}, and $\chi^{(3)}_{ijkl}$ obey several intrinsic symmetry relations that are similar to the relation $\chi^*_{ij}(\omega) = \chi_{ij}(-\omega)$ in linear optics:

$$\begin{aligned} \mathrm{d}^*_{ijk}(\omega_3;\omega_1,\omega_2) &= \mathrm{d}_{jki}(\omega_1;-\omega_2,\omega_3) \\ &= \mathrm{d}_{kij}(\omega_2;\omega_3,-\omega_1) \end{aligned} \tag{22.7-17}$$

$$\begin{aligned} \chi^{(3)}_{ijkl}{}^*(\omega_4;\omega_1,\omega_2,\omega_3) &= \chi^{(3)}_{jkli}(\omega_1;-\omega_2,-\omega_3,\omega_4) \\ &= \cdots \\ &= \chi^{(3)}_{jikl}(\omega_3;\omega_4,-\omega_1,-\omega_2). \end{aligned} \tag{22.7-18}$$

In these relations, the coefficient $\mathrm{d}_{jki}(\omega_1;-\omega_2,\omega_3)$, for example, represents a down-conversion process in which a wave of frequency ω_2 and polarization k mixes with a wave of frequency ω_3 and polarization i and generates a wave of frequency $\omega_1 = \omega_3 - \omega_2$ and polarization j. Other coefficients can be similarly interpreted. This type of intrinsic symmetry is of course supplemented by other structural symmetry relations that are obeyed for various classes of crystals.

READING LIST

Nonlinear Optics

See also the reading lists in Chapters 5, 6, 8, 21, and 23.

G. S. He and S. H. Liu, *Advanced Nonlinear Optics*, World Scientific, 2nd ed. 2018.

Y. Zhu, Z. Wang, Y. Chen, Y. Lu, and S. Zhu, *Superlattices and Microstructures of Dielectric Materials: Quasi-Phase-Matching in Nonlinear Optics and Quantum Optics*, De Gruyter, 2018.

A. Morita, *Theory of Sum Frequency Generation Spectroscopy*, Springer-Verlag, 2018.

G. S. He, *Nonlinear Optics and Photonics*, Oxford University Press, 2015.

L. Lugiato, F. Prati, and M. Brambilla, *Nonlinear Optical Systems*, Cambridge University Press, 2015.

P. D. Drummond and M. Hillery, *The Quantum Theory of Nonlinear Optics*, Cambridge University Press, 2014.

S. Guha and L. P. Gonzalez, *Laser Beam Propagation in Nonlinear Optical Media*, CRC Press/Taylor & Francis, 2014.

G. P. Agrawal, *Nonlinear Fiber Optics*, Academic Press/Elsevier, 5th ed. 2013.

P. E. Powers, *Field Guide to Nonlinear Optics*, SPIE Optical Engineering Press, 2013.

G. I. Stegeman and R. A. Stegeman, *Nonlinear Optics: Phenomena, Materials and Devices*, Wiley, 2012.

P. E. Powers, *Fundamentals of Nonlinear Optics*, CRC Press/Taylor & Francis, 2011.

G. New, *Introduction to Nonlinear Optics*, Cambridge University Press, 2011.

P. Mandel, *Nonlinear Optics*, Wiley–VCH, 2010.

M. E. Marhic, *Fiber Optical Parametric Amplifiers, Oscillators and Related Devices*, Cambridge University Press, 2008.

R. W. Boyd, *Nonlinear Optics*, Academic Press/Elsevier, 3rd ed. 2008.

R. Menzel, *Photonics: Linear and Nonlinear Interactions of Laser Light and Matter*, Springer-Verlag, 2nd ed. 2007.

M. Wegener, *Extreme Nonlinear Optics: An Introduction*, Springer-Verlag, 2005.

D. N. Nikogosyan, *Nonlinear Optical Crystals: A Complete Survey*, Springer-Verlag, 2005.

P. P. Banerjee, *Nonlinear Optics: Theory, Numerical Modeling, and Applications*, Marcel Dekker, 2004.

A. Brignon and J.-P. Huignard, eds., *Phase Conjugate Laser Optics*, Wiley, 2004.

T. Suhara and M. Fujimura, *Waveguide Nonlinear-Optic Devices*, Springer-Verlag, 2003.

F. Kajzar and R. Reinisch, eds., *Beam Shaping and Control with Nonlinear Optics*, Plenum Press, 1998.

N. Bloembergen, *Nonlinear Optics*, World Scientific, 1965, 4th ed., 1996.

C. L. Tang and L. K. Cheng, *Fundamentals of Optical Parametric Processes and Oscillators*, Harwood, 1995.

J.-Y. Zhang, J. Y. Huang, and Y. R. Shen, *Optical Parametric Generation and Amplification*, Harwood, 1995.

J. Zyss, ed., *Molecular Nonlinear Optics: Materials, Physics, and Devices*, Academic Press, 1994.

J.-I. Sakai, *Phase-Conjugate Optics*, McGraw–Hill, 1992.

P. N. Butcher and D. Cotter, *The Elements of Nonlinear Optics*, Cambridge University Press, 1990.

V. S. Butylkin, A. E. Kaplan, Yu. G. Khronopulo, and E. I. Yakubovich, *Resonant Nonlinear Interactions of Light with Matter*, Springer-Verlag, 1989.

M. Schubert and B. Wilhelmi, *Nonlinear Optics and Quantum Electronics*, Wiley, 1986.

B. Ya. Zel'dovich, N. F. Pilipetsky, and V. V. Shkunov, *Principles of Phase Conjugation*, Springer-Verlag, 1985.

Y. R. Shen, *The Principles of Nonlinear Optics*, Wiley, 1984.

R. A. Fisher, ed., *Optical Phase Conjugation*, Academic Press, 1983.

Seminal Articles and Reprint Collections

T. R. Gosnell, ed., *Selected Papers on Upconversion Lasers*, SPIE Optical Engineering Press (Milestone Series Volume 161), 2000.

J. H. Hunt, ed., *Selected Papers on Optical Parametric Oscillators and Amplifiers and Their Applications*, SPIE Optical Engineering Press (Milestone Series Volume 140), 1997.

J. Peřina, ed., *Selected Papers on Photon Statistics and Coherence in Nonlinear Optics*, SPIE Optical Engineering Press (Milestone Series Volume 39), 1991.

H. E. Brandt, ed., *Selected Papers on Nonlinear Optics*, SPIE Optical Engineering Press (Milestone Series Volume 32), 1991.

N. Bloembergen, Nonlinear Optics and Spectroscopy (Nobel Lecture in Physics, 1981), *Reviews of Modern Physics*, vol. 54, pp. 685–695, 1982.

G. H. C. New and J. F. Ward, Optical Third-Harmonic Generation in Gases, *Physical Review Letters*, vol. 19, pp. 556–559, 1967.

J. A. Armstrong, N. Bloembergen, J. Ducuing, and P. S. Pershan, Interactions Between Light Waves in a Nonlinear Dielectric, *Physical Review*, vol. 127, pp. 1918–1939, 1962.

W. H. Louisell, A. Yariv, and A. E. Siegman, Quantum Fluctuations and Noise in Parametric Processes. I., *Physical Review*, vol. 124, pp. 1646–1654, 1961.

P. A. Franken, A. E. Hill, C. W. Peters, and G. Weinreich, Generation of Optical Harmonics, *Physical Review Letters*, vol. 7, pp. 118–119, 1961.

PROBLEMS

22.2-2 **Power Exchange in Frequency Up-Conversion.** A $LiNbO_3$ crystal of refractive index $n = 2.2$ is used to convert light of free-space wavelength 1.3 μm to light of free-space wavelength 0.5 μm, using a three-wave mixing process. The three waves are collinear plane waves traveling in the z direction. Determine the wavelength of the third wave (the pump). If the power of the 1.3-μm wave decreases by 1 mW within an incremental distance Δz, what is the power gain of the up-converted wave and the power loss or gain of the pump within the same distance?

22.2-3 **Matching Conditions for Collinear Type-II SHG.** Determine the angle θ for a KDP crystal used in Type-II second-harmonic generation at $\lambda = 1.06$ μm for each of the o-e-o and o-e-e configurations. Use the Sellmeier equations in Table 5.5-1 to determine the wavelength dependence of the refractive indices.

22.2-4 **Phase Matching in a Degenerate Parametric Downconverter.** A degenerate parametric downconverter uses a KDP crystal to downconvert light from 0.6 μm to 1.2 μm. If the two

waves are collinear, what is the proper direction of propagation of the waves (in relation to the optic axis of the crystal), and their polarizations, so that the phase-matching condition is satisfied? KDP is a uniaxial crystal with the following refractive indices: at $\lambda_o = 0.6$ μm, $n_o = 1.509$ and $n_e = 1.468$; at $\lambda_o = 1.2$ μm, $n_o = 1.490$ and $n_e = 1.459$.

22.2-5 **Matching Conditions for Three-Wave Mixing in a Dispersive Medium.** The refractive index of a nonlinear medium is a function of wavelength approximated by $n(\lambda_o) \approx n_0 - \xi\lambda_o$, where λ_o is the free-space wavelength and n_0 and ξ are constants. Show that three waves of wavelengths λ_{o1}, λ_{o2}, and λ_{o3} traveling in the same direction cannot be efficiently coupled by a second-order nonlinear effect. Is efficient coupling possible if one of the waves travels in the opposite direction?

*22.2-6 **Tolerance to Phase Mismatching.**

(a) The Helmholtz equation with a source, $\nabla^2 E + k^2 E = -S$, has the solution [see (5.6-4)]

$$E(\mathbf{r}) = \int_V S(\mathbf{r}') \frac{\exp(-jk_o|\mathbf{r} - \mathbf{r}'|)}{4\pi|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}',$$

where V is the volume of the source and $k_o = 2\pi/\lambda_o$. This equation can be used to determine the field emitted at a point $\mathbf{r}$, given the source at all points $\mathbf{r}'$ within the source volume. If the source is confined to a small region centered about the origin $\mathbf{r} = \mathbf{0}$, and $\mathbf{r}$ is a point sufficiently far from the source so that $r' \ll r$ for all $\mathbf{r}'$ within the source, then $|\mathbf{r} - \mathbf{r}'| = (r^2 + r'^2 - 2\mathbf{r}\cdot\mathbf{r}')^{1/2} \approx r(1 - \mathbf{r}\cdot\mathbf{r}'/r^2)$ and

$$E(\mathbf{r}) \approx \frac{\exp(-jk_o r)}{4\pi r} \int_V S(\mathbf{r}')\, \exp(jk_o \hat{\mathbf{r}}\cdot\mathbf{r}')\, d\mathbf{r}',$$

where $\hat{\mathbf{r}}$ is a unit vector in the direction of $\mathbf{r}$. Assuming that the volume V is a cube of side L and the source is a harmonic function $S(\mathbf{r}) = \exp(-j\mathbf{k}_s \cdot \mathbf{r})$, show that if $L \gg \lambda_0$, the emitted light is maximized when $k_o\hat{\mathbf{r}} = \mathbf{k}_s$ and decreases sharply when this condition is not met. Thus, a harmonic source of dimensions much greater than a wavelength emits a plane wave with approximately the same wavevector.

(b) Use the relation in (a) and the first Born approximation to determine the scattered field when the field incident on a second-order nonlinear medium is the sum of two waves with wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$. Derive the phase-matching condition $\mathbf{k}_3 = \mathbf{k}_1 + \mathbf{k}_2$ and determine the smallest magnitude of $\Delta\mathbf{k} = \mathbf{k}_3 - \mathbf{k}_1 - \mathbf{k}_2$ at which the scattered field E vanishes.

22.2-7 **Backward SHG with QPM.** Show that a periodically poled crystal may be used to generate a second-harmonic wave traveling in a direction opposite to that of the fundamental wave. Write the phase-matching equation for this quasi-phase-matched process. If the equation is satisfied for the 7th-order harmonic of the periodic function, determine the ratio of the poling period to the wavelength of the fundamental wave in the medium.

22.3-4 **Invariants in Four-Wave Mixing.** Derive equations for energy and photon-number conservation (the Manley–Rowe relations) for four-wave mixing.

22.3-5 **Power of a Spatial Soliton.** Determine an expression for the integrated intensity of the spatial soliton described by (22.3-12) and show that it is inversely proportional to the beam width W_0.

22.3-6 **An Opto-Optic Phase Modulator.** Design a system for modulating the phase of an optical beam of wavelength 546 nm and width $W = 0.1$ mm using a CS_2 Kerr cell of length $L = 10$ cm. The modulator is controlled by light from a pulsed laser of wavelength 694 nm. CS_2 has a refractive index $n = 1.6$ and a coefficient of third-order nonlinearity $\chi^{(3)} = 4.4 \times 10^{-32}$ C·m/V^3. Estimate the optical power P_π of the controlling light required to modulate the phase of the controlled light by π.

22.3-7 **SHG in a Third-Order Nonlinear Medium via a Static Electric Field.** Show that SHG can take place in a third-order nonlinear medium with an applied static electric field. What physical parameters determine the efficiency of this SHG process?

*22.4-7 **Gain of a Parametric Amplifier.** A parametric amplifier uses a 4-cm-long KDP crystal ($n \approx 1.49, \mathrm{d} = 8.3 \times 10^{-24}$ C/V^2) to amplify light of wavelength 550 nm. The pump wavelength is 335 nm and its intensity is 10^6 W/cm^2. Assuming that the signal, idler, and pump waves are collinear, determine the amplifier gain coefficient and the overall gain.

*22.4-8 **Degenerate Parametric Downconverter.** Write and solve the coupled equations that describe wave mixing in a parametric downconverter with a pump at frequency $\omega_3 = 2\omega$ and signals at $\omega_1 = \omega_2 = \omega$. All waves travel in the z direction. Derive an expression for the photon-flux densities at 2ω and ω and the conversion efficiency for an interaction length L. Verify that energy conservation and photon conservation are maintained.

*22.4-9 **Threshold Pump Intensity for Parametric Oscillation.** A parametric oscillator makes use of a 5-cm-long $LiNbO_3$ crystal with a second-order nonlinear optical coefficient $\mathrm{d} = 4 \times 10^{-23}$ C/V^2 and refractive index $n = 2.2$ (assumed to be approximately constant at all frequencies of interest). The pump is obtained from a 1.06-μm Nd:YAG laser that is frequency doubled using a second-harmonic generator. The crystal is placed in a resonator using identical mirrors with reflectances $\mathcal{R} = 0.98$. Phase matching is satisfied when the signal and idler of the parametric amplifier are of equal frequencies. Determine the minimum pump intensity required to achieve parametric oscillation.

*22.5-2 **Combined SHG and SFG.** Two waves of angular frequencies ω_1 and ω_2, along with their second-harmonic counterparts of angular frequencies $2\omega_1$ and $2\omega_2$, and their sum-frequency wave of angular frequency $\omega_1 + \omega_2$, interact simultaneously in a second-order nonlinear medium. Assuming that phase matching is satisfied for the two SHG processes and for the SFG process, write coupled equations for this five-wave-mixing process. Solve these equations numerically and demonstrate that the presence of the second wave may suppress the SHG process for the first.

*22.5-3 **Coupled-Wave Equations for Degenerate Four-Wave Mixing.** Consider the collinear four-wave-mixing problem in a third-order nonlinear medium, in the degenerate case $\omega_4 = \omega_3$ and $\omega_1 + \omega_2 = 2\omega_3$. Derive coupled wave equations for the amplitudes A_1, A_2, and A_3 assuming that the phase-matching condition is fully met.

*22.6-1 **Collinear Type-II Three-Wave Mixing in a BBO Crystal.** Repeat the analysis carried out in Example 22.6-1 to demonstrate that the effective nonlinear coefficient $\mathrm{d}_{\mathrm{eff}}$ for Type-II o-e-e three-wave mixing for a crystal in the $3m$ group, such as BBO, is $\mathrm{d}_{\mathrm{eff}} = \mathrm{d}_{22} \cos^2\theta \cos 3\phi$.

*22.6-2 **Relation Between Nonlinear-Optical and Electro-Optic Coefficients.** Show that the electro-optic coefficients are related to the coefficients of optical nonlinearity by $\mathrm{r}_{ijk} = -4\epsilon_o \mathrm{d}_{ijk}/\epsilon_{ii}\epsilon_{jj}$ and $\mathrm{s}_{ijkl} = -12\epsilon_o \chi^{(3)}_{ijkl}/\epsilon_{ii}\epsilon_{jj}$. These relations are generalizations of (22.2-11) and (22.3-2), respectively. *Hint:* If two matrices $\mathbf{A}$ and $\mathbf{B}$ are related by $\mathbf{B} = \mathbf{A}^{-1}$, the incremental matrices $\Delta\mathbf{A}$ and $\Delta\mathbf{B}$ are related by $\Delta\mathbf{B} = -\mathbf{A}^{-1}\Delta\mathbf{A}\mathbf{A}^{-1}$.

CHAPTER

23

ULTRAFAST OPTICS

23.1 PULSE CHARACTERISTICS 1079
- A. Temporal and Spectral Characteristics
- B. Gaussian and Chirped Gaussian Pulses
- C. Spatial Characteristics

23.2 PULSE SHAPING AND COMPRESSION 1088
- A. Chirp Filters
- B. Implementations of Chirp Filters
- C. Pulse Compression
- D. Pulse Shaping

23.3 PULSE PROPAGATION IN OPTICAL FIBERS 1102
- A. The Optical Fiber as a Chirp Filter
- B. Propagation of a Gaussian Pulse in an Optical Fiber
- *C. Slowly Varying Envelope Diffusion Equation
- *D. Analogy Between Dispersion and Diffraction

23.4 ULTRAFAST LINEAR OPTICS 1115
- A. Ray Optics
- *B. Wave and Fourier Optics
- *C. Beam Optics

23.5 ULTRAFAST NONLINEAR OPTICS 1126
- A. Pulsed Parametric Processes
- B. Optical Solitons
- *C. Supercontinuum Light
- *D. High-Harmonic Generation and Attosecond Optics

23.6 PULSE DETECTION 1146
- A. Measurement of Intensity
- B. Measurement of Spectral Intensity
- C. Measurement of Phase
- *D. Measurement of Spectrogram

Paul B. Corkum (born 1943) (left) was the principal progenitor of the field of attosecond optics; **James P. Gordon (1928–2013)** greatly advanced our understanding of optical solitons; and **Gérard Mourou (born 1944)** (right) implemented optical chirped-pulse amplification with Donna Strickland, for which they shared the Nobel Prize with Arthur Ashkin in 2018.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

The study of optical pulses began in earnest with the invention of the laser because the earliest lasers could only emit light in the form of pulses; the development of CW lasers required significant additional effort. Interest in the characteristics and behavior of ultrashort optical pulses flourished as progress was made in generating optical pulses of shorter and shorter duration, with concomitantly larger and larger peak intensity. The generation of nanosecond optical pulses was followed by the generation of picosecond pulses, then femtosecond pulses, and finally attosecond pulses. Progress in generating ultrashort and ultrahigh-intensity (or ultrahigh-field-strength) optical pulses was fueled by the many important applications spawned by their availability. These include nonlinear frequency conversion; the probing of ultrafast physical, chemical, and biological processes; the generation of spatially coherent X-ray beams; multiphoton imaging; materials processing; and ultrafast optical fiber communications.

In the context of optics, the terms *ultrafast* and *ultrashort* are generally used to describe pulse durations that lie in the range of hundreds-of-femtoseconds to attoseconds. In electronics, however, these same terms refer to much longer pulses, namely durations of nanoseconds to tens-of-picoseconds, since the speed at which electronics operates lies well below that for optics. A nanosecond electrical pulse roughly has a GHz spectral width and so must be guided by a broadband microwave circuit. A picosecond electrical pulse, on the other hand, has a THz spectral width, which cannot be sustained by conventional electronic or microwave circuitry. Were a femtosecond electrical pulse to be generated, it would extend over a spectral band that spans the entire frequency range from DC to a PHz, beyond the ultraviolet edge of the visible band. By virtue of the time–energy uncertainty relation $\sigma_E\,\sigma_t \geq \hbar/2$ provided in (13.1-17), the energy uncertainty of such a pulse would exceed the bandgap energy of a typical semiconductor ($E_g \approx 1.5$ eV), rendering conventional electronics unreliable.

This Chapter

Ultrashort optical pulses may be directly generated by specially designed lasers that incorporate particular switching schemes or mode-locking methods (see Sec. 16.4). However, the shortest pulses capable of being directly generated by such devices are often not sufficiently short for certain applications. In this chapter, we demonstrate how ultrashort pulses can be shortened yet further, as well as reshaped, by making use of nonlinear dispersive optical components and systems.

The chapter begins with a description of the basic temporal and spectral characteristics of optical pulses (Sec. 23.1) and then considers their filtering by: (1) linear dispersive bulk optical components such as prisms and gratings (Sec. 23.2); and (2) transmission through linear dispersive media such as optical fibers (Sec. 23.3). Spatial effects, and the optics of pulsed waves with ultrabroad spectral widths, are examined in Sec. 23.4. The nonlinear optics of pulsed waves is addressed in Sec. 23.5, where several of the CW nonlinear optical phenomena introduced in Chapter 22, including parametric wave mixing, self-phase modulation, and optical solitons, are generalized to pulsed waves. Finally, various methods of measuring ultrashort optical pulses using "slow" detectors are reported in Sec. 23.6.

23.1 PULSE CHARACTERISTICS

A. Temporal and Spectral Characteristics

A pulse of light is described by an optical field of finite time duration. In this chapter we rely on the scalar wave theory of Chapter 2, representing the field components

with a generic complex wavefunction $U(\mathbf{r}, t)$ normalized such that the optical intensity is $I(\mathbf{r}, t) = |U(\mathbf{r}, t)|^2$ (W/m^2). When we are concerned with only the temporal or spectral properties of a pulse at a fixed position $\mathbf{r}$, we simply use the functions $U(t)$ and $I(t)$.

Temporal and Spectral Representations

The complex wavefunction describing an optical pulse of central frequency ν_0 is written in the form $U(t) = \mathcal{A}(t)\exp(j\omega_0 t)$, where $\mathcal{A}(t)$ is the **complex envelope** and $\omega_0 = 2\pi\nu_0$ is the central angular frequency. The complex envelope itself is characterized by its magnitude $|\mathcal{A}(t)|$ and phase $\varphi(t) = \arg\{\mathcal{A}(t)\}$, so that $U(t) = |\mathcal{A}(t)|\exp\,(j[\omega_0 t + \varphi(t)])$. The optical intensity $I(t) = |U(t)|^2 = |\mathcal{A}(t)|^2$ (W/m^2) and the area under the intensity function $\int I(t)dt$ is the energy density (J/m^2).

The intensity profiles of typical pulses include the Gaussian function, $I(t) \propto \exp(-2t^2/\tau^2)$ (which is examined in detail in Sec. 23.1B), the Lorentzian function $I(t) \propto 1/(1 + t^2/\tau^2)$, and the hyperbolic-secant function $I(t) \propto \text{sech}^2(t/\tau)$ (which appears in Sec. 23.5B in connection with optical solitons). The width of each of these pulses is proportional to the time constant τ.

In the spectral domain, the pulse is described by the Fourier transform $V(\nu) = \int U(t)\exp(-j2\pi\nu t)dt$, which is a complex function $V(\nu) = |V(\nu)|\exp[j\psi(\nu)]$. The squared magnitude $S(\nu) = |V(\nu)|^2$ is called the **spectral intensity** and $\psi(\nu)$ is the **spectral phase**. The function $V(\nu)$ is centered at the central frequency ν_0 and vanishes for negative ν since $U(t)$ is a complex analytic signal (Sec. 2.6A). The Fourier transform of the complex envelope, $A(\nu) = \int \mathcal{A}(t)\exp(-j2\pi\nu t)dt = V(\nu - \nu_0)$, is centered at $\nu = 0$. If the pulse has a narrow spectral width, then the complex envelope is a slowly varying function of time (i.e., varies slightly within an optical cycle $1/\nu_0$), but this is not the case for ultranarrow pulses with ultrawide spectral distributions. Figure 23.1-1 illustrates the various temporal and spectral functions that characterize an optical pulse.

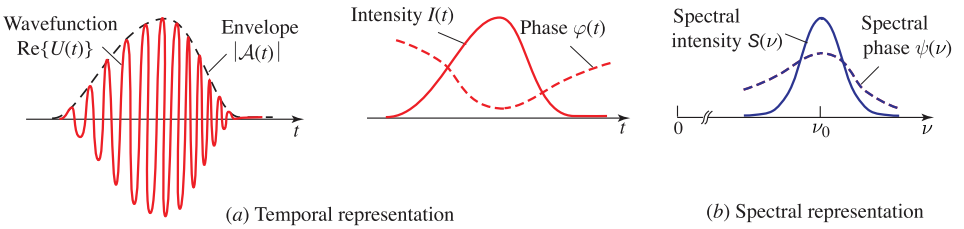


Figure 23.1-1 Temporal and spectral representations of an optical pulse. (*a*) The real part of the wavefunction $\text{Re}\{U(t)\} = |\mathcal{A}(t)|\cos[\omega_0 t + \varphi(t)]$, the magnitude of the envelope $|\mathcal{A}(t)|$, the intensity $I(t)$, and the phase $\varphi(t)$. (*b*) Spectral intensity $S(\nu)$ and spectral phase $\psi(\nu)$.

Temporal and Spectral Widths

The temporal and spectral widths of a pulse are the widths of the intensity $I(t) = |U(t)|^2$ and the spectral intensity $S(\nu) = |V(\nu)|^2$, respectively, as defined by any of the measures of width set forth in Appendix A.2. Unless otherwise specified, we will use the full-width at half-maximum (FWHM) definition and denote the temporal and spectral widths as τ_{FWHM} and $\Delta\nu$, respectively.

Because of the Fourier-transform relation between $U(t)$ and $V(\nu)$, the spectral width is inversely proportional to the temporal width. The coefficient of proportionality depends on the pulse shape and the definition of width. This inverse relation is illustrated in Fig. 23.1-2(*a*) for a Gaussian pulse for which $\tau_{\text{FWHM}}\Delta\nu = 0.44$.

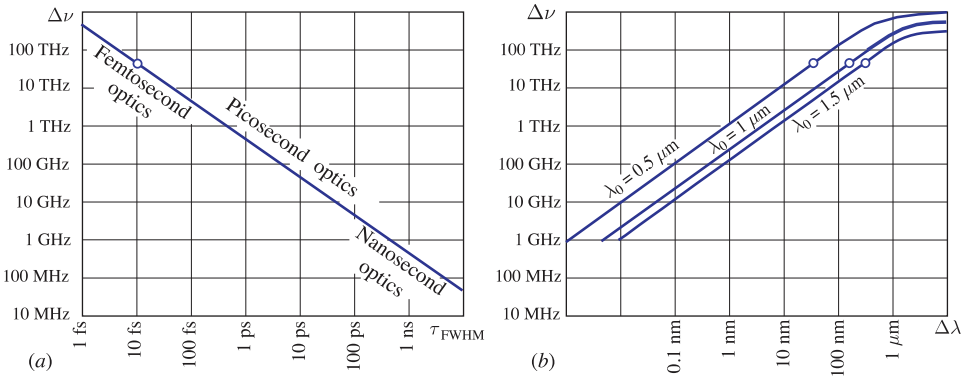


Figure 23.1-2 (*a*) The relation $\Delta\nu = 0.44/\tau_{\text{FWHM}}$ between the spectral width $\Delta\nu$ and the temporal width τ_{FWHM} for a Gaussian pulse. (*b*) The corresponding width $\Delta\lambda$ for a pulse of central frequency ν_0 corresponding to the central wavelengths $\lambda_0 = c/\nu_0$ = 0.5 μm, 1 μm, and 1.5 μm. As an example, a 10-fs pulse has a spectral width $\Delta\nu = 44$ THz, corresponding to $\Delta\lambda$ = 37 nm, 147 nm, and 331 nm, if the central wavelength is λ_0 = 0.5 μm, 1 μm, and 1.5 μm, respectively, as indicated by the open circles in the graph. This relation is linear if $\Delta\nu \ll \nu_0$ [see (23.1-1)].

The spectral intensity $S(\nu)$ is often plotted as a function of the wavelength, $S_\lambda(\lambda)$. This conversion is obtained by use of the relation $S_\lambda(\lambda) = S(\nu)|d\nu/d\lambda| = (c/\lambda^2)S(c/\lambda)$. The spectral width $\Delta\nu$ may also be converted into wavelength units. If $\Delta\nu \ll \nu_0$, then the spectral width in wavelength units is approximately $\Delta\lambda \approx |d\lambda/d\nu|\,\Delta\nu$, or

$$\Delta\lambda \approx \frac{\lambda_0^2}{c}\Delta\nu, \tag{23.1-1}$$

Spectral Width

where $\lambda_0 = c/\nu_0$ is the wavelength corresponding to the central frequency. If $\Delta\nu$ is in units of THz, λ_0 in μm, and $\Delta\lambda$ in nm, then

$$\Delta\lambda \approx 3.3\lambda_0^2\,\Delta\nu \qquad \Delta\lambda\ [\text{nm}];\ \ \lambda_0\ [\mu\text{m}];\ \ \Delta\nu\ [\text{THz}]. \tag{23.1-2}$$

For example, a spectral width $\Delta\nu = 1$ THz corresponds to $\Delta\lambda$ = 1 nm at λ_0 = 0.55 μm, and to 4 nm at λ_0 = 1.1 μm. This relation is illustrated in 23.1-2(*b*).

For ultranarrow pulses with large $\Delta\nu$, the exact expression for $\Delta\lambda$ is

$$\Delta\lambda = \frac{c}{\nu_0 - \Delta\nu/2} - \frac{c}{\nu_0 + \Delta\nu/2} = \frac{\lambda_0^2}{c}\frac{\Delta\nu}{1 - (\Delta\nu/2\nu_0)^2}. \tag{23.1-3}$$

However, under these conditions, the concept of spectral width loses its significance. A 2-fs pulse, e.g., has spectral width $\Delta\nu$ = 220 THz, corresponding to $\Delta\lambda$ = 847 nm at $\lambda_0 = 1$ μm, i.e., the spectrum is quite broad and extends from visible through infrared.

Instantaneous Frequency

Another descriptor of the optical pulse is the time dependence of its instantaneous frequency. The instantaneous angular frequency ω_{i} is the derivative of the phase of $U(t)$, and the instantaneous frequency $\nu_{\text{i}} = \omega_{\text{i}}/2\pi$, so that

$$\omega_{\text{i}} = \omega_0 + \frac{d\varphi}{dt}, \qquad \nu_{\text{i}} = \nu_0 + \frac{1}{2\pi}\frac{d\varphi}{dt}. \tag{23.1-4}$$

Instantaneous Frequency

If the phase is a linear function of time, $\varphi(t) = 2\pi f t$, then the instantaneous frequency $\nu_{\text{i}} = \nu_0 + f$; i.e., a linearly varying phase corresponds to a fixed frequency shift. Nonlinear time dependence of the phase corresponds to time-dependent instantaneous frequency.

Chirped Pulses

A pulse is said to be **chirped**, or frequency modulated (FM), if its instantaneous frequency is time varying. If ν_{i} is an increasing function of time at the pulse center ($t = 0$), i.e., $\varphi'' = d^2\varphi/dt^2 > 0$, then the pulse is said to be **up-chirped**. If ν_{i} is a decreasing function of time at the pulse center, i.e., $\varphi'' < 0$, it is said to be **down-chirped**.

In particular, if the phase of an optical pulse of width τ is a quadratic function of time $\varphi(t) = at^2/\tau^2$, where a is a constant, then $\varphi'' = 2a/\tau^2$ so that the instantaneous frequency $\nu_{\text{i}} = \nu_0 + (a/\pi\tau^2)t$ is a linear function of time. The pulse is then said to be **linearly chirped** and the parameter

$$a = \tfrac{1}{2}\varphi''\tau^2 \qquad (23.1\text{-}5)$$

Chirp Parameter

is called the **chirp parameter**. The pulse is up-chirped if $a > 0$ and down-chirped if $a < 0$. At $t = \tau/2$, the instantaneous frequency increases by $a/2\pi\tau$, which is of the order of magnitude of $a\Delta\nu$. Thus, the chirp parameter is indicative of the ratio between the instantaneous frequency change at the pulse half-width point and the spectral width $\Delta\nu$. Examples of linearly chirped pulses and their instantaneous frequencies are illustrated in Fig. 23.1-3.

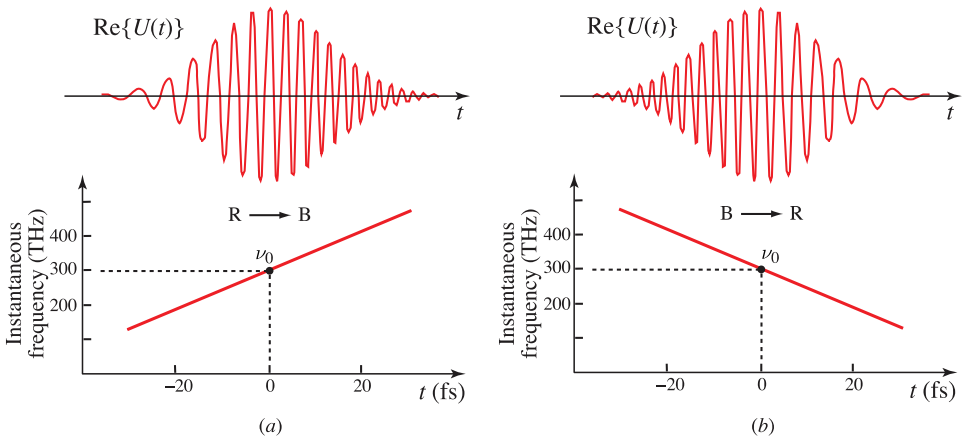


Figure 23.1-3 Linearly up-chirped and down-chirped optical pulses. (*a*) An up-chirped pulse has an increasing instantaneous frequency. (*b*) A down-chirped pulse has a decreasing instantaneous frequency. In this figure, the pulse width is 20 fs and the central frequency $\nu_0 = 300$ THz. The letters R and B, which represented red and blue, are generic indicators of long and short wavelengths, respectively.

If the dependence of the phase φ on time is an arbitrary nonlinear function, as in Fig. 23.1-1, then it can be approximated by a Taylor-series expansion in the vicinity of the pulse center, and the chirp coefficient a defined by (23.1-5) then represents the lowest-order chirping effect resulting from the quadratic term of the expansion.

Time-Varying Spectrum

It is often useful to trace the spectral changes of a time-varying pulse throughout its time course. Such changes are obscured in the Fourier transform, which only provides an average spectral representation of the entire signal without noting which frequencies occur at which times. This is particularly evident if the signal is composed of a sequence of segments each with a different spectral composition. A good example is a musical signal for which the spectral changes indicate changes of the musical score as time progresses.

While the instantaneous frequency can be a measure of the time-dependent nature of the spectrum, it is not always adequate since it is based only on the phase and ignores the amplitude. A commonly used measure is based on a sliding window, or gate, that selects only one short time segment at a time, and obtains the Fourier transform of the pulse within the window duration. This is repeated at different locations of the sliding window, as illustrated in Fig. 23.1-4, and the result is plotted as a function of both frequency and time delay. The resultant 2D function is called the **short-time Fourier transform**. Its squared magnitude is called the **spectrogram** and is often plotted as a picture with the horizontal and vertical axes representing time and frequency, respectively, as illustrated in Fig. 23.1-4.

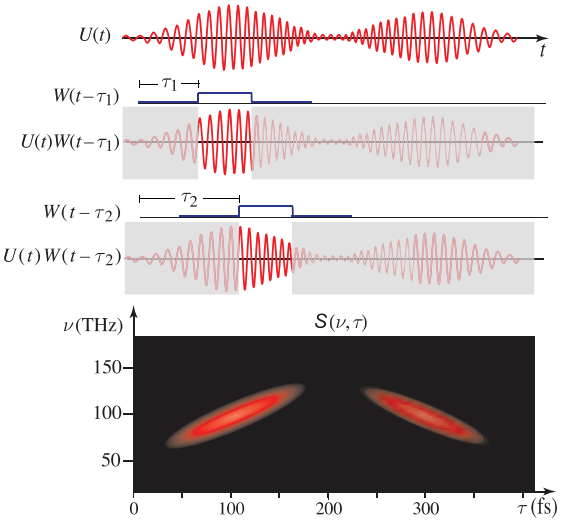


Figure 23.1-4 The short-time Fourier transform of $U(t)$ is constructed by making use of a sequence of Fourier transforms of $U(t)$ multiplied by a moving window $W(t-\tau)$. The spectrogram $S(\nu,\tau)$ is the squared magnitude of these Fourier transforms. In this example, $U(t)$ comprises two Gaussian pulses, each of time constant τ = 60 fs and central frequency 100 THz. The first pulse is up-chirped ($a = 5$) whereas the second, which has a smaller amplitude, is down-chirped ($a = -5$). The window function $W(t)$ is Gaussian with time constant τ = 20 fs.

If $W(t)$ is a **window function** of short duration T beginning at $t = 0$, and if $U(t)$ is the pulse wavefunction, then the product $U(t)W(t-\tau)$ is a segment of the pulse of duration T beginning at time τ. The Fourier transform of the segment is

$$\Phi(\nu,\tau) = \int U(t)W(t-\tau)\exp(-j2\pi\nu t)dt. \tag{23.1-6}$$

Short-Time Fourier Transform

The function $\Phi(\nu,\tau)$ is the short-time Fourier transform and its squared magnitude $S(\nu,\tau) = |\Phi(\nu,\tau)|^2$ is the spectrogram.

B. Gaussian and Chirped Gaussian Pulses

Transform-Limited Gaussian Pulse

A **transform-limited** Gaussian pulse has a complex envelope with constant phase and Gaussian magnitude, so that

$$\mathcal{A}(t) = A_0 \exp(-t^2/\tau^2), \tag{23.1-7}$$

where τ is a real time constant. The intensity $I(t) = I_0 \exp(-2t^2/\tau^2)$ is also a Gaussian function, with peak value $I_0 = |A_0|^2$, $1/e$ full width $\sqrt{2}\,\tau$, and FWHM

$$\tau_{\text{FWHM}} = \sqrt{2 \ln 2}\,\tau = 1.18\,\tau. \tag{23.1-8}$$

The Fourier transform of the complex envelope, $A(\nu) \propto \exp\left(-\pi^2\tau^2\nu^2\right)$, is a Gaussian function, and so is the spectral intensity

$$S(\nu) \propto \exp\left[-2\pi^2\tau^2(\nu - \nu_0)^2\right]. \tag{23.1-9}$$

The FWHM of the spectral intensity is

$$\Delta\nu = 0.375/\tau = 0.44/\tau_{\text{FWHM}}, \tag{23.1-10}$$

so that the product of the FWHM temporal and spectral widths is $\tau_{\text{FWHM}}\Delta\nu = 0.44$. Figure 23.1-5($a$) illustrates the temporal and spectral characteristics of the transform-limited Gaussian pulse.

As discussed in Appendix A.2, the transform-limited Gaussian pulse has a minimum temporal- and spectral-width product, and this is why it is called transform limited (also called **Fourier-transform limited** or **bandwidth limited**).

Though the Gaussian pulse has an ideal shape that is not encountered exactly in practice, it is a useful approximation that lends itself to analytical studies.

Chirped Gaussian Pulse

A more general Gaussian pulse has a complex envelope $\mathcal{A}(t) = A_0 \exp\left(-\alpha t^2\right)$, where $\alpha = (1 - ja)/\tau^2$ is a complex parameter and τ and a are real parameters, so that

$$\mathcal{A}(t) = A_0 \exp(-t^2/\tau^2) \exp(jat^2/\tau^2). \tag{23.1-11}$$

The magnitude of the complex envelope is a Gaussian function $|A_0| \exp(-t^2/\tau^2)$ and the intensity is also Gaussian. The phase is a quadratic function $\varphi = at^2/\tau^2$ so that the instantaneous frequency $\nu_{\text{i}} = \nu_0 + at/\pi\tau^2$ is a linear function of time; i.e., the pulse is linearly chirped with chirp parameter a. The pulse is up-chirped for positive a, down-chirped for negative a, and transform-limited (unchirped) for $a = 0$. The Fourier transform of the complex envelope $\mathcal{A}(t) = A_0 \exp\left(-\alpha t^2\right)$ is proportional to $\exp\left(-\pi^2\tau^2\nu^2/\alpha\right)$, which is also a Gaussian function of frequency. The spectral intensity $S(\nu)$ is proportional to $\exp\left[-2\pi^2\tau^2(\nu - \nu_0)^2/(1 + a^2)\right]$, which is Gaussian with FWHM $\Delta\nu = (0.375/\tau)\sqrt{1 + a^2} = (0.44/\tau_{\text{FWHM}})\sqrt{1 + a^2}$. This is a factor of $\sqrt{1 + a^2}$ greater than that of an unchirped pulse (a = 0) of the same time constant τ. The product of the FWHM temporal and spectral widths is $\tau_{\text{FWHM}}\Delta\nu = 0.44\sqrt{1 + a^2}$, so that the unchirped Gaussian pulse (a = 0) has the smallest temporal- and spectral-width product. The spectral phase $\psi(\nu) \propto a\nu^2$ is a quadratic function of frequency.

Key equations characterizing the chirped Gaussian pulse are provided in Table 23.1-1. Figure 23.1-5 illustrates the temporal and spectral characteristics of transform-limited and chirped Gaussian pulses.

Table 23.1-1 Temporal and spectral properties of a chirped Gaussian pulse of peak amplitude A_0, peak intensity $I_0 = |A_0|^2$, central frequency ν_0, time constant τ, and chirp parameter a.

$\mathcal{A}(t) = A_0 \exp[-(1-ja)t^2/\tau^2]$	Complex envelope	(23.1-12)
$I(t) = I_0 \exp(-2t^2/\tau^2)$	Intensity	(23.1-13)
$\int I(t)dt = \sqrt{\pi/2}\, I_0 \tau$	Energy density	(23.1-14)
$\tau_{1/e} = \sqrt{2}\,\tau$	$1/e$ half width	(23.1-15)
$\tau_{\text{FWHM}} = 1.18\,\tau$	FWHM width	(23.1-16)
$\varphi(t) = at^2/\tau^2$	Phase	(23.1-17)
$A(\nu) = \dfrac{\sqrt{\pi}\, A_0 \tau}{\sqrt{(1-ja)}} \exp\left[-\dfrac{\pi^2\tau^2\nu^2}{1-ja}\right]$	Fourier transform	(23.1-18)
$S(\nu) = \dfrac{\pi\, I_0 \tau^2}{\sqrt{1+a^2}} \exp\left[-\dfrac{2\pi^2\tau^2(\nu-\nu_0)^2}{1+a^2}\right]$	Spectral intensity	(23.1-19)
$\Delta\nu_{1/e} = \dfrac{2}{\tau}\sqrt{1+a^2}$	$1/e$ half width	(23.1-20)
$\Delta\nu = \dfrac{0.375}{\tau}\sqrt{1+a^2} = \dfrac{0.44}{\tau_{\text{FWHM}}}\sqrt{1+a^2}$	FWHM spectral width	(23.1-21)
$\psi(\nu) = -2\pi^2\tau^2[a/(1+a^2)]\nu^2$	Spectral phase	(23.1-22)
$\nu_{\text{i}} = \nu_0 + (a/\pi\tau^2)t$	Instantaneous frequency	(23.1-23)

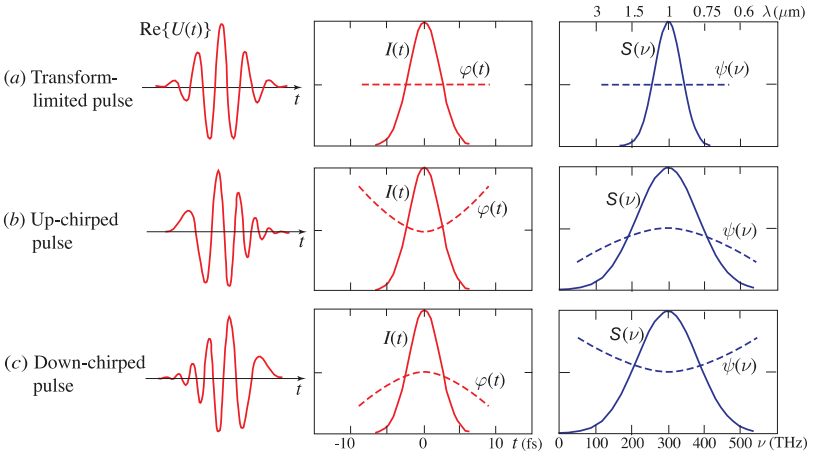


Figure 23.1-5 Temporal and spectral profiles of three Gaussian pulses of central frequency ν_0= 300 THz (corresponding to a wavelength of 1 μm and a 3.3-fs optical cycle) and width τ_{FWHM} = 5 fs (τ = 4.23 fs). (*a*) Transform-limited pulse; the spectral width $\Delta\nu = 88$ THz ($\Delta\lambda$ = 73 nm). (*b*) Up-chirped pulse of chirp parameter $a = 2$; the spectral width is a factor of $\sqrt{1+a^2} = \sqrt{5}$ greater than in (*a*), so that $\Delta\nu = 197$ THz. The instantaneous frequency is a linearly increasing function of time with value ν_0= 300 THz at $t = 0$ (center of the pulse) and values $\nu_{\text{i}} = \nu_0(1 \pm at/\pi\nu_0\tau) = 300(1 \pm 0.497)$ THz at $t = \pm\tau$. The frequency is swept between 151 THz and 449 THz as t changes from $-\tau$ to $+\tau$. This corresponds to a change of the wavelength between 0.67 μm and 1.99 μm. (*c*) Same as in (*b*) but the pulse is down-chirped with chirp parameter $a = -2$.

C. Spatial Characteristics

In this section we examine a few simple examples of pulsed optical waves traveling in free space or, alternatively, in a linear, homogeneous, and nondispersive medium. In such media, in accordance with (2.2-4), the complex wavefunction $U(\mathbf{r}, t)$ obeys the wave equation $\nabla^2 U - (1/c^2)\partial^2 U/\partial t^2 = 0$. The simplest exact solutions of this equation are the pulsed plane wave and the pulsed spherical wave. We discuss these solutions and also introduce the pulsed Gaussian beam. A more detailed study of the spatial properties of pulsed light is deferred to Sec. 23.4.

Pulsed Plane Wave

As discussed in Sec. 2.6A, a pulsed plane wave traveling in the z direction has a complex wavefunction of the form $U(r, t) = \mathcal{A}(t - z/c)\exp[j\omega_0(t - z/c)]$, where $\mathcal{A}(t)$ is an arbitrary function. The corresponding intensity is $I(t - z/c)$, where $I(t) = |\mathcal{A}(t)|^2$. If the width of $I(t)$ is τ, then the traveling pulse occupies a distance $\Delta z = c\tau$ at any time and travels without change at a velocity c, as illustrated in Fig. 23.1-6. Numerical values of the pulse temporal and spatial widths in free space are:

Temporal width τ	1 ns	1 ps	1 fs	1 as
Spatial width $c\tau$	30 cm	0.3 mm	0.3 μm	0.3 nm

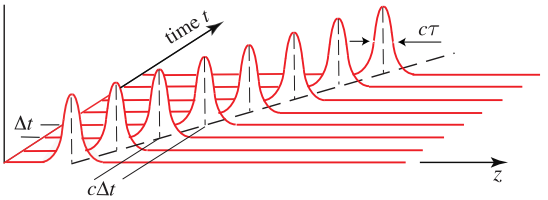


Figure 23.1-6 The envelope of a plane-wave pulse of width τ traveling in the z direction with velocity c. The pulse occupies a distance $c\tau$ at any time.

A pulsed plane wave traveling at an angle θ with the z axis has a complex wavefunction $U(r, t) = \mathcal{A}[t - (x\sin\theta + z\cos\theta)/c]\exp[-jk_0(x\sin\theta + z\cos\theta)]\exp(j\omega_0 t)$ and intensity $I[t - (x\sin\theta + z\cos\theta)/c]$, where $I(t) = |\mathcal{A}(t)|^2$. If this intensity is recorded as a function of x and z in a sequence of snapshots (each at a fixed time), then the result is as illustrated in Fig. 23.1-7(a). The bright stripe in each snapshot represents the traveling pulse at a given time. For example, a 100-fs pulse in free space appears as a stripe of width 30 μm. Note that a single vertical line (fixed z) intercepting the stripe in a single snapshot (fixed t) provides a complete record of the pulse temporal profile since it records the function $I(-x\sin\theta/c + \text{constant})$. Thus, the temporal profile may be measured by observing the spatial profile of a snapshot of the pulse. This can be utilized for pulse detection, as will be discussed in Sec. 23.5B.

Pulsed Spherical Wave

Another simple solution of the wave equation is the pulsed spherical wave $U(r, t) = (1/r)g(t - r/c)\exp[j\omega_0(t - r/c)]$, where $g(t)$ is an arbitrary function. The pulse expands radially and its wavefronts are concentric spheres, as illustrated in Fig. 23.1-7(b). At any fixed time, it occupies a spherical shell of radial width $c\tau$, where τ is the width of $g(t)$.

⋆Paraxial Wave Modulated by Slowly Varying Pulse

When the envelope of a pulsed wave varies slowly with time so that it is approximately constant within an optical cycle, it is said to have a **slowly varying envelope** (**SVE**). Because of the associated narrow spectral width, $\Delta\nu \ll \nu_0$, the spatial behavior is

approximately the same as that of a monochromatic (CW) wave at the central frequency ν_0 or the wavelength $\lambda_0 = c/\nu_0$. The wave may therefore be regarded as a **quasi-CW** pulsed wave.

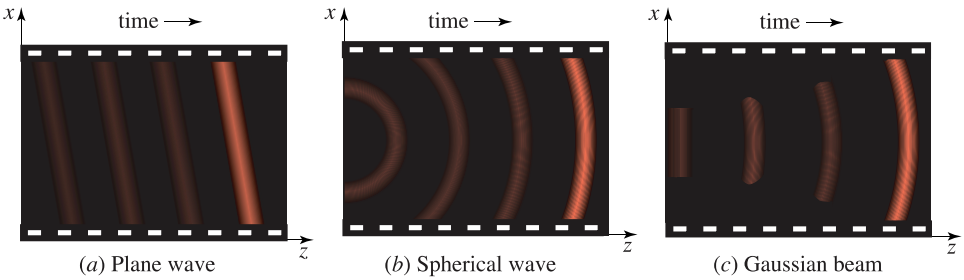
If the wave is also paraxial (see Sec. 2.2C), it may be expressed in terms of its envelope in the general form $U(r,t) = \mathcal{A}(\mathbf{r},t)\exp(-jk_0z)\exp(j\omega_0t)$, where the envelope varies slowly with z so that it is approximately constant within a distance equal to a wavelength $\lambda_0 = 2\pi/k_0$; i.e., the condition $\partial^2\mathcal{A}/\partial z^2 \ll k_0^2\mathcal{A}$ is satisfied. Since the envelope is also slowly varying in time, the approximation $\partial^2\mathcal{A}/\partial t^2 \ll \omega_0^2\mathcal{A}$ is also applicable. Under such conditions, the wave equation $\nabla^2 U - (1/c^2)\partial^2 U/\partial t^2 = 0$ leads to an approximate equation for the envelope,

$$\nabla_T^2\mathcal{A} - j\frac{4\pi}{\lambda_o}\left(\frac{\partial\mathcal{A}}{\partial z} + \frac{1}{c}\frac{\partial\mathcal{A}}{\partial t}\right) = 0, \tag{23.1-24}$$

Paraxial SVE Equation

where $\nabla_T^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the transverse Laplacian operator. Equation (23.1-24) is known as the paraxial SVE equation. For a CW wave, $\partial\mathcal{A}/\partial t = 0$ and (23.1-24) reproduces the paraxial Helmholtz equation (2.2-23).

As can be seen by direct substitution, (23.1-24) is satisfied by $\mathcal{A}(\rho,z,t) = g(t - z/c)\mathcal{A}_0(\mathbf{r})$, where g is an arbitrary function of the retarded time $t - z/c$ and $\mathcal{A}_0(\mathbf{r})$ satisfies the paraxial Helmholtz equation $\nabla_T^2\mathcal{A}_0 - j(4\pi/\lambda_0)\partial\mathcal{A}_0/\partial z = 0$, which is applicable in the CW case. It follows that in this approximation a paraxial wave at the wavelength λ_0 may be modulated by a slowly varying pulse of arbitrary shape, without altering its spatial behavior.



(*a*) Plane wave (*b*) Spherical wave (*c*) Gaussian beam

Figure 23.1-7 (*a*) Four snapshots (taken at equal time intervals) of a pulsed plane wave traveling at an angle. Each snapshot contains a single line of width $c\tau$ (in the z direction), where τ is the pulse width. The line moves from left to right as the pulsed wave propagates. (*b*) Same as (*a*) but for a spherical wave; (*c*) Same as (*a*) but for a Gaussian beam.

Pulsed Gaussian Beam

One of the solutions of the paraxial Helmholtz equation is the Gaussian beam described by (3.1-5). In the pulsed quasi-CW case, the Gaussian beam is given by

$$\mathcal{A}(\rho,z,t) = g(t - z/c)\frac{jz_0}{z + jz_0}\exp\left(-j\frac{\pi}{\lambda_0}\frac{\rho^2}{z + jz_0}\right), \tag{23.1-25}$$

where $g(t)$ is an arbitrary slowly varying function of the retarded time $t - z/c$ and z_0 is the Rayleigh range (also called the diffraction length). In this approximation, except for the retardation effect, there is no coupling between space and time; i.e., the beam maintains its Gaussian spatial profile at all times, and the pulse maintains its initial temporal profile at all positions. Snapshots of such a beam are illustrated in Fig. 23.1-7(*c*).

It will be shown in Sec. 23.4 that for ultranarrow pulses, for which the SVE approximation is not applicable, space–time coupling can be significant, and a wave that is Gaussian in time and space in a given transverse plane becomes non-Gaussian in both time and space as it propagates in free space.

23.2 PULSE SHAPING AND COMPRESSION

The temporal profile of a short optical pulse is unavoidably altered as it travels through a dispersive optical system. This is because the individual spectral components that constitute the pulse are attenuated and/or phase shifted by different amounts. The shorter the optical pulse, the greater is its spectral width and thus the more dramatic is the effect of dispersion. On the other hand, dispersive optical elements may be designed to effect desired changes in the shape of an optical pulse, such as compression or stretching.

In this section, we study only temporal effects in linear dispersive optical media, i.e., only pulsed plane waves are considered. A discussion of spatial effects in linear optical media, including diffraction and beam propagation in dispersive media, is reserved for Sec. 23.4. Dispersion in nonlinear systems is examined in Sec. 23.5.

A. Chirp Filters

Linear Filtering of an Optical Pulse

The transmission of an optical pulse through an arbitrary linear optical system is generally described by the theory of linear systems (see Appendix B). A linear time-invariant system is characterized by a transfer function $H(\nu)$, which is the factor by which the Fourier component of the input pulse at frequency ν is multiplied to generate the output component at the same frequency. If $U_1(t)$ and $U_2(t)$ are the complex wavefunctions of the original and filtered pulses, respectively, then their Fourier transforms $V_1(\nu)$ and $V_2(\nu)$ are related by

$$V_2(\nu) = H(\nu)\, V_1(\nu). \tag{23.2-1}$$

In using (23.2-1) we only need to know $H(\nu)$ at frequencies within the spectral band of the pulse, which is a region of width $\Delta\nu$ surrounding the central frequency ν_0, as illustrated in Fig. 23.2-1. When $\Delta\nu \ll \nu_0$, it is convenient to work with the complex envelope instead of the wavefunction. Using the relation $U(t) = \mathcal{A}(t)\exp(j2\pi\nu_0 t)$ and the shift property of the Fourier transform, $V(\nu) = A(\nu - \nu_0)$, where $A(\nu)$ is the Fourier transform of $\mathcal{A}(t)$, it follows from (23.2-1) that $A_2(\nu-\nu_0) = H(\nu)A_1(\nu-\nu_0)$, where the subscripts 1 and 2 denote the input and output pulses, respectively. Defining the frequency difference $f = \nu - \nu_0$, we obtain $A_2(f) = H(\nu_0 + f)A_1(f)$, or

$$A_2(f) = H_e(f)A_1(f), \tag{23.2-2}$$

where

$$\boxed{H_e(f) = H(\nu_0 + f)} \tag{23.2-3}$$

Envelope Transfer Function

is called the **envelope transfer function**. Working with (23.2-2) is generally more convenient than working with (23.2-1), since the frequency f is typically much smaller than ν. These relations are illustrated in Fig. 23.2-1.

The transfer functions $H(\nu)$ and $H_e(f)$ are complex functions, $H(\nu) = |H(\nu)|\exp[-j\Psi(\nu)]$ and $H_e(f) = |H_e(f)|\exp[-j\Psi_e(f)]$, where $\Psi_e(f) = \Psi(\nu_0 + f)$ are

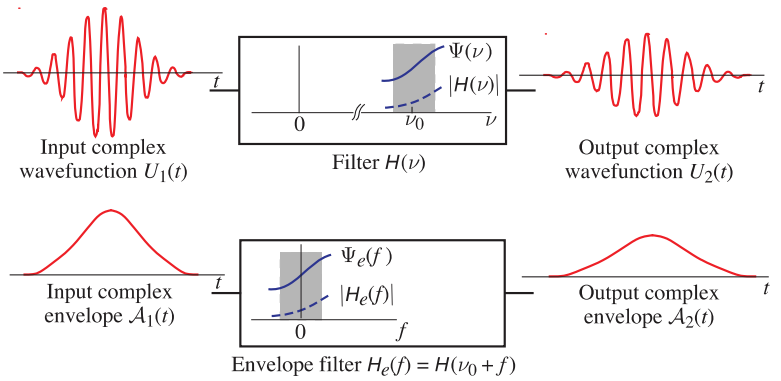


Figure 23.2-1 Filtering the wavefunction with a filter $H(\nu)$ (upper figure) is equivalent to filtering the envelope with a filter $H_e(f) = H(\nu_0 + f)$ (lower figure). The shaded area represents the spectral band of interest.

real functions representing the phase transfer. The phase introduced by the filter often plays a more important role than the magnitude in the reshaping of pulses. Throughout this chapter we will deal with **phase filters**, i.e., filters for which the magnitude $|H(\nu)|$ is approximately constant within the frequency range of interest.

When transformed to the time domain, (23.2-2) becomes the convolution relation

$$\mathcal{A}_2(t) = \int_{-\infty}^{\infty} h_e(t - t')\, \mathcal{A}_1(t')dt', \tag{23.2-4}$$

where $h_e(t)$ is the inverse Fourier transform of $H_e(f)$.

The Ideal Filter

An **ideal filter** preserves the shape of the input pulse envelope; it merely multiplies it by a constant (of magnitude < 1 for an attenuator and > 1 for an amplifier), and possibly delays it by a fixed time. The transfer function has the form

$$H_e(f) = H_0 \exp(-j2\pi\tau_d f), \tag{23.2-5}$$

where H_0 is a constant, $G = |H_0|^2$ is the intensity reduction or gain factor, and τ_d is the time delay. The phase is a linear function of frequency $\Psi_e(f) = \Psi_0 + 2\pi\tau_d f$, where $\Psi_0 = \arg\{H_0\}$ is a constant phase [see Fig. 23.2-2(*a*)]. Using a basic Fourier-transform property (see Appendix A), the phase $2\pi\tau_d f$ is equivalent to a time delay τ_d. The input and output envelopes are related by $\mathcal{A}_2(t) = H_0\mathcal{A}_1(t - \tau_d)$, and the intensities are related by $I_2(t) = GI_1(t - \tau_d)$. For a distributed attenuator/amplifier of attenuation/gain coefficient α, velocity c, and length d, the transfer function is $H_e(f) = \exp(-\alpha d/2)\exp(-j2\pi f d/c)$ so that $G = \exp(-\alpha d)$ and $\tau_d = d/c$. A slab of ideal nondispersive material with attenuation coefficient α and refractive index n is an example of such filter, where $c = c_o/n$. Here, the transfer function $H(\nu) = \exp(-\alpha d/2)\exp(-j\beta d)$, where $\beta = 2\pi\nu/c$ is the propagation constant (see Sec. 5.5A), and $H_e(f) = \exp(-\alpha d/2)\exp(-j2\pi f d/c)$. When α and n are frequency dependent, i.e., the medium is dispersive, the filter is not ideal and the pulse shape may be significantly altered, as will be shown in Sec. 23.3.

The Chirp Filter

Perhaps the most important filter in ultrafast optics is the Gaussian chirp filter, often simply called the **chirp filter**. It is a phase filter whose phase is a quadratic function of

frequency $\Psi_e(f) = b\pi^2 f^2$ [see Fig. 23.2-2(*b*)] so that the envelope transfer function is Gaussian,

$$H_e(f) = \exp\left(-jb\pi^2 f^2\right), \tag{23.2-6}$$

Chirp-Filter Transfer Function

where b is a real parameter (units of s^2) called the **chirp coefficient** of the filter. For $b > 0$ the filter is said to be **up-chirping**, and for $b < 0$ it is **down-chirping**.

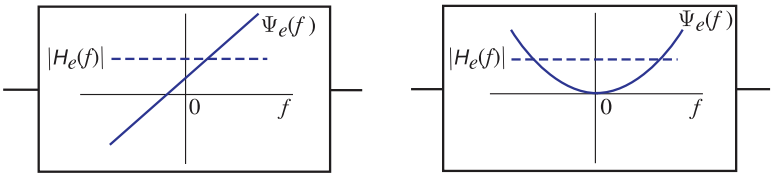
The corresponding **impulse response function** is the inverse Fourier transform of (23.2-6) (see Table A.1-1), which is another Gaussian function

$$h_e(t) = \frac{1}{\sqrt{j\pi b}} \exp(jt^2/b). \tag{23.2-7}$$

Chirp-Filter Impulse Response Function

It too has a phase that is a quadratic function of time, i.e., it is a linearly chirped function, which is up-chirped for positive b and down-chirped for negative b.

A cascade of two chirp filters with coefficients b_1 and b_2 is equivalent to a single chirp filter with coefficient $b = b_1 + b_2$, since the transfer functions multiply. Thus, a down-chirping filter may compensate the effect of an up-chirping filter, so that the action of a chirp filter is reversible.



(*a*) Ideal filter (*b*) Chirp filter

Figure 23.2-2 Magnitude and phase of the envelope transfer functions for (*a*) an ideal filter, and (*b*) a chirp filter (with $b > 0$).

As can be seen by substituting (23.2-7) into (23.2-4), the pulse envelopes at the output and input of a chirp filter are related by

$$\mathcal{A}_2(t) = \frac{1}{\sqrt{j\pi b}} \int_{-\infty}^{\infty} \mathcal{A}_1(t') \exp\left[j\frac{(t-t')^2}{b}\right] dt'. \tag{23.2-8}$$

This transformation is mathematically similar to Fresnel diffraction [see (4.3-12) in Sec. 4.3B], and for a sufficiently large chirp parameter b it becomes similar to Fraunhofer diffraction, i.e., equivalent to a Fourier transform (see Sec. 4.3A). The analogy between diffraction in space and dispersion in time, which is described by a chirp filter, is formally established in Sec. 23.3D.

Approximation of Arbitrary Phase Filter by a Chirp Filter

When the filter magnitude and phase vary slowly within the narrow spectral width of a pulse, we may assume that the magnitude is approximately constant at its central-frequency value, $|H(\nu_0 + f)| \approx |H(\nu_0)| \equiv |H_0|$, and expand the phase function $\Psi(\nu)$ in a Taylor series centered at the frequency ν_0. Retaining only the first three terms, $\Psi(\nu_0 + f) \approx \Psi_0 + \Psi' f + \frac{1}{2}\Psi'' f^2$, where $\Psi_0 = \Psi(\nu_0), \Psi' = d\Psi/d\nu|_{\nu_0}, \Psi'' = d^2\Psi/d\nu^2|_{\nu_0}$, we obtain $H(\nu_0 + f) \approx |H_0| \exp[-j(\Psi_0 + \Psi' f + \frac{1}{2}\Psi'' f^2)]$.

It follows from (23.2-3) that the envelope transfer function may therefore be approximated by

$$H_e(f) \approx |H_0| \exp\left[-j(\Psi_0 + \Psi' f + \tfrac{1}{2}\Psi'' f^2)\right]. \tag{23.2-9}$$

This filter is equivalent to a cascade of an ideal filter and a chirp filter (see Fig. 23.2-3). The ideal filter is composed of a constant multiplier $H_0 = |H_0| \exp(-j\Psi_0)$, which does not alter the shape of the pulse and may be ignored, and a phase shift $\exp(-j2\pi\tau_d f)$, which is equivalent to a time delay

$$\boxed{\tau_d = \Psi'/2\pi.} \tag{23.2-10}$$

Group Delay

The chirp filter has a transfer function $\exp\left(-jb\pi^2 f^2\right)$ with chirp coefficient

$$\boxed{b = \frac{\Psi''}{2\pi^2}.} \tag{23.2-11}$$

Chirp Coefficient

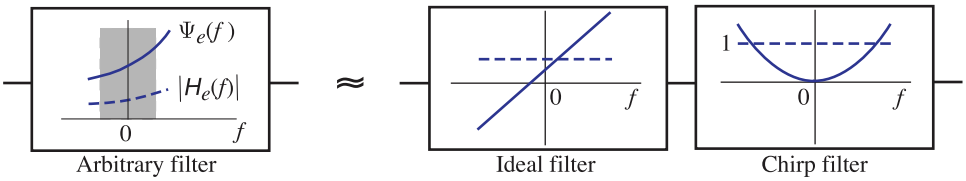


Figure 23.2-3 Approximation of an arbitrary filter with slowly varying transfer function as a cascade of an ideal filter (including a time delay) and a chirp filter.

We conclude that the principal source of distortion in a dispersive system with slowly varying phase is described by a chirp filter. Examples of such systems based on angular dispersion and Bragg gratings are considered subsequently in this section. Dispersive media are also described by chirp filters, as will be shown in Sec. 23.3. A more accurate approximation of the phase filter would require the inclusion of additional terms in the Taylor-series expansion of the phase $\Psi(\nu)$. The third-order term corresponds to a phase filter $\exp(-j\frac{1}{6}\Psi''' f^3)$, and higher-order terms can be similarly defined.

Chirp Filtering of a Transform-Limited Gaussian Pulse

We now consider the effect of a chirp filter with a transfer function given by $H_e(f) = \exp\left(-jb\pi^2 f^2\right)$ and chirp coefficient b on an unchirped (transform-limited) Gaussian pulse of complex envelope $\mathcal{A}_1(t) = A_{10}\exp(-t^2/\tau_1^2)$. Since the Fourier transform of $\mathcal{A}_1(t)$ is $A_1(f) = (A_{10}\tau_1/2\sqrt{\pi})\exp(-\pi^2\tau_1^2 f^2)$, by virtue of (23.2-2) the filtered pulse has a complex envelope with Fourier transform

$$A_2(f) = A_{10}\frac{\tau_1}{2\sqrt{\pi}}\exp[-\pi^2(\tau_1^2 + jb)f^2]. \tag{23.2-12}$$

This expression may be cast as the Fourier transform of a chirped Gaussian pulse of width τ_2 and chirp parameter a_2, which, in accordance with (23.1-18), has a Fourier transform

$$A_2(f) = A_{20}\frac{\tau_2}{2\sqrt{\pi(1-ja_2)}}\exp\left(-\frac{\pi^2\tau_2^2 f^2}{1-ja_2}\right). \tag{23.2-13}$$

Equating the exponents in (23.2-12) and (23.2-13), we obtain

$$\tau_1^2 + jb = \frac{\tau_2^2}{1-ja_2}, \tag{23.2-14}$$

and equating the amplitudes we obtain $A_{20} = A_{10}\sqrt{1-ja_2}\,\tau_1/\tau_2$. Equating the real and imaginary parts of (23.2-14) leads to the expressions that relate the parameters of the output pulse to those of the input pulse:

$$\text{Width} \quad \tau_2 = \tau_1\sqrt{1+b^2/\tau_1^4}, \tag{23.2-15}$$

$$\text{Chirp parameter} \quad a_2 = b/\tau_1^2, \tag{23.2-16}$$

$$\text{Amplitude} \quad A_{20} = \frac{A_{10}}{\sqrt{1+jb/\tau_1^2}}. \tag{23.2-17}$$

We conclude that upon transmission through a chirp filter, an unchirped Gaussian pulse remains Gaussian and its properties are modified as follows:

- *The pulse width is increased* by a factor $\sqrt{1+a_2^2} = \sqrt{1+b^2/\tau_1^4}$. For $|b| = \tau_1^2$, this factor is $\sqrt{2}$. Thus, the filter begins to have a significant effect when its chirp coefficient is of the order of the squared width of the original pulse. For $|b| \gg \tau_1^2$, i.e., for large chirp coefficient or narrow original pulse, $\tau_2 \approx |b|/\tau_1$, indicating that the width of the filtered pulse is directly proportional to $|b|$ and inversely proportional to τ_1, so that narrower pulses undergo greater broadening.
- *The initially transform-limited pulse becomes chirped* with a chirp parameter a_2 that is directly proportional to the filter chirp coefficient b and inversely proportional to the square of the original pulse width. The filtered pulse will be up-chirped if b is positive, i.e., if the filter is up-chirping, and will be down-chirped if b is negative, i.e., the filter is down-chirping. For $b = \tau_1^2$, the chirp parameter $a_2 = 1$.
- *The spectral width of the pulse remains unchanged.* The original pulse has a spectral width $\Delta\nu = 0.375/\tau_1$, and the filtered pulse has an equal spectral width $(0.375/\tau_2)\sqrt{1+a_2^2} = 0.375/\tau_1 = \Delta\nu$. This is not surprising since the chirp filter is a phase filter that does not alter the spectral intensity of the original pulse. The invariance of the spectral width may also be viewed as follows: The temporal width of the pulse is expanded by a factor $\sqrt{1+a_2^2}$, so that the associated spectral width must be compressed by the same factor. However, because the filtered pulse is chirped this is accompanied by a spectral broadening by the very same factor, resulting in an unchanged spectral width.

The dependence of the pulse broadening ratio τ_2/τ_1 and the chirp parameter a_2 on the ratio b/τ_1^2 is illustrated in Fig. 23.2-4.

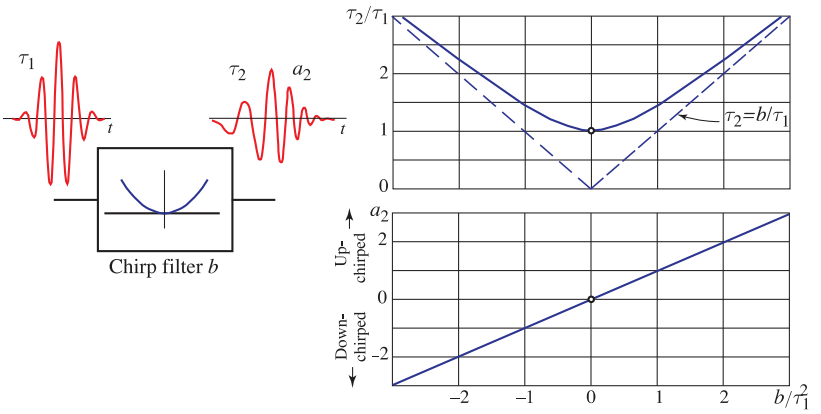


Figure 23.2-4 A chirp filter with coefficient b converts an unchirped Gaussian pulse of width τ_1, marked by an open circle, into a chirped Gaussian pulse of width τ_2 and chirp parameter a_2. The pulse width increases as $|b|$ increases, and is greater for smaller τ_1 . The chirp parameter is directly proportional to b and is greater for smaller τ_1.

Chirp Filtering of a Chirped Gaussian Pulse

When a chirped Gaussian pulse is transmitted through a chirp filter, the outcome is also a chirped Gaussian pulse, with altered parameters. The pulse will be either expanded or compressed and its chirp parameter will be altered, and may under certain conditions diminish to zero so that the new pulse may become unchirped (transform limited). This compression property offers a technique for generation of picosecond and femtosecond optical pulses, as will be shown in subsequent sections.

If the original pulse has width τ_1, chirp parameter a_1, and complex envelope $\mathcal{A}_1(t) = A_{10}\exp[-(1-ja_1)t^2/\tau_1^2]$, then upon filtering with a chirp filter $H_e(f) = \exp\left(-jb\pi^2 f^2\right)$, the result is a chirped Gaussian pulse $\mathcal{A}_2(t) = A_{20}\exp[-(1-ja_2)t^2/\tau_2^2]$, where

$$\frac{\tau_2^2}{1-ja_2} = \frac{\tau_1^2}{1-ja_1} + jb. \tag{23.2-18}$$

Equating the real and imaginary parts of (23.2-18) leads to the following expressions for the width τ_2 and chirp parameter a_2:

$$\tau_2 = \tau_1\sqrt{1 + 2a_1\frac{b}{\tau_1^2} + (1+a_1^2)\frac{b^2}{\tau_1^4}}\,, \tag{23.2-19}$$

$$a_2 = a_1 + (1+a_1^2)\frac{b}{\tau_1^2}\,. \tag{23.2-20}$$

A sketch of the dependence of the pulse broadening ratio τ_2/τ_1 and the chirp parameter a_2 on the ratio b/τ_1^2 is shown in Fig. 23.2-5. To determine the value $b_{\min}$ of the filter's chirp parameter at which the filtered pulse has its minimum width τ_0, we equate to zero the derivative of τ_2 in (23.2-19) with respect to b. The result is

$$\text{Minimum width} \quad \tau_0 = \frac{\tau_1}{\sqrt{1+a_1^2}}\,, \tag{23.2-21}$$

$$\text{Chirp coefficient} \quad b_{\min} = -a_1\tau_0^2 = -\frac{a_1}{1+a_1^2}\tau_1^2. \tag{23.2-22}$$

Using (23.2-21) and (23.2-22) we rewrite (23.2-19) and (23.2-20) in terms of $b_{\min}$ and τ_0 as follows:

$$\text{Width} \quad \tau_2 = \tau_0\sqrt{1 + (b - b_{\min})^2/\tau_0^4}\,, \tag{23.2-23}$$

$$\text{Chirp parameter} \quad a_2 = (b - b_{\min})/\tau_0^2. \tag{23.2-24}$$

When $b = b_{\min}$, (23.2-23) and (23.2-24) give $\tau_2 = \tau_0$ and $a_2 = 0$, so that the pulse is both maximally compressed and unchirped. Based on (23.2-22), if the original pulse is up-chirped ($a_1 > 0$), then $b_{\min} < 0$, so that a down-chirping filter is necessary for maximal compression. If the original pulse is unchirped ($a_1 = 0$), no chirp filter can compress it further, since it is already at its minimum width ($b_{\min} = 0$ and $\tau_0 = \tau_1$).

Note that (23.2-23) and (23.2-24) are identical to (23.2-15) and (23.2-16), which were derived for the initially unchirped pulse, except that b is replaced by $b - b_{\min}$. Thus, the graphs in Fig. 23.2-4 also apply to the case of an initially chirped pulse except for a shift in the horizontal direction by the value $b_{\min}$, as determined from (23.2-22).

EXAMPLE 23.2-1. ***Compression/Expansion of a Chirped Pulse Using a Chirp Filter.***

(a) A Gaussian pulse of width τ_1 and negative chirp parameter $a_1 = -1$ is filtered by a chirp filter of coefficient b. The filtered pulse is also Gaussian and has width τ_2 and chirp parameter a_2. In this case, the filtered pulse becomes maximally compressed and unchirped when $b = b_{\min} = \frac{1}{2}\tau_1^2$ and the compression factor is $\sqrt{1 + a_1^2} = \sqrt{2}$, so that the compressed pulse width $\tau_0 = \tau_1/\sqrt{2}$. The normalized pulse width τ_2/τ_0 is plotted in Fig. 23.2-5(a) versus the ratio b/τ_0^2. For small positive values of b, the pulse is compressed and acquires positive chirp. It becomes maximally compressed (and unchirped) when $b/\tau_0^2 = 1$ (i.e., $b/\tau_1^2 = 0.5$). As b increases further, the pulse is expanded. For negative b, the pulse is expanded and acquires additional down-chirp.

(b) An initially up-chirped pulse with chirp parameter $a_1 = 1$ is expanded with the application of an up-chirping filter ($b > 0$); its chirp parameter $a_2 > 1$. Application of a down-chirping filter ($b < 0$) results in compression. Maximal compression is achieved at $b/\tau_0^2 = -1$ (or $b/\tau_1^2 = -0.5$), as illustrated in Fig. 23.2-5(b).

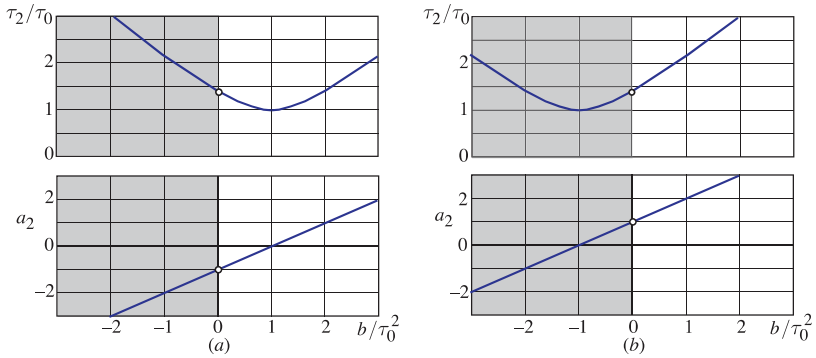


Figure 23.2-5 Filtering a Gaussian pulse of width τ_1 and chirp parameter a_1 with a chirp filter of chirp coefficient b, which is positive/negative in the unshaded/shaded areas. Two values of the original chirp parameter are considered: (a) $a_1 = -1$, and (b) $a_1 = 1$. The filtered pulse has width τ_2 and chirp parameter a_2. Parameters of the original pulse ($b = 0$) are indicated by open circles. The minimum pulse width $\tau_0 = \tau_1/\sqrt{1 + a_1^2}$ is used for normalization. The upper graphs show the dependence of the normalized pulse width on the ratio b/τ_0^2. The lower graphs show the dependence of the chirp parameter a_2 on b/τ_0^2.

EXAMPLE 23.2-2. ***Chirped-Pulse Amplification (CPA).*** The amplification of an ultrashort high-peak-power optical pulse is often limited by nonlinear effects such as saturation and self-focusing in the optical amplifier. Such limitations may be alleviated if the pulse is stretched by use of a chirp filter prior to amplification, and compressed by filtering through a second chirp filter after it has been amplified, as illustrated in Fig. 23.2-6. The first filter lowers the peak power by stretching the pulse, while maintaining its total energy. The second chirp filter, which has a chirp parameter of equal magnitude and opposite sign, compresses the pulse back to its original width. The amplification process is distributed over a longer time duration so that the peak power does not exceed the amplifier limits.

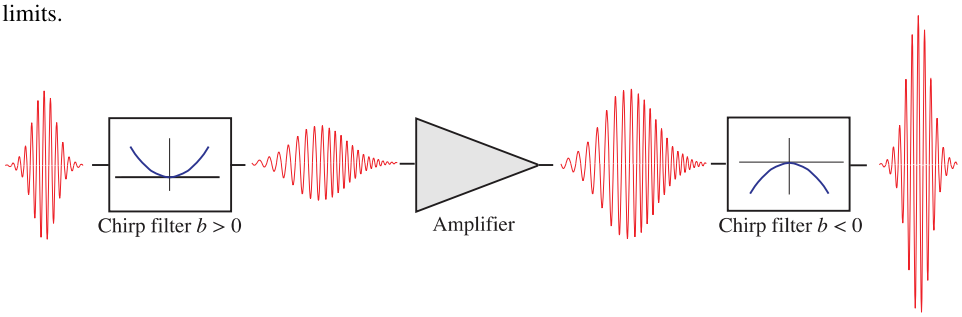


Figure 23.2-6 Chirped-pulse amplifier.

EXAMPLE 23.2-3. ***Ultrafast, Petawatt, High-Repetition-Rate Laser Using CPA.*** The HAPLS laser system[†] makes use of chirped-pulse amplification to generate ultrafast, high-power pulses at a central wavelength of 800 nm and a repetition rate of 10 pulses/s. Each pulse delivers an energy of 30 J in a duration of 30 fs, corresponding to a peak power of 1 PW. The 17-m-long system comprises a frequency-doubled, diode-pumped Nd^{3+}:glass laser, which in turn serves as the pump for a Ti^{3+}:sapphire femtosecond laser (Sec. 16.3) that incorporates a chirped-pulse amplifier (Example 23.2-2). The power-amplifier portion of the pump laser contains a collection of Nd^{3+}:glass laser-amplifier slabs similar to those used at the National Ignition Facility (Sec. 15.3B). The slabs are pumped by four AlGaAs laser-diode stacks (Sec. 18.4A) (comprising $> 500\,000$ laser diodes) that collectively deliver 1000-J pulses of 0.3-ms duration, at a wavelength of 888 nm and at a repetition rate of 10 pulses/s, with a peak power of 3.2 MW and an average power of 10 kW. This high repetition rate is enabled by laser-diode pumping, which is a factor of 20 more efficient than flashlamp pumping and thus reduces heating considerably. The Nd^{3+}:glass pump laser produces 200-J pulses at a repetition rate of 10 pulses/s, with an average power of 2 kW, at a wavelength of 1.053 μm. After passage through a second-harmonic-generation (SHG) module (Sec. 22.2A), which halves the wavelength to 526.5 nm, the pulse train pumps the Ti^{3+}:sapphire femtosecond laser that incorporates chirped-pulse amplification using a diffraction-grating chirp filter for pulse compression (Example 23.2-5). The net result is that the 1000-J, 3.2-MW pulses of 0.3-ms duration delivered by the laser-diode stacks are converted into 30-J, 1-PW pulses of 30-fs duration, at the same repetition rate of 10 pulses/s. The 10-kW average optical power generated by the laser-diode stacks is reduced to 300 W at the output of the HAPLS, providing an optical-to-optical efficiency $\eta_o \approx 3\%$. Since the HAPLS consumes 150 kW of electrical power, its wall-plug efficiency $\eta_e \approx 0.2\%$. Multi-petawatt systems generally fall into two classes: 1) those with pulse durations < 50 fs that are principally based on Ti^{3+}-doped sapphire lasers; and 2) those with pulse durations > 100 fs that are principally based on Nd^{3+}- and Yb^{3+}-doped glass or crystals (see, e.g., Example 15.3-1).

B. Implementations of Chirp Filters

Chip filters are implemented by use of dispersive optical systems. The following are some of the various origins of dispersion in optical components (see also Sec. 10.3B):

[†] The *High-Repetition-Rate Advanced Petawatt Laser System* (HAPLS), developed at LLNL, is located at the *European Extreme Light Infrastructure* (ELI) *Beamlines* facility in Dolní Břežany, Czech Republic.

- *Material dispersion* results from the frequency/wavelength dependence of the index of refraction and/or absorption coefficient of optical materials.
- *Spatial dispersion* takes a variety of forms:
  - *Angular dispersion* has its origin at the frequency/wavelength dependence of the deflection angle of certain optical components. This is most pronounced in diffractive optical elements such as diffraction gratings and holographic optical elements. Refractive elements such as prisms exhibit angular dispersion as a result of their material dispersion.
  - *Multipath dispersion* is associated with the existence of multiple paths with different optical pathlengths. An example is *modal dispersion* in optical waveguides, which results from the different propagation constants of the waveguide modes (Sec. 9.2).
  - Optical systems dominated by interferometric effects are wavelength dependent and therefore exhibit *interferometric dispersion*. For example, stratified media and periodic structures such as Bragg gratings have frequency-dependent reflectance and transmittance. Optical resonators have strong frequency selectivity, and are therefore highly dispersive.
  - Likewise, diffraction from small apertures is wavelength dependent and can therefore be responsible for significant changes in the profiles of short optical pulses; this is a form of *diffractive dispersion*. In general, propagation through, or scattering from, spatial structures or inhomogeneities of size comparable to a wavelength contribute to this type of dispersion. Even single-mode waveguides exhibit *waveguide dispersion*, which is associated with the confinement of light in small structures (see Sec. 9.2).
- *Polarization dispersion* is a result of the wavelength dependence of the anisotropic properties of optical materials, components, and systems.
- *Nonlinear dispersion* also plays an important role in the reshaping of intense optical pulses, because of the wavelength dependence of nonlinear optical effects such as self-phase modulation and parametric interactions governed by frequency-dependent energy conservation and phase-matching conditions.

Any of these dispersive effects may be used to implement the chirp filter, as demonstrated by the following examples.

Angular-Dispersion Chirp Filters

Optical elements that introduce angular dispersion, such as prisms and diffraction gratings, may function as chirp filters. A generic element with such behavior, illustrated schematically in Fig. 23.2-7(a), disperses the monochromatic components that constitute a pulsed plane wave into different directions. Assume that the component with frequency ν is directed at an angle $\theta(\nu)$ measured from the direction of the component at the central frequency ν_0, i.e., $\theta(\nu_0) = 0$. If ℓ_0 is the optical pathlength of the central-frequency component, then the optical pathlength of the component at frequency ν is $\ell_0 \cos\theta(\nu)$, as can be seen from Fig. 23.2-7(a). The phase shift encountered by the spectral component ν is

$$\Psi(\nu) = \frac{2\pi\nu}{c}\ell_0 \cos\theta(\nu), \tag{23.2-25}$$

and the corresponding phase filter has a transfer function $H(\nu) = \exp[-j\Psi(\nu)]$.

A pulsed beam is typically filtered by use of four identical dispersive elements arranged as shown in Fig. 23.2-7(b). One element separates the spectral components of the optical pulse into separate directions. A second inverted element brings back the rays into parallelism, as illustrated in the left block of Fig. 23.2-7(b). The process is reversed by two identical elements in the reverse order, as illustrated in the right block

of the figure. The overall system is a phase filter with $\Psi(\nu) = (2\pi\nu/c)\ell_0 \cos\theta(\nu)$, where ℓ_0 is the overall optical pathlength of the central-frequency component.

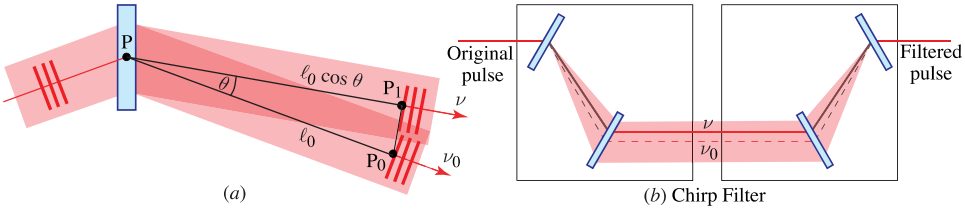


Figure 23.2-7 (a) An optical element exhibiting angular dispersion. The component at frequency ν is separated from that at the central frequency ν_0 by a deflection angle $\theta(\nu)$. At the observation point P_0, the pathlength of the central-frequency component is ℓ_0 (distance $\overline{PP_0}$). The pathlength of the component at frequency ν is the distance $\overline{PP_1}$, where P_1 is determined by lining up the wavefront to pass through the observation point P_0. Therefore, the distance $\overline{PP_1}$ in the triangle PP_1P_0 is $\ell_0 \cos\theta(\nu)$. (b) A chirp filter made with a combination of four of the elements in (a).

The function $\theta(\nu)$ depends on the dispersive element used, as will be shown in subsequent examples. Typically, $\theta(\nu)$ is sufficiently small so that $\cos\theta(\nu) \approx 1 - \frac{1}{2}\theta^2(\nu)$ and

$$\Psi(\nu) \approx \frac{2\pi\nu}{c}\ell_0 \left[1 - \tfrac{1}{2}\theta^2(\nu)\right]. \tag{23.2-26}$$

If $\theta(\nu)$ is slowly varying within the pulse spectral width, then it may be approximated by a few terms of a Taylor-series expansion about the central frequency ν_0. The derivatives of $\Psi(\nu)$ evaluated at $\nu = \nu_0$, where $\theta(\nu_0) = 0$, are:

$$\Psi' \approx \frac{2\pi}{c}\ell_0, \quad \Psi'' \approx -\frac{2\pi\nu}{c}\ell_0 \left(\frac{d\theta}{d\nu}\right)^2. \tag{23.2-27}$$

Based on (23.2-10) and (23.2-11), the filter is equivalent to a time delay $\tau_d = \ell_0/c$ and a chirp filter with chirp coefficient

$$\boxed{b \approx -\frac{\ell_0}{\pi\lambda_o}\alpha_\nu^2,} \tag{23.2-28}$$

Angular-Dispersion Chirp Coefficient

where $\alpha_\nu = d\theta/d\nu$ is the **angular dispersion coefficient**. Since b is always negative in this approximation, regardless of the sign of α_ν, such filters are always down-chirping. Higher-order terms of the series expansion of the phase do, of course, introduce additional pulse shaping effects.

EXAMPLE 23.2-4. ***Prism Chirp Filter.*** The angle of deflection $\theta_d(\nu)$ of a ray incident on a prism is a function of the refraction geometry and the refractive index $n(\nu)$ (see Fig. 23.2-8). Since $\theta(\nu) = \theta_d(\nu) - \theta_d(\nu_0)$ the angular dispersion coefficient $\alpha_\nu = d\theta/d\nu = (d\theta_d/dn)(dn/d\nu)$. Using the relations $dn/d\nu = -(\lambda_o/\nu_0)dn/d\lambda_o = (n - N)/\nu_0$, where $N = n - \lambda_o dn/d\lambda_o$ is the group index of the material (see Sec. 5.7), we obtain

$$\alpha_\nu = \frac{n - N}{\nu_0}\frac{d\theta_d}{dn}. \tag{23.2-29}$$

For a thin prism with apex angle α, the deflection angle $\theta_d = (n-1)\alpha$ [see (1.2-7)] so that $d\theta_d/dn = \alpha$ and

$$\alpha_\nu = \frac{n-N}{\nu_0}\alpha. \tag{23.2-30}$$

As an example, for BK7 glass at wavelength λ_o = 800 nm, n = 1.511 and N = 1.527. For a prism with $\alpha = 15^\circ$, $\alpha_\nu = -1.11 \times 10^{-17} = -0.011$ fs. For ℓ_0 = 1 m, the chirp coefficient given by (23.2-28) is $b = -\alpha^2(n-N)^2\ell_0\lambda_o/\pi c_o^2 = -5 \times 10^{-29}$ s$^2 \approx -$ (7.1 fs)^2. In accordance with (23.2-15) and (23.2-16), an unchirped pulse of width τ_1 = 5 fs transmitted through this device is broadened by a factor $(1 + b^2/\tau_1^4)^{1/2} \approx 2.23$ and becomes chirped with chirp parameter $a_2 = b/\tau_1^2 = 2$.

The thick-prism chirp filter is considered in Prob. 23.2-1.

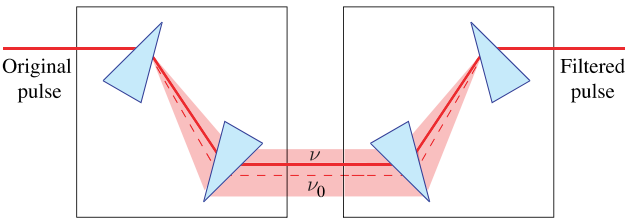


Figure 23.2-8 Prism chirp filter.

EXAMPLE 23.2-5. ***Diffraction-Grating Chirp Filter.*** In a diffraction grating system (Fig. 23.2-9) the angles of incidence and diffraction, θ_1 and θ_2, from a grating with period Λ are related by the diffraction condition (2.4-13). If $\theta_2 = \theta_{20} + \theta(\nu)$, where θ_{20} is the angle of the central-frequency component, then for first-order diffraction,

$$\sin\theta_1 + \sin[\theta_{20} + \theta(\nu)] = \frac{\lambda}{\Lambda} = \frac{c}{\nu\Lambda}. \tag{23.2-31}$$

Taking the derivatives of both sides at $\nu = \nu_0$, we obtain

$$\alpha_\nu = \frac{d\theta}{d\nu} = \frac{-c}{\nu_0^2\Lambda\cos\theta_{20}} = \frac{-\lambda_o^2}{c\Lambda\cos\theta_{20}}. \tag{23.2-32}$$

In the symmetrical case in which $\theta_1 = \theta_{20}$, $\sin\theta_{20} = \lambda_o/2\Lambda$, and therefore

$$\alpha_\nu = -\frac{1}{\nu_0}\frac{\lambda_o}{\sqrt{\Lambda^2 - (\lambda_o/2)^2}} \tag{23.2-33}$$

so that

$$b = -\frac{\lambda_o\ell_0}{\pi c^2}\frac{\lambda_o^2}{\Lambda^2 - (\lambda_o/2)^2}. \tag{23.2-34}$$

For λ_o = 800 nm and Λ = 1.6 μm, $\alpha_\nu = -2.72 \times 10^{-15}$ s $= -2.72$ fs. For $\ell_0 = 10$ cm, $b = -2.94 \times 10^{-25} = -(542$ fs$)^2$.

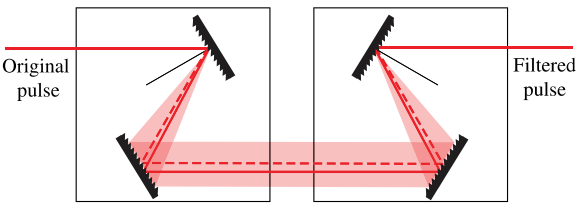


Figure 23.2-9 The diffraction grating as a down-chirping filter.

Bragg-Grating Chirp Filters

Variable-pitch (or chirped) Bragg gratings (Fig. 23.2-10) are often used as chirp filters. As described in Sec. 7.1C, a Bragg grating is a periodic structure that reflects optical waves selectively. A grating with period Λ reflects only waves with wavelength λ satisfying the Bragg condition $\Lambda = m\lambda/2$, where m is an integer; waves at other wavelengths are transmitted without change. The grating can therefore serve as a narrowband filter. If the grating has a pitch that varies with position, then each segment of the grating reflects the wave with a wavelength matching the local pitch. The reflected waves travel different distances depending on the location from which they are reflected, so that the system acts as a frequency-sensitive phase filter. If the frequency of the periodic structure varies linearly with distance, the grating is said to be linearly chirped, and it functions as a linear chirp filter.

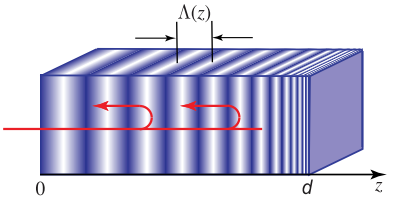


Figure 23.2-10 A Bragg grating with decreasing period serves as a positive chirp filter.

Assume that the period of a Bragg grating is a function $\Lambda(z)$ of the position z selected such that the frequency varies linearly with z, i.e., $\Lambda^{-1}(z) = \Lambda_o^{-1} + \xi z$ where Λ_o is the period at $z = 0$ and ξ is a constant. To determine the effect of the grating on an optical pulse, we decompose the pulse into its spectral components and examine the effect of the grating on each component. The component of frequency ν is reflected from the grating at the location z for which $\Lambda = m\lambda/2$, i.e., $\Lambda(z) = m\lambda/2 = mc/2\nu$ or $z = 2\nu/mc\xi - 1/\xi\Lambda_o$. That component travels a distance $2z$ and undergoes a phase shift $\Psi = (2\pi\nu/c)(2z)$ so that

$$\Psi = (8\pi/mc^2\xi)\nu^2 + (4\pi/c\xi\Lambda_o)\,\nu. \tag{23.2-35}$$

It follows from (23.2-10) and (23.2-11) that the chirped Bragg grating is equivalent to a time delay $\tau_d = 2/c\xi\Lambda_o$ and a chirp filter with chirp coefficient

$$\boxed{b = \frac{8}{m\pi\xi c^2}.} \tag{23.2-36}$$

Bragg-Grating Chirp Coefficient

If $\xi > 0$, i.e., the grating has an increasing frequency, as illustrated in Fig. 23.2-10, and the chirp coefficient $b > 0$, i.e., the filter is up-chirping. Likewise, a chirped Bragg grating with a decreasing frequency is a down-chirping filter.

C. Pulse Compression

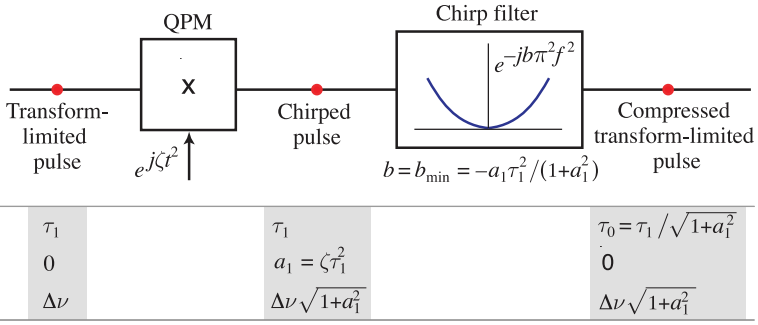
A transform-limited pulse cannot be compressed by use of a chirp filter alone. Such a filter introduces chirp accompanied by temporal broadening, but it does not alter the spectral width of the pulse. However, compression may be achieved by making use of a phase modulator followed by a chirp filter. The phase modulator *multiplies* the pulse by a time-dependent phase factor, which introduces chirp accompanied by spectral broadening, but it does not alter the temporal width of the pulse. The chirped pulse

is subsequently compressed by a chirp filter, which preserves the broadened spectral width while compressing the temporal width in the course of generating a transform-limited compressed pulse.

To compress an unchirped pulse $\mathcal{A}(t) = A_0 \exp(-t^2/\tau_1^2)$, we first convert it into a chirped pulse by *multiplication* with a quadratic phase factor $\exp(j\zeta t^2)$, where ζ is a constant, using a **quadratic phase modulator** (QPM). The result is a chirped pulse $\mathcal{A}_1(t) = A_{10} \exp[-(1 - ja_1)t^2/\tau_1^2]$ with chirp parameter

$$a_1 = \zeta \tau_1^2. \tag{23.2-37}$$

If $\zeta > 0$, the pulse becomes up-chirped, and subsequent filtering with a down-chirping filter can result in compression. Alternatively, if $\zeta < 0$, the pulse becomes down-chirped and subsequent filtering with an up-chirping filter can result in compression. In either case, the pulse is compressed by a factor $\sqrt{1 + a_1^2} = \sqrt{1 + \zeta^2 \tau_1^4}$. The system is illustrated in Fig. 23.2-11.



Pulse width:	τ_1	τ_1	$\tau_0 = \tau_1/\sqrt{1+a_1^2}$
Chirp parameter:	0	$a_1 = \zeta\tau_1^2$	0
Spectral width:	$\Delta\nu$	$\Delta\nu\sqrt{1+a_1^2}$	$\Delta\nu\sqrt{1+a_1^2}$

Figure 23.2-11 Compression of a transform-limited pulse by use of a quadratic phase modulator (QPM) followed by a chirp filter.

If the original pulse is a chirped pulse $A_1(t) = A_{10} \exp[-(1 - ja_1)t^2/\tau_1^2]$, then modulation by a quadratic phase $\exp(j\zeta t^2)$ converts it into another chirped pulse $A_2(t) = A_{10} \exp[-(1 - ja_2)t^2/\tau_1^2]$ with the same width but with an altered chirp parameter

$$\boxed{a_2 = a_1 + \zeta \tau_1^2.} \tag{23.2-38}$$

Effect of QPM on Chirp Parameter

Thus, a quadratic phase modulator for which the sign of ζ is opposite to that of a_1 may unchirp the initial pulse or even reverse its chirp sign.

Summary

The quadratic phase modulator (QPM) and the chirp filter serve dual functions. One operation is the Fourier-transform analog of the other:

QPM	= multiplication by a quadratic phase function:	Alters spectral width	Preserves temporal width
Chirp Filter	= convolution with a quadratic phase function:	Preserves spectral width	Alters temporal width

QPMs may be implemented by use of electro-optic modulators (see Sec. 21.1B), although the production of the appropriate signal $\exp(j\zeta t^2)$ is not simple. Passive phase modulation occurs when intense pulses are transmitted through nonlinear media exhibiting the optical Kerr effect, as will be described in Sec. 23.5C in connection with self-phase modulation, and this effect may be used to implement QPMs.

D. Pulse Shaping

The pulse-shaping methods discussed thus far are based on chirp filters implemented by dispersive optical components. Though chirp filters can be used for the purposes of pulse stretching and compression, they cannot be used to alter the optical pulse shape in an arbitrary manner, nor can they be used to generate optical pulses of prescribed shape, as is often desired in optical-communications and signal-processing applications. The general shaping of ultrashort pulses can, however, be achieved by making use of optical frequency-to-space mapping or time-to-space mapping, together with the use of spatial light modulators, as described in this section.

Frequency-to-Space Mapping

Frequency-to-space mapping of an optical pulse is achieved by means of a diffraction grating and a lens, which direct each constituent spectral component to a unique point in the lens's focal plane, as illustrated in the left side of Fig. 23.2-12. This system in effect projects the Fourier transform of the temporal profile of the pulse as a spatial pattern in the focal plane. A modulator modifies the magnitude and phase in accordance with the transfer function of the desired pulse-shaping linear filter. This is accomplished by use of a microlithographic or holographic mask, or a programmable spatial light modulator (SLM) (see Sec. 21.3B). The inverse operation of spatial-spectral mapping is subsequently implemented by a second lens and grating, which recombine the modified spectral components to form the reshaped pulse. This amounts to an inverse Fourier transform, and the overall operation is similar to spatial filtering in Fourier optics (see Sec. 4.4B). This technique has become an established tool for the general shaping of ultrashort pulses.

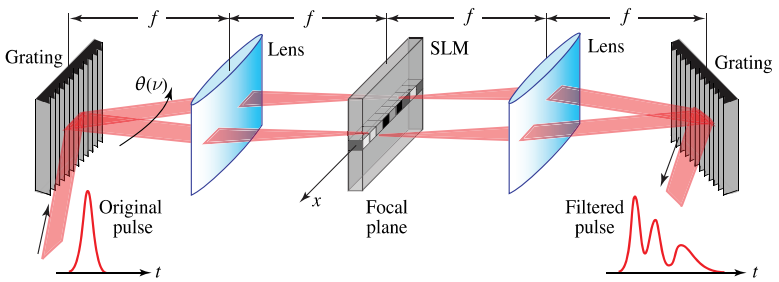


Figure 23.2-12 A system for pulse shaping includes: (1) frequency-to-space mapping — a grating and a lens display the Fourier transform of the pulse as a spatial pattern in the Fourier plane; (2) modulation by a spatial light modulator (SLM); and (3) space-to-frequency mapping using a lens and a grating generating the inverse Fourier transform.

The system depicted in Fig. 23.2-12 is described quantitatively as follows. If $\theta(\nu)$ is the deflection angle introduced by the grating at frequency ν, then the Fourier component at that frequency will be focused at a position $x = \theta(\nu)f$ in the lens focal plane (the Fourier plane), where f is the lens focal length and the angle is assumed to be small. A mask with amplitude transmittance $p(x)$ is therefore equivalent to a filter with transfer function $H(\nu) = p[\theta(\nu)f]$. If $\theta(\nu)$ is approximated by a linear function

of frequency, $\theta(\nu) \approx \alpha_\nu \nu$, where $\alpha_\nu = d\theta/d\nu$ is the angular dispersion coefficient of the grating [given by (23.2-32)], then the shape of the filter transfer function $H(\nu)$ is a scaled version of the profile of the mask function $p(x)$, i.e.,

$$H(\nu) = p(\alpha_\nu f \nu). \tag{23.2-39}$$

In this frequency-to-space mapping, the position x in the Fourier plane corresponds to the frequency $\nu = x/\alpha_\nu f$, and the spectral width $\Delta\nu$ extends over a width $X = \alpha_\nu f \Delta\nu$.

The preceding simplified analysis was based on the assumption that the original pulse is a plane wave, so that diffraction plays no role. For an original beam of finite width W in the plane of the grating, the spectral component at frequency ν is deflected by an angle $\theta(\nu) \approx \alpha_\nu \nu$, but has an angular spread proportional to $\lambda/W = c/\nu W$, which corresponds to a spatial spread $\delta x = f\lambda_0/W = cf/\nu W$. This frequency-dependent spread limits the spatial resolution of the system. A mask of total width X has approximately $M = X/\delta x = X/(\lambda_0 f/W)$ independent points, where λ_0 is the central wavelength. The spatial spread δx corresponds to a spectral spread $\delta\nu = (\lambda_0 f/W)/(\alpha_\nu f) = \lambda_0/(\alpha_\nu W)$. This limits the spectral resolution of the pulse filtering system to $M = XW/\lambda_0 f$ independent points.

The reshaping of picosecond and femtosecond pulses has been successfully demonstrated using a number of SLM technologies, including deformable mirrors, multielement liquid-crystal modulator arrays (millisecond to submillisecond response times, high duty cycle), acousto-optic deflectors (microsecond reprogramming, low duty cycle), and semiconductor optoelectronic modulator arrays (nanosecond reprogramming times).

Time-to-Space Mapping

Another configuration for arbitrary pulse shaping uses a spatial light modulator (SLM) butted against a diffraction grating and followed by a 2-f lens system with an on-axis pinhole in the Fourier plane, as illustrated in Fig. 23.2-13. The grating multiplies the spectral component of frequency ν, which has complex envelope $A_1(\nu)$, by the frequency-dependent and position-dependent phase factor $\exp(j2\pi\gamma\nu x)$, where γ is a constant. The SLM modulates it by a controllable spatial pattern $p(x)$, and the lens system functions as a spatial integrator producing an amplitude

$$A_2(\nu) \propto A_1(\nu) \int p(x) \exp(j2\pi\gamma\nu x)\, dx \propto A_1(\nu) P(-\gamma\nu), \tag{23.2-40}$$

where $P(\nu_x)$ is the spatial Fourier transform of $p(x)$. The overall system therefore acts as a linear system with transfer function $H(\nu) \propto P(-\gamma\nu)$, which corresponds to an impulse response function

$$h(t) \propto p\,(-t/\gamma). \tag{23.2-41}$$

It follows that the transmittance of the SLM at the position x controls the value of the impulse response function at one-and-only-one time $t = -\gamma x$. Thus, the system serves as a direct time-to-space mapping that may be exploited to reshape or synthesize a femtosecond pulse with arbitrary temporal profile.

23.3 PULSE PROPAGATION IN OPTICAL FIBERS

This section examines the propagation of an optical pulse in an extended linear dispersive medium, such as an optical fiber, by regarding the process of propagation as a

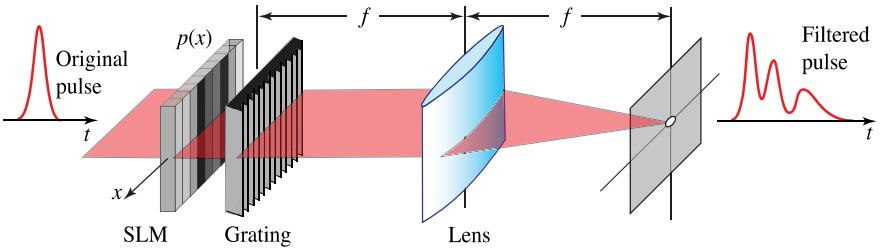


Figure 23.2-13 Pulse shaping based on time-to-space mapping. The system has an impulse response function $h(t)$ that is a scaled version of the SLM transmittance function $p(x)$.

linear filter with a transfer function governed by the frequency-dependent propagation constant. For optical pulses with a slowly varying envelope, such as picosecond pulses, the filter may be approximated by a combination of a time delay and a chirp filter, in which case the mathematics of pulse propagation is encompassed in the analysis provided in Sec. 23.2A. A differential equation describing the evolution of the pulse envelope as it travels through the medium is derived and an analogy between dispersion and ordinary optical diffraction is established.

A. The Optical Fiber as a Chirp Filter

The Dispersive Medium as a Filter

Upon propagation in a linear lossless dispersive medium, a monochromatic plane wave of frequency ν traveling a distance z in the z direction (Fig. 23.3-1) undergoes a phase shift $\beta(\nu)z$, where $\beta(\nu) = 2\pi\nu n(\nu)/c_o$ is the propagation constant, and $n(\nu)$ is the refractive index. Propagation is therefore mathematically equivalent to multiplication by the phase factor $\exp[-j\beta(\nu)z]$. Since a pulsed wave of wavefunction $U(z,t)$ is a superposition of many monochromatic waves, the phase factor $H(\nu) = \exp[-j\beta(\nu)z]$ is the transfer function of the linear system that represents propagation, i.e., $V(z,\nu) = H(\nu)V(0,\nu) = \exp[-j\beta(\nu)z]\,V(0,\nu)$, where $V(z,\nu)$ is the Fourier transform of $U(z,t)$.

For pulses with narrow spectral distribution, the complex wavefunction is written in terms of the complex envelope, $U(z,t) = \mathcal{A}(z,t)\exp(-j\beta_0 z)\,\exp(j2\pi\nu_0 t)$, where ν_0 is the central frequency and $\beta_0 = \beta(\nu_0)$. In the Fourier domain, this translates to $V(z,\nu) = A(z,\nu-\nu_0)\exp(-j\beta_0 z)$, and hence the relation $V(z,\nu) = \exp[-j\beta(\nu)z]\,V(0,\nu)$ becomes $A(z,\nu-\nu_0) = A(0,\nu-\nu_0)\exp(-j[\beta(\nu)-\beta(\nu_0)]z)$. In terms of the frequency difference $f = \nu - \nu_0$,

$$A(z,f) = H_e(f)A(0,f), \tag{23.3-1}$$

where

$$\boxed{H_e(f) = \exp\{-j[\beta(\nu_0+f)-\beta(\nu_0)]z\}} \tag{23.3-2}$$

Envelope Transfer Function

is the envelope transfer function.The effect of the dispersive medium on the pulse envelope is therefore modeled as a phase filter $H_e(f) = \exp[-j\Psi(f)]$ with phase $\Psi(f) = [\beta(\nu_0+f)-\beta(\nu_0)]z$.

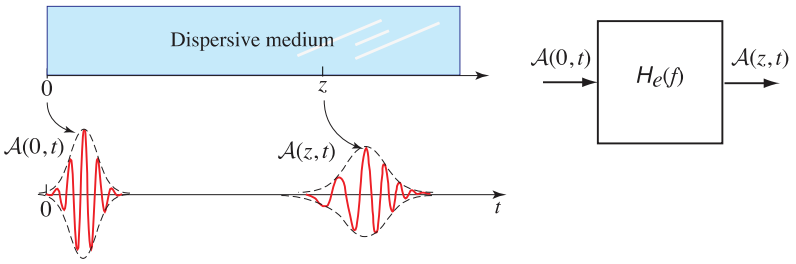


Figure 23.3-1 Transmission of an optical pulse through a dispersive medium is equivalent to a phase filter.

Approximation of a Dispersive Medium by Time Delay and Chirp Filter

If the propagation constant $\beta(\nu)$ varies slowly within the pulse spectral width, we may use the results of the Taylor-series expansion $\Psi(f) \approx \Psi' f + \frac{1}{2}\Psi'' f^2$, described in Sec. 23.2A, where Ψ' and Ψ'' are the first and second derivatives of $\Psi(\nu)$ with respect to ν at ν_0, and $\Psi(0) = 0$. The envelope transfer function can then be approximated by

$$H_e(f) \approx \exp[-j(\Psi' f + \tfrac{1}{2}\Psi'' f^2)] = \exp(-j2\pi\tau_d f)\exp(-jb\pi^2 f^2), \qquad (23.3\text{-}3)$$

where $\tau_d = \Psi'/2\pi$ and $b = \Psi''/2\pi^2$. It follows that the process of pulse propagation is equivalent to a combination of a time delay and a chirp filter.

The factor $\exp(-j2\pi\tau_d f)$ is a time delay $\tau_d = \Psi'/2\pi = (1/2\pi)(d\beta/d\nu)z = (d\beta/d\omega)z = z/v$, called the **group delay**, where

$$\boxed{v = 1/\beta' = \frac{c_o}{N}} \qquad \begin{matrix}(23.3\text{-}4)\\ \text{Group Velocity}\\ (\beta' = d\beta/d\omega)\end{matrix}$$

is the group velocity, and $N = n - \lambda_o dn/d\lambda_o$ is the group index. These parameters have been previously defined in the simplified analysis provided in Sec. 5.7.

The factor $\exp(-jb\pi^2 f^2)$ represents a chirp filter with chirp coefficient $b = \Psi''/2\pi^2 = (1/2\pi^2)(d^2\beta/d\nu^2)z = 2\beta'' z$, where $\beta'' = d^2\beta/d\omega^2$. The chirp coefficient is proportional to the distance z and is usually written in the form

$$\boxed{b = 2\beta'' z = \frac{D_\nu}{\pi} z,} \qquad \begin{matrix}(23.3\text{-}5)\\ \text{Chirp Coefficient}\end{matrix}$$

where

$$\boxed{D_\nu = 2\pi\beta'' = \frac{d}{d\nu}\left(\frac{1}{v}\right) = \frac{\lambda_o^3}{c_o^2}\frac{d^2 n}{d\lambda_o^2}} \qquad \begin{matrix}(23.3\text{-}6)\\ \text{GVD Coefficient}\\ (\beta'' = d^2\beta/d\omega^2)\end{matrix}$$

is the group velocity dispersion (GVD) coefficient. It is the derivative of the group delay per unit length with respect to the frequency ν, as described previously in Sec. 5.7. A medium with $\beta'' > 0$ (or $D_\nu > 0$) is said to have normal dispersion or positive

GVD, and it functions as an up-chirping chirp filter ($b > 0$). Conversely, a medium with $\beta'' < 0$ (or $D_\nu < 0$) is said to have anomalous dispersion or negative GVD, and it corresponds to a down-chirping filter ($b < 0$).

EXAMPLE 23.3-1. ***Adjustable Chirp Filter Using Combined Angular and Material Dispersion in a Prism.*** In Example 23.2-4, it was shown that when a pulsed beam is refracted by a prism, a chirping effect is introduced as a result of angular dispersion. In this example, we consider the effect of material dispersion, which was neglected in the previous example. If the central ray crossing the prism in Fig. 23.2-8 travels a distance L through the prism, then material dispersion amounts to a chirp filter with chirp coefficient $b = 2\beta''L = D_\nu L/\pi$ [see (23.3-5)]. For a prism made of BK7 glass at $\lambda = 800$ nm, the dispersion coefficient $D_\nu = 0.284 \times 10^{-24}$ s^2/m so that, for $L = 1$ m, the chirp coefficient $b = D_\nu L/\pi = +9 \times 10^{-26}$ s^2 $= +(300 \text{ fs})^2$. In Example 23.2-4, it was shown that the chirp coefficient arising from angular dispersion alone for a thin prism with 15° apex angle was $b \approx -(7.1 \text{ fs})^2$. The total chirp coefficient is the sum of the contributions of material and angular dispersion so the net value of b is positive. The distance L can be adjusted by moving the prism in a direction orthogonal to its base, as illustrated in Fig. 23.3-2.

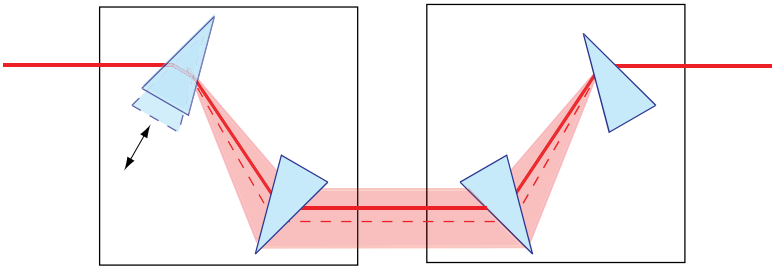


Figure 23.3-2 Prism chirp filter with adjustable chirp coefficient.

Summary

The propagation of a pulse in a dispersive medium may be approximated by two effects: a time delay associated with the group velocity $v = 1/\beta' = c_o/N$ and a chirp filter with chirp parameter $b = 2\beta''z = D_\nu z/\pi$ proportional to the propagation distance z. The parameters β' and β'' are the derivatives of the propagation constant β with respect to the angular frequency ω, and $D_\nu = 2\pi\beta''$ is the GVD coefficient.

B. Propagation of a Gaussian Pulse in an Optical Fiber

Since a linear dispersive medium may be approximated by a time delay and a chirp filter, the propagation of a Gaussian optical pulse in such a medium may be understood in terms of the general results presented in Sec. 23.2A, as discussed above.

Transform-Limited-Gaussian Input Pulse

Consider first a transform-limited (unchirped) Gaussian pulse of width τ_0 at $z = 0$. At a distance z the pulse is delayed by a time $\tau_d = z/v$ and is filtered by a chirp filter of chirp coefficient $b = D_\nu z/\pi$. In accordance with (23.2-15)–(23.2-17), the pulse remains Gaussian, but its width increases to $\tau(z) = \tau_0[1 + (D_\nu^2/\pi^2\tau_0^4)z^2]^{1/2}$, and it becomes chirped with chirp parameter $a(z) = (D_\nu/\pi\tau_0^2)z$ and amplitude $A_0(z) =$

$A_0[1 - j(D_\nu/\pi\tau_0^2)z]^{-1/2}$. By defining the parameter z_0 such that $D_\nu/\pi\tau_0^2 = 1/z_0$, these equations may be expressed in the simpler forms provided in Table 23.3-1, which also includes an expression for the complex envelope based on (23.1-12). The magnitude of z_0 is called the **dispersion length** and is a characteristic of the medium and the initial pulse width. The following observations emerge:

Table 23.3-1 Characteristics of a Gaussian pulse traveling through a dispersive medium with group velocity v, dispersion coefficient D_ν, and dispersion parameter z_0. At $z = 0$ the pulse is transform-limited with width τ_0, amplitude A_0, and intensity $I_0 = |A_0|^2$.

$\mathcal{A}(z,t) = A_0\sqrt{\dfrac{-jz_0}{z-jz}}\exp\left[j\dfrac{\pi}{D_\nu}\dfrac{(t-z/v)^2}{z-jz_0}\right]$	Complex envelope	(23.3-7)
$I(z,t) = I_0\dfrac{\tau_0}{\tau(z)}\exp\left[-2\dfrac{(t-z/v)^2}{\tau^2(z)}\right]$	Intensity	(23.3-8)
$\int I(t)dt = \sqrt{\pi/2}\, I_0\tau_0$	Energy density	(23.3-9)
$\tau(z) = \tau_0\sqrt{1+(z/z_0)^2}$	Pulse width	(23.3-10)
$a(z) = z/z_0$	Chirp parameter	(23.3-11)
$z_0 = \pi\dfrac{\tau_0^2}{D_\nu} = \dfrac{\tau_0^2}{2\beta''}$	Dispersion length $\lvert z_0\rvert$	(23.3-12)
$\Delta\nu = \dfrac{0.375}{\tau_0}$	Spectral width	(23.3-13)

- The pulse center is delayed by a time z/v; i.e., the pulse travels with the group velocity $v = 1/\beta'$.
- The width of the pulse $\tau(z)$ has its minimum value τ_0 at $z = 0$ and increases with increasing $|z|$, as illustrated in Fig. 23.3-3. At $z = |z_0|$ the pulse expands by a factor of $\sqrt{2}$, and at $z = \sqrt{3}\,|z_0|$ its width doubles. For $z \gg |z_0|$, $\tau(z) \approx \tau_0 z/|z_0| = (|D_\nu|/\pi\tau_0)z$; i.e., the pulse expands linearly at a rate inversely proportional to its initial pulse width, τ_0. In terms of the spectral width $\Delta\nu = 0.375/\tau_0$, the pulse width behaves as $\tau(z) \approx (1/0.375\pi)\,|D_\nu|\,\Delta\nu\, z = 0.85\,|D_\nu|\,\Delta\nu\, z$, which is consistent with the fact that D_ν is the pulse broadening rate per unit distance per unit spectral width (s/m-Hz). This relation may also be written in terms of the dispersion coefficient D_λ [ps/km-nm] as $\tau(z) \approx 0.85\,|D_\lambda|\,\Delta\lambda\, z$, which is an approximate version of (5.7-8).
- The chirp parameter $a(z)$ is 0 at $z = 0$, by definition, and increases linearly with the distance z, reaching a magnitude of unity at $z = |z_0|$, as illustrated in Fig. 23.3-3. The chirp sign is the same as the sign of D_ν. For normal dispersion, $D_\nu > 0$ and $a(z) > 0$ for $z > 0$, meaning that the pulse is up-chirped. In the visible region, normal dispersion means that "blue" is slower than "red," which is consistent with an up-chirped pulse. The opposite occurs for anomalous dispersion.
- The dispersion length $|z_0|$ depends on the magnitude of the medium dispersion coefficient D_ν and the initial pulse width τ_0. It is the distance at which the pulse width increases by a factor of $\sqrt{2}$ and the chirp parameter reaches a magnitude of unity.
- The spectral width $\Delta\nu = 0.375/\tau_0$ remains the same as the pulse travels. The spectral compression that accompanies temporal expansion of the pulse is fully compensated by an equal spectral broadening that accompanies chirping. This is

to be expected since propagation in the dispersive medium is modeled as a phase filter, which does not alter the spectral intensity.

- The energy density carried by the pulse is independent of z, as one would expect in a lossless medium.

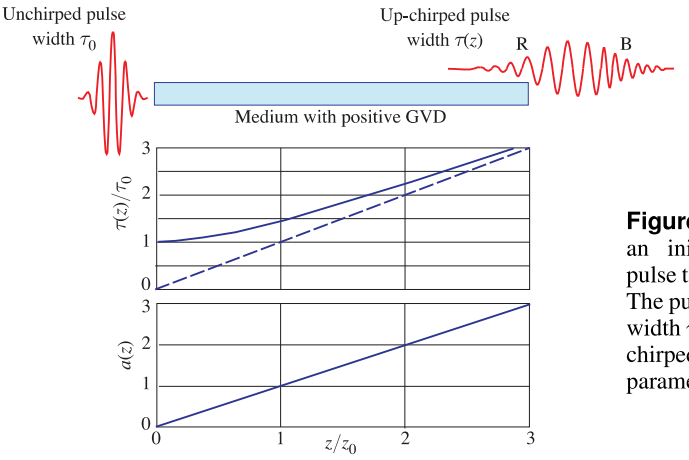


Figure 23.3-3 Propagation of an initially unchirped Gaussian pulse through a dispersive medium. The pulse remains Gaussian, but its width $\tau(z)$ expands, and it becomes chirped with an increasing chirp parameter $a(z)$.

EXAMPLE 23.3-2. ***Pulse Broadening in BK7 Glass.*** The dispersion coefficient of BK7 glass at λ= 620 nm is $\beta'' = 1.02 \times 10^{-25}$ s^2/m. For a slab of thickness 1 cm, this corresponds to a chirp coefficient $b = 2\beta'' z = 2.04 \times 10^{-23}$ s^2/m = (4.5 ps)^2. This means that when a Gaussian pulse of width 4.5 ps crosses the slab, its width expands by a factor of $\sqrt{2}$. For a shorter Gaussian pulse of time constant τ_0 = 100 fs and central wavelength λ_o= 620 nm, the dispersion length is $|z_0| = \tau_0^2/2|\beta''|$= 0.5 mm. The pulse doubles its width upon crossing a slab of thickness $\sqrt{3}\, z_0$ mm = 0.87 mm.

Chirped Gaussian Input Pulse

Based on (23.2-21), upon propagation through the dispersive medium a chirped Gaussian pulse of width τ_1 and chirp parameter a_1 at z = 0 reaches a minimum width

$$\tau_0 = \frac{\tau_1}{\sqrt{1 + a_1^2}} \tag{23.3-14}$$

Minimum Width

at a distance $z_{\min}$ for which $(D_\nu/\pi) z_{\min} = b_{\min}$. From (23.2-22),

$$z_{\min} = -a_1 \frac{\pi}{D_\nu} \tau_0^2, \tag{23.3-15}$$

which may be written in terms of the dispersion parameter $z_0 = \pi\tau_0^2/D_\nu$ as

$$z_{\min} = -a_1 z_0. \tag{23.3-16}$$

Location of Minimum Width

Finally (23.2-23) and (23.2-24) translate to the following expressions for the pulse width and chirp parameter as functions of the distance z,

$$\tau(z) = \tau_0\sqrt{1 + (z - z_{\min})^2/z_0^2}, \tag{23.3-17}$$

$$a(z) = (z - z_{\min})/z_0. \tag{23.3-18}$$

Equations (23.3-17) and (23.3-18) are identical to (23.3-10) and (23.3-11) for the initially unchirped case, except for a shift by a distance $z_{\min}$, which is the location of the minimum width.

The expressions in Table 23.3-1 are therefore universally valid for both positive or negative z and may be used for initially chirped pulses by placing the beginning of the medium at the location z corresponding to the matching value of the initial chirp parameter. This is illustrated in Fig. 23.3-4, which is another plot of $\tau(z)$ and $a(z)$ based on (23.3-10) and (23.3-11) for positive and negative values of z. As an example, for a medium with positive z_0 (positive GVD, or normal dispersion), when the initial chirp parameter is $a_1 = -1$, then $z_{\min} = z_0$, so that the medium begins at the position $z = -z_0$. The process of pulse compression and subsequent spreading is now clear. The pulse is maximally compressed by a factor of $\sqrt{2}$ and becomes unchirped at a distance $z_{\min} = z_0$. Upon further propagation through the medium the pulse is broadened and becomes up-chirped.

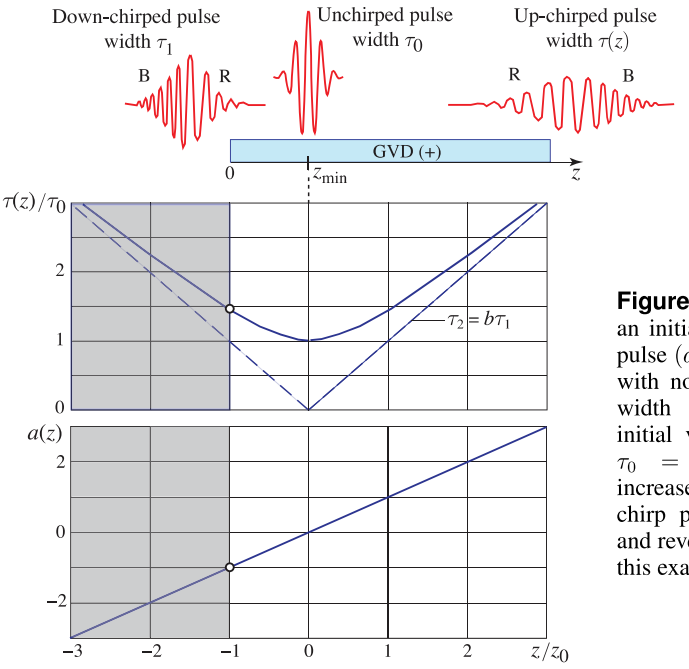


Figure 23.3-4 Propagation of an initially down-chirped Gaussian pulse ($a_1 = -1$) through a medium with normal dispersion. The pulse width $\tau(z)$ decreases from an initial value of τ_1 to a minimum $\tau_0 = \tau_1/\sqrt{2}$, and subsequently increases. The initially negative chirp parameter increases linearly and reverses sign when $z > z_{\min}$. In this example $z_{\min} = z_0$.

Since the initial chirp parameter a_1 and the dispersion coefficient D_ν may be positive or negative, we have a number of possibilities:

- For a medium with normal dispersion ($D_\nu > 0$) the filter is up-chirping and the parameter z_0 is positive. For an initially down-chirped pulse ($a_1 < 0$), $z_{\min}$ is positive so that the pulse is indeed compressed as it travels in the positive z direction. For an initially up-chirped pulse ($a_1 > 0$), $z_{\min}$ is negative and the pulse will not be compressed.
- For a medium with anomalous dispersion ($D_\nu < 0$) the filter is down-chirping and the parameter z_0 is negative. For an initially up-chirped pulse ($a_1 > 0$), $z_{\min}$ is positive so that the pulse is indeed compressed as it travels in the positive z direction. For an initially down-chirped pulse ($a_1 < 0$), $z_{\min}$ is negative, so that the pulse will not be compressed.

In summary, compression can occur if an up-chirped pulse travels through a down-chirping (anomalous) medium, or if a down-chirped pulse travels through an up-chirping (normal) medium.

Pulse Compression by Use of a QPM and a Dispersive Medium

As described in Sec. 23.2C, a transform-limited pulse may be compressed by use of a combination of a quadratic phase modulator (QPM) and a chirp filter. The chirp filter may be implemented by a dispersive medium, as illustrated in Fig. 23.3-5. If the width of the initial pulse is τ_1, then modulation by the phase factor $\exp(j\zeta t^2)$ is equivalent to a chirp coefficient $a_1 = \zeta\tau_1^2$. The spectral width of the chirped pulse is expanded by the factor $\sqrt{1+a_1^2}$. If ζ is negative, the pulse is down-chirped, and subsequent travel through a medium with positive GVD (normal dispersion) results in pulse compression to a minimum width

$$\tau_0 = \frac{\tau_1}{\sqrt{1+a_1^2}} = \frac{\tau_1}{\sqrt{1+\zeta^2\tau_1^4}}. \tag{23.3-19}$$

The pulse will also be compressed if ζ is positive and the medium has negative GVD. Using (23.3-16) and (23.3-19), we conclude that the minimum width occurs at a distance

$$z_{\min} = -a_1 z_0 = -\frac{\pi\tau_0^2}{D_\nu}a_1 = -\frac{\pi}{D_\nu}\frac{\tau_1^2 a_1}{1+a_1^2} = -\frac{\pi\zeta}{D_\nu}\frac{\tau_1^4}{1+\zeta^2\tau_1^4}, \tag{23.3-20}$$

which is positive if ζ and D_ν have opposite signs.

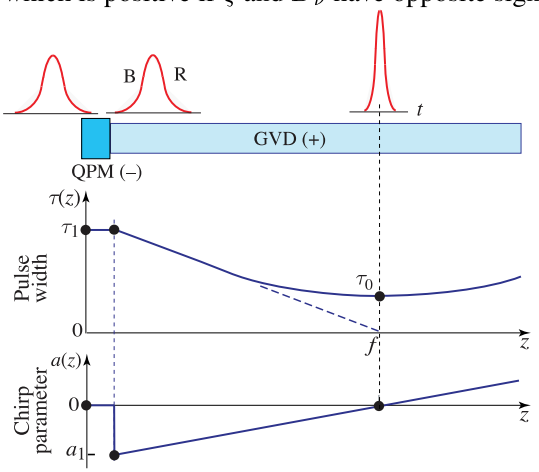


Figure 23.3-5 Pulse compression by a quadratic phase modulator (QPM) and a medium with group velocity dispersion (GVD).

In the limit when $a_1 = \zeta\tau_1^2 \gg 1$,

$$\tau_0 \approx \frac{\tau_1}{a_1} = \frac{1}{\zeta\tau_1} \tag{23.3-21}$$

and $z_{\min} \approx f$, where

$$f = \frac{\pi}{-\zeta D_\nu}. \tag{23.3-22}$$

This distance may be regarded as the **focal length** of this pulse focusing system.

EXAMPLE 23.3-3. ***Pulse Compression in a Silica-Glass Optical Fiber.***

(a) A Gaussian pulse of time constant τ = 100 fs and central wavelength λ_0= 850 nm (generated, e.g., by a Ti-sapphire laser) travels through a silica-glass optical fiber. At this wavelength, silica glass has normal dispersion (positive GVD) with $D_\lambda = -200$ ps/km-nm (see Fig. 5.7-5), corresponding to $D_\nu = -(\lambda_o^2/c_o)D_\lambda = +4.82 \times 10^{-25}$ s^2/m. If the pulse is initially unchirped, then τ_0 = 100 fs and therefore the dispersion length is $z_0 = \pi\tau_0^2/D_\nu = 6.52$ cm. At this distance the pulse expands by a factor of $\sqrt{2}$ and has a chirp coefficient $a = 1$. At a distance $z = 6.52$ m, the pulse width increases by a factor of approximately $z/z_0 = 100$, becoming 10 ps and the chirp parameter $a = 100$.

(b) If the initial pulse is phase modulated by a factor $\exp(j\zeta t^2)$, then $a_1 = \zeta\tau^2$. For $\zeta = -10^{-2}$ fs^{-2} the pulse becomes down-chirped with parameter $a = -1$. Upon subsequent propagation through the fiber, the initial 100-fs pulse is compressed to τ_0 =100/$\sqrt{2}$ = 77 fs at a distance $z_{\min} = -a_1 z_0 = \pi\tau_0^2/D_\nu = 3.26$ cm. Since the pulse is now narrower, it expands more rapidly upon further propagation through the fiber. At the distance z = 6.52 m, the width increases by a factor of approximately $z/z_0 \approx 200$, reaching a width of 14.1 ps.

EXERCISE 23.3-1

Dispersion Compensation in Optical Fibers. Pulse broadening in an optical fiber may be eliminated by balancing normal and anomalous dispersion.

(a) An unchirped pulse of central wavelength $\lambda_0 = 1.55$ μm and width $\tau_1 = 10$ ps is transmitted through a silica-glass optical fiber. At this wavelength, silica glass has anomalous dispersion with D_λ = +20 ps/km-nm. Determine the pulse width τ and chirp parameter a at a distance $d_1 = 100$ km.

(b) If the pulse is to be compressed back to the original width of 10 ps by use of another fiber of length d_2 (see Fig. 23.3-6) made of some material exhibiting normal dispersion with $D_\lambda = -100$ ps/km-nm, determine d_2.

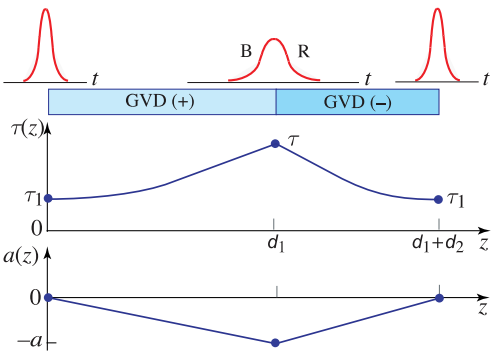


Figure 23.3-6 Dispersion compensation in optical fibers.

EXERCISE 23.3-2

Dispersion Compensation by Use of a Periodic Sequence of Phase Modulators. Pulse broadening in an optical fiber may be reduced by use of a periodic set of quadratic phase modulators spaced at a distance $2d$. Each modulator introduces a quadratic phase $\exp(j\zeta t^2)$. If the dispersion coefficient ζ is positive and the fiber material has negative GVD, then the pulse width and chirp parameter increase and decrease periodically as illustrated in Fig. 23.3-7. Show that the condition for this periodic pattern is

$$\zeta = -\frac{2a}{\tau^2} = -\frac{2d/z_0}{\tau_0^2\,[1+(d/z_0)^2]} = -\frac{2D_\nu d}{\pi\tau_0^4\,[1+(D_\nu d/\pi\tau_0^2)^2]}, \tag{23.3-23}$$

where τ_0 and τ are the minimum and maximum pulse widths, a is the chirp parameter, and $z_0 = \pi\tau_0^2/D_\nu$.

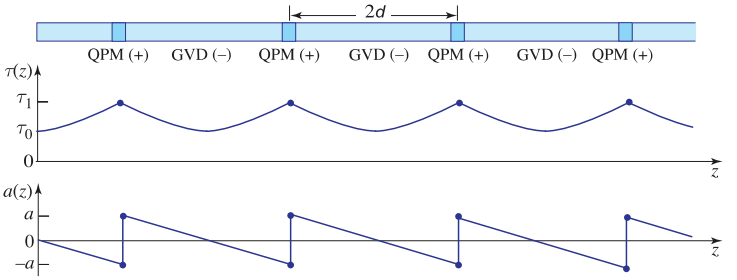


Figure 23.3-7 Dispersion compensation by use of periodic positive QPM and negative GVD.

*C. Slowly Varying Envelope Diffusion Equation

It was shown in Sec. 23.3A that a dispersive medium with propagation constant approximated by a Taylor-series expansion up to the quadratic term is equivalent to a pulse-envelope filter with transfer function $H_e(f) = \exp(-j2\pi zf/v)\exp(-j\pi D_\nu z f^2)$, where v is the group velocity and D_ν is the dispersion coefficient. We now demonstrate that under such conditions the envelope $\mathcal{A}(z,\,t)$ satisfies the partial differential equation

$$\frac{\partial^2\mathcal{A}}{\partial t^2} + j\frac{4\pi}{D_\nu}\left(\frac{\partial\mathcal{A}}{\partial z} + \frac{1}{v}\frac{\partial\mathcal{A}}{\partial t}\right) = 0. \tag{23.3-24}$$

If the time delay z/v is ignored (or a coordinate system moving with the pulse velocity v is used), then (23.3-24) simplifies to

$$\boxed{\frac{\partial^2\mathcal{A}}{\partial t^2} + j\frac{4\pi}{D_\nu}\frac{\partial\mathcal{A}}{\partial z} = 0,} \tag{23.3-25}$$

SVE Diffusion Equation

which is recognized as the **diffusion equation**.

□ **Proof of the SVE Diffusion Equation in a Dispersive Medium.** The proof begins with the filter equation $A(z,f) = A(0,f)H_e(z,f)$ from which $A(z,f) \approx A(0,f)\exp[-j2\pi(z/v)f - j\pi D_\nu z f^2]$, where $A(z,f)$ is the Fourier transform of $\mathcal{A}(z,t)$. Taking the derivative with respect to z we obtain the differential equation $(d/dz)A(z,f) \approx [-j2\pi f/v - j\pi D_\nu f^2]A(z,f)$. Forming the inverse Fourier transform of both sides with respect to f, and noting that the inverse Fourier transforms of $A(z,f)$, $j2\pi fA(z,f)$, and $(j2\pi f)^2A(z,f)$ are $\mathcal{A}(z,t)$, $\partial\mathcal{A}(z,t)/\partial t$, and $\partial^2\mathcal{A}(z,t)/\partial t^2$, respectively, we obtain (23.3-24). ■

The impulse response function associated with the diffusion equation is

$$h_e(t) = \frac{1}{\sqrt{jD_\nu z}} \exp\left(j\frac{\pi t^2}{D_\nu z}\right), \tag{23.3-26}$$

which is identical to that of a chirp filter (23.2-7) with $b = D_\nu z/\pi$.

For an initial Gaussian distribution $\mathcal{A}(0,t) = A_0 \exp(-t^2/\tau_0^2)$, the diffusion equation is known to have a Gaussian solution, $\mathcal{A}(z,t) = A_0\sqrt{-jz_0/(z-jz_0)}$ $\exp\left[j(\pi/D_\nu)t^2/(z-jz_0)\right]$, where $z_0 = \pi\tau_0^2/D_\nu$. Accounting for the time delay, we replace t with $t - z/v$ and reproduce (23.3-7).

□ *__Derivation of the SVE Diffusion Equation from the Helmholtz Equation.__ Equation (23.3-24) may also be directly derived from the Helmholtz equation $[d^2/dz^2 + \beta^2(\nu)]V(z,\nu) = 0$. Since $U(z,t) = \mathcal{A}(z,t)\exp(-j\beta_0 z)\exp(j2\pi\nu_0 t)$, its Fourier transform is $V(z,\nu) = A(z,\nu-\nu_0)\exp(-j\beta_0 z)$ where $A(z,\nu)$ is the Fourier transform of $\mathcal{A}(z,t)$. Substituting $\nu = \nu_0 + f$, the Helmholtz equation yields $[d^2/dz^2 + \beta^2(\nu_0+f)][A(z,f)\exp(-j\beta_0 z)] = 0$. Using the SVE approximation $d^2/dz^2[A\exp(-j\beta_0 zt)] \approx [-j2\beta_0 dA/dz - \beta_0^2 A]\exp(-j\beta_0 z)$, Helmholtz equation becomes $-j2\beta_0 dA/dz + [\beta^2(\nu_0+f) - \beta_0^2]A = 0$. For weak dispersion, $\beta^2(\nu_0+f) - \beta_0^2 \approx 2\beta_0[\beta(\nu_0+f) - \beta_0]$. As before, we approximate the propagation constant $\beta(\nu)$ by a 3-term Taylor-series expansion $\beta(\nu_0+f) \approx \beta_0 + 2\pi f\beta' + 2\pi^2 f^2\beta''$, where $\beta_0 = \beta(\omega_0)$, $\beta' = d\beta/d\omega|_{\omega_0}$, and $\beta'' = d^2\beta/d\omega^2|_{\omega_0}$. With this, Helmholtz equation now becomes $-jdA/dz + [2\pi f\beta' + 2\pi^2 f^2\beta'']A = 0$. Performing an inverse Fourier transform and noting that the multipliers $j2\pi f$ and $-4\pi^2 f^2$ are equivalent to the derivatives $\partial/\partial t$ and $\partial^2/\partial t^2$, respectively, we obtain $-j\partial A/\partial z - j\beta'\partial A/\partial t - \frac{1}{2}\beta''\partial^2 A/\partial t^2 = 0$. Finally, substituting $\beta' = 1/v$ and $\beta'' = D_\nu/2\pi$, we obtain (23.3-24). ■

*D. Analogy Between Dispersion and Diffraction

A striking mathematical similarity is observed between the SVE diffusion equation $\partial^2\mathcal{A}/\partial t^2 + j(4\pi/D_\nu)\,\partial\mathcal{A}/\partial z = 0$, which describes the propagation of a pulse $\mathcal{A}(z,t)$ in a dispersive medium (in a frame moving with velocity v, and neglecting dispersion terms higher than the quadratic term), and the paraxial Helmholtz equation $\nabla_T^2 A - j(4\pi/\lambda)\,\partial A/\partial z = 0$, which describes the diffraction of an optical beam $A(x,y,z)$ through free space in the paraxial approximation. Both are diffusion equations (the former is 1D and the latter is 2D). This similarity indicates that the temporal spreading (dispersion) of a pulse as it travels through the dispersive medium obeys the same mathematical law that governs the spatial spreading (diffraction) of a beam in the transverse plane as it travels through free space, with time t playing the role of the transverse coordinate $\rho = (x,y)$ and the dispersion coefficient $-D_\nu$ playing the role of the wavelength λ. Various features of this analogy are summarized in Table 23.3-2.

The analogy between the dispersion coefficient $-D_\nu$ and the wavelength λ is appreciated more fully if time t is measured in units of distance traveled at the speed of light, ct. In these units $c^2 D_\nu$ has units of distance and its role in determining the scale of pulse dispersion is quantitatively similar to the role played by the wavelength in determining the scale of diffraction. For example, if $D_\nu = 10^{-23}$ s^2/m, then $c^2 D_\nu \approx 0.9$ μm, which is equivalent to 3 fs.

Another interesting analogy relates the role of a lens in altering the wavefront curvature and the role of a quadratic phase modulator (QPM) in chirping a pulse. A thin lens introduces multiplication by a phase factor $\exp(j\pi\rho^2/\lambda f)$ [see (2.4-9)], while a QPM introduces multiplication by a phase factor $\exp(j\zeta t^2)$ (see Sec. 23.2C). Writing $\exp(j\zeta t^2) = \exp[j\pi t^2/(-D_\nu f)]$, where $\zeta = -\pi D_\nu f$, we see that the QPM

Table 23.3-2 Comparison between diffraction in space (paraxial approximation) and dispersion in a dispersive medium (second-order approximation). The dispersion coefficient $-D_\nu$ in pulse dispersion plays the role of the wavelength λ in diffraction. The quadratic phase modulator (QPM) is analogous to a temporal lens.

Diffraction		**Dispersion**	
Complex envelope	$A(\rho, z)$	Complex envelope	$\mathcal{A}(z, t)$
Transverse coordinate	$\rho = \sqrt{x^2 + y^2}$	Time	t
Axial coordinate	z	Axial coordinate	z
Paraxial Helmholtz equation	$\nabla_T^2 A - j\frac{4\pi}{\lambda}\frac{\partial A}{\partial z} = 0$	SVE diffusion (moving frame)	$\frac{\partial^2 \mathcal{A}}{\partial t^2} + j\frac{4\pi}{D_\nu}\frac{\partial \mathcal{A}}{\partial z} = 0$
Wavelength	λ	Dispersion coefficient	$-D_\nu$
Impulse response function $h_e(\rho)$	$\frac{j}{\lambda z}\exp\left(-j\frac{\pi\rho^2}{\lambda z}\right)$	Impulse response function $h_e(t)$	$\frac{1}{\sqrt{jD_\nu z}}\exp\left(j\frac{\pi t^2}{D_\nu z}\right)$
Lens	$\exp(j\pi\rho^2/\lambda f)$	QPM	$\exp(j\zeta t^2)$
Focal length	f	Focal length	$f = \pi/(-D_\nu\zeta)$

is equivalent to a **time lens** that compresses the pulse to a minimum width at $z = f$, where $f = \pi/(-D_\nu\zeta)$ is a focal length, confirming (23.3-22).

The mathematical analogy between the temporal spreading of a Gaussian pulse in a dispersive medium (Sec. 23.3B) and the spatial spreading of a Gaussian beam in free space (Sec. 3.1B) is summarized in Table 23.3-3. The dispersion length z_0 is analogous to the diffraction length (Rayleigh range) z_0. Though the latter is always positive, the former is defined such that it is positive for normal dispersion and negative for anomalous dispersion. This explains the negative sign in the parameter z_0 that appears in the expression for the complex envelope of the Gaussian pulse.

Table 23.3-3 Comparison between the diffraction of a Gaussian beam in free space and the dispersion of a Gaussian pulse in a dispersive medium.

Gaussian Beam		**Gaussian Pulse**	
Beam width	$W(z) = W_0\sqrt{1 + (z/z_0)^2}$	Pulse width	$\tau(z) = \tau_0\sqrt{1 + (z/z_0)^2}$
Diffraction length	$z_0 = \pi W_0^2/\lambda$	Dispersion length	$\lvert z_0\rvert = \pi\tau_0^2/\lvert D_\nu\rvert$
Divergence half angle	$\theta_0 = \lambda/\pi W_0$	Spreading rate (s/m)	$\lvert D_\nu\rvert/\pi\tau_0$
Wavefront curvature	$\varphi'' = \frac{-k}{R(z)} = -\frac{2\pi}{\lambda}\frac{z}{z^2 + z_0^2}$	Chirping rate	$\varphi'' = \frac{2\pi}{D_\nu}\frac{z}{z^2 + z_0^2}$
Spatial chirp	$a(z) = \frac{W^2(z)}{2R(z)} = \frac{z}{z_0}$	Chirp parameter	$a(z) = \frac{1}{2}\varphi''\tau^2(z) = \frac{z}{z_0}$

Because of the mathematical analogy between spatial diffraction and temporal dispersion, and between the lens and the quadratic phase modulator (QPM), each conventional optical system comprising combinations of free space and lenses has an analogous temporal system comprising combinations of dispersive media and QPMs. Figure 23.3-8 provides a number of examples:

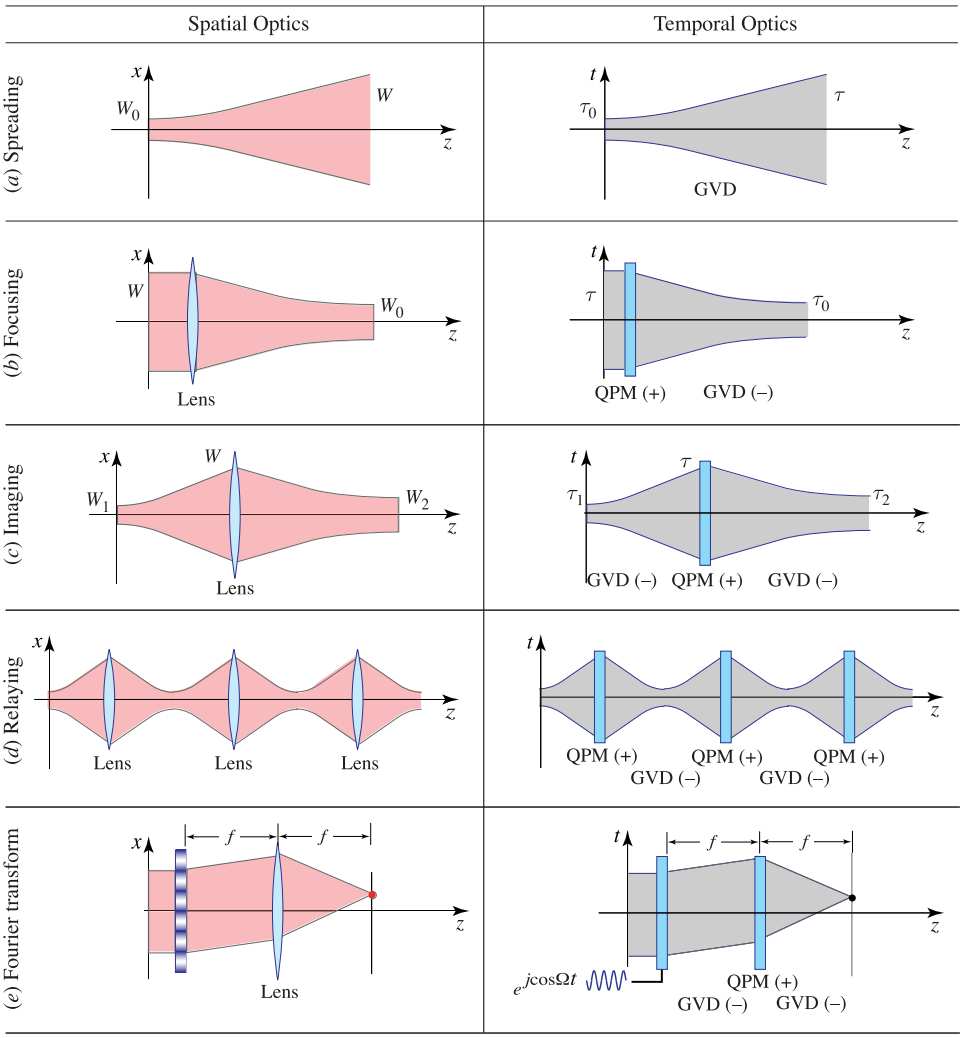


Figure 23.3-8 Analogy of spatial optics (left column) and temporal optics (right column). The quadratic phase modulator (QPM) plays the role of the lens. The shaded areas represent the spatial width of a wave (left) and the temporal width of a pulse (right) as functions of z. In the right column, time delays are ignored and only time spread is shown. The optical pulse (right) is assumed to travel in a medium with negative GVD. The figures in the right column are also applicable for a medium with positive GVD, but in this case the QPM must be negative.

- Temporal spreading of a pulse in a dispersive medium is analogous to spatial diffraction of a beam, or a wave transmitted through an aperture.
- Temporal compression of a pulse by a QPM is analogous to spatial focusing of a beam by a lens. For example, since a Gaussian beam is focused by a lens of focal

length f into a width $W_1 = (\lambda/\pi W_0)f$, it follows by analogy that a Gaussian pulse is compressed by a QPM into a temporal width $\tau_1 = 1/\zeta\tau_0 = (-D_\nu/\pi\tau_0)f$, where $f = \pi/(-D_\nu\zeta)$ is the focal length of the QPM. Another example of the time-focusing effect of the QPM is the focusing of two separated narrow pulses at $z = 0$ into one single pulse at $z = f$.

- The counterpart to single-lens imaging in conventional optics is a system using a QPM as a temporal lens that generates a magnified or minified replica of the pulse temporal profile, i.e., a temporal image (see Prob. 23.3-4).
- A periodic sequence of QPMs, designed to maintain the width of a pulse, is analogous to a periodic set of relaying lenses.
- The counterpart to a 2-f Fourier transform system (see Sec. 4.2) is a 2-f temporal Fourier transform system using a QPM. The system, for example, transforms a phase modulated optical pulse into an amplitude modulated pulse whose temporal profile is the Fourier transform of the original pulse.

One primary difference between spatial diffraction and temporal dispersion is that the wavelength λ is always positive, while its counterpart $-D_\nu$ may be positive or negative. The implication of this difference may be appreciated by examining the impulse response functions in Table 23.3-2. The positivity of the wavelength λ implies that a point of light must spread into a diverging phase front (a spherical wave). By analogy, in a medium with negative D_ν (i.e., positive $-D_\nu$), an impulse of light spreads into a down-chirped pulse. Conversely, in a medium with positive D_ν (i.e., normal dispersion), an impulse of light spreads into an up-chirped pulse. Both signs of chirp are permitted, whereas spatial diffraction admits only diverging waves.

23.4 ULTRAFAST LINEAR OPTICS

The spatial and temporal characteristics of pulsed waves are inherently coupled. Spatial spreading (or focusing) depends on the initial temporal profile, and the temporal pulse shape is influenced by the initial spatial pattern. These effects are particularly pronounced for ultranarrow optical pulses and for optical systems that exhibit angular dispersion. Only in a few special cases does a pulsed optical wave *exactly* maintain its temporal profile as it travels (examples are the plane wave and the spherical wave, as discussed in Sec. 23.1C). For optical pulses with a slowly varying envelope, the quasi-CW approximation is applicable, in which case the temporal and spatial changes are *approximately* decoupled. This approximation is not applicable for ultranarrow pulses, however. In this section we consider the propagation of ultranarrow pulsed beams in simple imaging systems. We begin with a simplified analysis based on ray optics and subsequently proceed to a theory based on wave optics using a Fourier-optics approach.

A. Ray Optics

Ray optics is based on the description of light by rays that are reflected and refracted at optical boundaries in accordance with Snell's law (Sec. 1.1). Temporal effects are accommodated in this theory since rays are assumed to travel with a medium-dependent velocity $c = c_o/n$. We used this theory in Sec. 10.3B to estimate the spreading of the time of arrival of optical rays inside an optical fiber by determining the time of travel for each of the optical paths and estimating the difference between the longest and shortest delays.

If some of the components of the optical system are dispersive, then the delay introduced by these components must be based on the group velocity $v = c_o/N$, rather than the phase velocity $c = c_o/n$, where $N = n - \lambda_o\, dn/d\lambda_o$ is the group index [see

(5.7-2)]. Estimating the broadening of an optical pulse as it travels through an optical system is therefore an exercise in determining the difference between the longest and shortest group delays for all possible optical paths.

Pulse Broadening in a Single-Lens Imaging System

In the single-lens imaging system illustrated in Fig. 23.4-1, an optical pulse is emitted at point P_1 in the form of multiple rays that meet at the conjugate point P_2. Each ray travels through air and the glass of the lens and is delayed accordingly. If the glass is nondispersive, then in accordance with Fermat's principle (Sec. 1.1) all rays arrive at the same time, and the pulse is not broadened. To account for the effect of dispersion, it is convenient to define the differential delay as the difference between the group delay (based on the group velocity v) and the phase delay (based on the phase velocity c). The difference between the longest and shortest differential delays then constitutes the pulse broadening. The differential delay is, of course, zero for the nondispersive portions of the optical path, so that attention need be directed only to the differential delay in the lens material.

Marking each ray by its position (x, y) in the plane of the lens, and denoting the lens thickness at position (x, y) as $d(x, y)$, the differential delay is written as

$$\tau(x,y) = \left|\frac{1}{c} - \frac{1}{v}\right| d(x,y) = \frac{|n-N|}{c_o}\, d(x,y)\,. \tag{23.4-1}$$

The width of the broadened pulse is the difference between the maximum and minimum values of $\tau(x, y)$, so that

$$\Delta\tau(x,y) = |n-N|\,\Delta d/c_o\,, \tag{23.4-2}$$

where Δd is the difference between the maximum and minimum widths of the lens. For a thin lens of focal length f and maximum thickness d_0, (2.4-8) and (2.4-10) provide $d(x,y) \approx d_0 - (x^2+y^2)/2R = d_0 - (x^2+y^2)/2(n-1)f$, with $(x^2+y^2) = (D/2)^2$, where D is the diameter of the lens aperture. Thus, $\Delta d = (D/2)^2/2(n-1)f$ and $\Delta\tau = [|n-N|/(n-1)](D/2)^2/2c_o f$, from which we have

$$\boxed{\Delta\tau = \frac{|n-N|}{n-1}\,\frac{1}{8F_\#^2}\,\frac{f}{c_o}\,,} \tag{23.4-3}$$

Pulse Spreading

where $F_\# = f/D$ is the lens F-number.

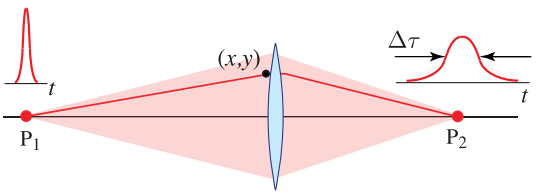


Figure 23.4-1 Pulse broadening in a single-lens imaging system resulting from material (chromatic) dispersion.

As an example, for a BK7-glass lens at $\lambda_o = 400$ nm, $n = 1.53$, and $n - N = \lambda_o\, dn/d\lambda_o = -0.052$. If $f = 30$ mm and $F_\# = 2$, the pulse spreads to a width $\Delta\tau \approx 307$ fs.

In this system, the pulse broadening is a result of the differential material dispersion associated with the multiple spatial paths of the rays. Without material dispersion, the existence of multiple paths would not result in pulse broadening, thanks to Fermat's principle.

*B. Wave and Fourier Optics

The wave nature of light dictates that a monochromatic narrow optical beam spreads into a wide cone with an angle directly proportional to the wavelength and inversely proportional to the original beam width. When the beam is modulated by an ultrashort pulse with a broad spectrum, each of its wavelength components spreads into its own cone, with the short-wavelength components occupying cones of smaller angles. Consequently, the spectral composition of the propagated light at each point in space is altered, with the points farther from the axis having less energy at the shorter wavelengths, as illustrated in Fig. 23.4-2. At off-axis points, the spectrum is therefore shifted to a lower central frequency (red shift) and the spectral width is reduced and accompanied by temporal broadening. This example demonstrates that the spatial and temporal characteristics of light are entwined through the very process of wave propagation, particularly when the beam is ultranarrow and the pulse is ultrashort.

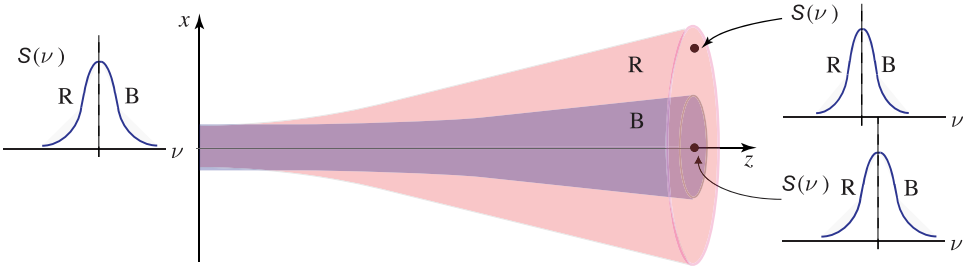


Figure 23.4-2 Spreading of a pulsed beam. The long-wavelength components (R) spread into cones with angles greater than those of the short-wavelength components (B). This results in the suppression of the short-wavelength components at off-axis points, and hence a red shift and a reduction of the spectral width accompanied by an increase in the pulse duration.

Though the propagation of ultrashort light pulses through arbitrary optical systems is complicated by the inherent space–time coupling, the analysis is conceptually simple when the system is linear since a Fourier approach can be used to reduce the problem to one of a superposition of solutions for each of the constituent monochromatic components. An arbitrary pulsed wave $U(\mathbf{r}, t)$ is decomposed into a sum of monochromatic components with amplitudes given by the Fourier transform $V(\mathbf{r}, \nu) = \int U(\mathbf{r}, t)\, \exp(-j2\pi\nu t)\, dt$. The propagation of each monochromatic component through the system is determined using the tools developed in Chapters 2–4, and the overall solution is subsequently obtained by superposition, i.e., by an inverse Fourier transform $U(\mathbf{r}, t) = \int V(\mathbf{r}, \nu)\, \exp(j2\pi\nu t)\, d\nu$.

Fourier Optics of Pulsed Waves

The propagation of monochromatic light between two parallel planes, denoted 1 and 2, with an arbitrary linear optical system sandwiched between, may be described by the linear transformation [see (B.2-1) of Appendix B and Chapter 4]:

$$V_2(x, y, \nu) = \iint h(x, x', y, y', \nu) V_1(x', y', \nu)\, dx' dy', \tag{23.4-4}$$

where h is the impulse response function of the system at frequency ν. For a pulsed input wavefunction $U_1(x, y, t)$, the output wavefunction $U_2(x, y, t)$ may be readily determined by computing its Fourier transform $V_2(x, y, \nu)$ via (23.4-4), and then computing the inverse Fourier transform.

The impulse response function h has been determined in Chapter 4 for various optical components. The results are reproduced here with the dependence on the frequency ν made explicit:

- **Free space**. In accordance with (4.1-18), which is valid in the Fresnel approximation, propagation through a distance z of free space is equivalent to a system with impulse response function

$$h(x, x', y, y', \nu) \approx \frac{j\nu}{cz} \exp\left[-j\frac{2\pi\nu}{c}\frac{(x-x')^2+(y-y')^2}{2z}\right]. \tag{23.4-5}$$

  We have ignored the factor $\exp(-j2\pi\nu z/c)$ that belongs to h_0 in (4.1-18) since it represents an inconsequential constant time delay z/c.

- **Aperture**. Transmission through a planar aperture is equivalent to multiplication by the aperture function (which is, in accordance with (4.3-2), unity within the aperture and zero outside it).

- **Lens**. Transmission through a thin double-convex lens of focal length f is, according to (2.4-9) and (2.4-11), equivalent to multiplication by the quadratic phase factor

$$\mathrm{t}(x, y, \nu) \approx \exp\left(-jn\frac{2\pi\nu}{c_o}\,d_0\right)\exp\left(j\frac{2\pi\nu}{c_o}\frac{\rho^2}{2f}\right), \tag{23.4-6}$$

  where $\rho = \sqrt{x^2+y^2}$ is the radial distance and the focal length f is given by

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{23.4-7}$$

  Here R_1 and R_2 are the two radii of the spherical lens. If the refractive index n of the lens material is wavelength dependent, then the focal length f depends on the optical frequency ν. Material dispersion thus results in chromatic aberration, which in turn contributes to the distortion of ultrashort optical pulses.

With the help of these equations, it is in principle possible to determine the space–time dependence of the output wave for any input pulsed wave transmitted through any system comprising combinations of free space, apertures, and lenses.

Optical Fourier-Transform System

Consider, for example, an optical system involving the propagation of a monochromatic wave between the front and back focal planes of a lens. Ignoring the y dependence for simplicity, this system is described by an impulse response function

$$h(x, x', \nu) \approx h_l \exp\left(j\frac{2\pi\nu}{cf}xx'\right), \tag{23.4-8}$$

where $h_l = (j\nu/cf)\exp(-j4\pi\nu f/c)$ is a constant [see (4.2-8)]; this corresponds to a spatial Fourier transform for monochromatic light, as described in Sec. 4.2. This system exhibits strong temporal-spatial coupling — the temporal waveform at a fixed point at the output plane is heavily influenced by the spatial distribution of the wave at the input plane. Similarly, the spatial distribution at the output plane is sensitive to the temporal waveform of the input field.

To illustrate this point, consider the special case in which the input wavefunction is separable in time and space, $U_1(x,t) = g(t)p(x)$. This wavefunction may be generated by transmitting a pulsed plane wave of amplitude $g(t)$ through a spatial light modulator (SLM) with frequency-independent transmittance $p(x)$, as illustrated in Fig. 23.4-3. The Fourier transform of $U_1(x,t)$ is $V_1(x,\nu) = G(\nu)p(x)$, where $G(\nu)$ is the Fourier transform of $g(t)$. Substituting $V_1(x,\nu)$, together with the impulse response function of the optical Fourier-transform system given in (23.4-8), into (23.4-4) reveals that the field at the output plane is characterized by

$$V_2(x,\nu) \propto j\nu G(\nu) P(\nu x/cf), \tag{23.4-9}$$

where $P(\nu_x) = \int p(x) \exp(j2\pi\nu_x x)\, dx$ is the spatial Fourier transform of $p(x)$.

It is evident from (23.4-9) that the output field is no longer time-space separable. Since the temporal waveform of the field at a fixed position x_0 in the output plane is governed by (23.4-9), the transfer function of the linear system that relates $U_2(x_0,t)$ to the input pulse $g(t)$ is

$$H(\nu) \propto j\nu P(\nu x_0/cf). \tag{23.4-10}$$

This *temporal* transfer function is seen to be a scaled version of the spatial Fourier transform of the input *spatial* distribution $p(x)$.

The corresponding temporal impulse response function is obtained by taking the temporal inverse Fourier transform of both sides of (23.4-10),

$$\boxed{h(t) \propto p(tcf/x_0),} \tag{23.4-11}$$

Space-to-Time Conversion

revealing that the value of the function $h(t)$ at time t is controlled by the transmittance of the SLM at one-and-only-one position, $x = (cf/x_0)t$. Equivalently, the transmittance of the mask at a point x controls the value of the impulse response function of the system at one-and-only-one time, $t = (x_0/cf)x$. This configuration thus serves as a direct space-to-time conversion system, which can be used for pulse shaping. A similar pulse-shaping system using a combination of a diffraction grating and an SLM was discussed in Sec. 23.2D.

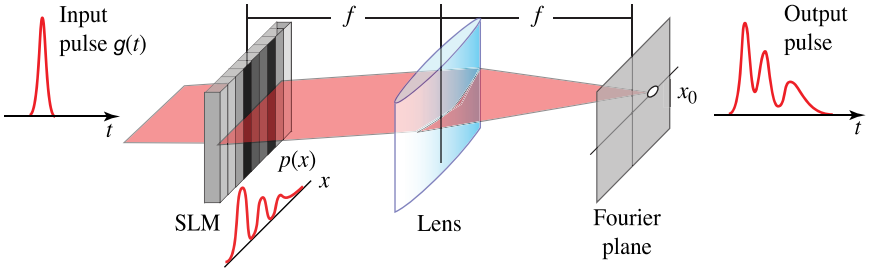


Figure 23.4-3 A spatial Fourier-transform system couples the temporal and spatial distributions of the input pulsed light. The shape of the output pulse at a fixed position is governed by the spatial distribution at the input, which is controlled by the SLM (see Sec. 21.1E).

*C. Beam Optics

The Fourier approach described in the previous section (23.4B) may be applied to the study of pulsed Gaussian beams. Consider first a Gaussian beam modulated in the plane of its waist by a pulse $g(t)$, i.e., $U_1(x,y,t) = g(t)\exp(-\rho^2/W_0^2)$ or $V_1(x,y,t) \propto G(\nu)\exp(-\rho^2/W_0^2)$, where W_0 is the beam radius and $G(\nu)$ is the Fourier transform of $g(t)$. At an arbitrary distance z, the spatiotemporal wavefunction is determined by use of (23.4-4) and (23.4-5),

$$V_2(x,y,\nu) \propto \nu\, G(\nu)\frac{jz_0(\nu)}{z+jz_0(\nu)}\exp\left(-j\frac{\pi\nu}{c}\frac{\rho^2}{z+jz_0(\nu)}\right), \tag{23.4-12}$$

where

$$z_0(\nu) = \frac{\pi W_0^2}{c}\nu \tag{23.4-13}$$

is the diffraction length (Rayleigh range) at frequency ν. Equation (23.4-12) is the standard expression of the wavefunction of a Gaussian beam [see (3.1-5)] with the frequency dependence of the diffraction length made explicit. The beam radius and the radius of curvature given by (3.1-8) and (3.1-9) are also frequency dependent.

If the spectral width is sufficiently narrow, then in accordance with the quasi-CW approximation the spatial distribution of the Gaussian beam may be approximated by its values at the central frequency $\nu \approx \nu_0$, and consequently the time-space dependence is separable, as described earlier by (23.1-25). For ultranarrow (i.e., broadband) pulses, this approximation is not applicable.

The temporal profile of the pulse may be determined at an arbitrary point (ρ, z) by evaluating the inverse Fourier transform of (23.4-12)). In general, a numerical solution is necessary.

Gaussian-Pulsed Gaussian Beam

If the original wave is modulated by a Gaussian pulse $g(t) = \exp(-t^2/\tau_0^2)\exp(j2\pi\nu_0 t)$, then $G(\nu) \propto \exp[-\pi^2\tau_0^2(\nu-\nu_0)^2]$ is also Gaussian. An approximate analytical expression for $V_2(x,y,\nu)$ in the far zone [$z \gg z_0(\nu)$ for all ν] may be obtained as follows: The factor $[z+jz_0(\nu)]^{-1} = z^{-1}[1+jz_0(\nu)/z]^{-1}$ in the exponent of (23.4-12) is approximated by $z^{-1}[1-jz_0(\nu)/z]$, and the same factor in the amplitude is approximated by z^{-1}. Using (23.4-13), we obtain the far-zone expression

$$V_2(x,y,\nu) \propto j\nu\exp\left[-\pi^2\tau_0^2(\nu-\nu_0)^2\right]\exp\left(-\pi^2\frac{W_0^2\rho^2}{c^2z^2}\nu^2\right)\exp\left(-j2\pi\nu\frac{\rho^2}{2cz}\right). \tag{23.4-14}$$

The inverse Fourier transform of (23.4-14) may now be determined. The phase factor in the exponent is equivalent to a time delay $\rho^2/2cz$. The factor $j\nu$ in the amplitude is equivalent to a derivative $\partial/\partial t$. The middle two Gaussian functions are combined into one Gaussian function of ν whose inverse Fourier transform is another Gaussian function. The result may be cast in the normalized form

$$U_2(x,y,t) \propto \frac{\exp[-\pi N\rho^2/(\rho^2+\rho_0^2)]}{1+\rho^2/\rho_0^2}\frac{\exp(-t_\rho^2/\tau_\rho^2)}{1+jt_\rho/\pi N\tau_0}\exp\left(-j2\pi\nu_\rho t_\rho\right), \tag{23.4-15}$$

Gaussian-Pulsed Gaussian Beam

where

$$t_\rho = t - \rho^2/2cz = t - \pi N\tau_0(z/z_0)(\rho^2/2\rho_0^2) \tag{23.4-16}$$

is a position-dependent delay time,

$$\tau_\rho = \tau_0\sqrt{1+\rho^2/\rho_0^2} \tag{23.4-17}$$

is a position-dependent time constant,

$$\nu_\rho = \frac{\nu_0}{1+\rho^2/\rho_0^2} = \frac{N/\tau_0}{1+\rho^2/\rho_0^2} \tag{23.4-18}$$

is a position-dependent central frequency,

$$N = \nu_0\tau_0 \tag{23.4-19}$$

is the number of optical cycles within the width τ_0 of the initial pulse, and

$$\rho_0 = \pi NW(z) = \pi NW_0 z/z_0, \tag{23.4-20}$$

where $W(z) = W_0 z/z_0$ is the far-zone beam radius for a CW wave at the central frequency ν_0 and $z_0 = \pi W_0^2/\lambda_o$ is the associated diffraction length. As a function of the normalized transverse distance ρ/ρ_0 and the normalized time t/τ_0, the far-zone wavefunction is completely described by two free parameters: N and the ratio z/z_0.

The intensity $I_2(x,y,t) = |U_2(x,y,t)|^2$ is

$$I_2(x,y,t) \propto \frac{\exp[-2\pi N\rho^2/(\rho^2+\rho_0^2)]}{1+\rho^2/\rho_0^2}\,\frac{\exp(-2t_\rho^2/\tau_\rho^2)}{1+t_\rho^2/\pi^2N^2\tau_0^2}. \tag{23.4-21}$$

This is a universal function of t/τ_0 and ρ/ρ_0 characterized by only one free parameter N. The spectral intensity $S_2(x,y,\nu) = |V_2(x,y,\nu)|^2$ is

$$S_2(x,y,\nu) \propto \frac{\nu^2}{\nu_0^2}\exp\left[-2\pi^2N^2\frac{\rho^2}{\rho^2+\rho_0^2}\right]\exp\left[-2\pi^2N^2\frac{(\nu-\nu_\rho)^2}{\nu_\rho^2}\right], \tag{23.4-22}$$

which is a universal function of ν/ν_0 and ρ/ρ_0, characterized by the free parameter N.

Based on (23.4-15)–(23.4-22), we conclude that the pulse at a point (ρ,z) in the far-zone has the following characteristics (see Fig. 23.4-4):

- The pulse is delayed by time $\rho^2/2cz$, which is the travel time between the center of the beam $(0,0)$ and the point (ρ,z).
- The pulse temporal profile is the product of a Gaussian function of width $\tau_\rho = \tau_0[1+\rho^2/\rho_0^2]^{1/2}$ and a Lorentzian function of width $\pi N\tau_0$. The width of the Gaussian function is τ_0 at $\rho = 0$, and increases with the transverse distance ρ, reaching the value $\sqrt{2}\,\tau_0$ at $\rho = \rho_0$. The phase shift $\arctan(t_p/\pi N\tau_0)$ introduced by the Lorentzian function is a manifestation of the Gouy effect (see Sec. 3.1B) for pulsed Gaussian beams.

- The pulse central frequency ν_ρ depends on the transverse distance ρ. Starting at the value ν_0 on axis ($\rho = 0$), it decreases monotonically with increase of ρ, reaching $\nu_0/2$ at $\rho = \rho_0$. This is a consequence of the fact that long-wavelength (low-frequency) components of the pulse spread into wider cones, as illustrated in Fig. 23.4-2. For the same reason, the farther the point is from the beam axis, the smaller the spectral width and the greater the temporal width.
- The initial Gaussian spatial distribution is altered dramatically as t increases. An initially single-peaked distribution builds up, is subsequently flattened, and eventually becomes double-peaked as it decays (see Fig. 23.4-4).

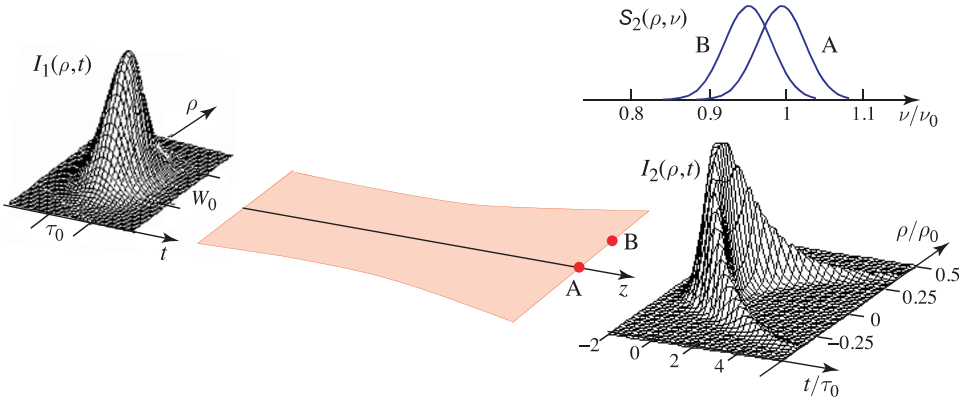


Figure 23.4-4 Temporal and spatial spreading of a Gaussian beam modulated by a Gaussian pulse. Initially, the beam has radius W_0 and temporal width τ_0 (left surface plot). The far-zone intensity $I_2(\rho,t)$ is illustrated in the right surface plot. Time is normalized to the initial pulse width τ_0, transverse distance is normalized to ρ_0, and the intensity has arbitrary units. At a fixed time t, $I_2(\rho,t)$ provides a snapshot of the intensity as a function of position. It changes from a single-peaked function at $t = 0$ to a double-peaked function at $t = \tau_0$, and eventually becomes two separate weak peaks at $t = 2\tau_0$ and beyond. The temporal profile at a fixed position is also depicted by this surface. In the center of the beam, the pulse has its shortest width. At off-axis points the pulse is weakened, delayed, and has longer duration. The spectral intensity $S_2(\rho,\nu)$ is shown (top right) as a function of the normalized frequency ν/ν_0 at two positions, A and B, and is normalized such that the peak value is unity for each position. In this plot, $N = \nu_0\tau_0 = 5$; i.e., the pulse has five optical cycles. As an example, N = 5 for a pulse of central frequency ν_0 = 750 THz (λ_0 = 400 nm) and width τ_0 = 6.67 fs. If W_0 = 1 mm, then z_0 = 7.85 m, $W_0 z/z_0$ = 5 mm, and $\rho_0 \approx 8$ cm.

Focusing of a Pulsed Beam

If a beam of arbitrary spatial distribution $p(x,y)$ modulated with a pulse of arbitrary temporal shape $g(t)$ is transmitted through a lens of focal length f followed by a distance z of free space, then by substituting $U_1(x,y,t) = g(t)p(x,y)$ into (23.4-4) and (23.4-5) we obtain

$$V_2(x,y,\nu) \propto G(\nu) \iint dx'dy'p(x',y') \exp\left(-j\tfrac{2\pi\nu}{c}d_0\right) \exp\left(j\tfrac{2\pi\nu}{c}\tfrac{x'^2+y'^2}{2f}\right) \exp\left(-j\tfrac{2\pi\nu}{c}\tfrac{(x-x')^2+(y-y')^2}{2z}\right), \quad (23.4\text{-}23)$$

where $G(\nu)$ is the Fourier transform of $g(t)$. We have here assumed that the lens has an aperture wider than the beam width.

If the lens material is nondispersive, so that n and f are independent of ν, then at points in the focal plane $z = f$, (23.4-23) simplifies to

$$V_2(x, y, \nu) \propto \nu\, G(\nu) P\left(\frac{\nu x}{cf}, \frac{\nu y}{cf}\right) \exp\left(-j\frac{2\pi\nu}{c}\frac{\rho^2}{2f}\right), \tag{23.4-24}$$

where $P(\nu_x, \nu_y) = \iint dxdyp(x, y) \exp\left[j2\pi\left(\nu_x x + \nu_y y\right)\right]$ is the spatial Fourier transform of $p(x, y)$. The factor $\exp(-j2\pi\nu d_0/c)$ has been ignored since it now represents a simple time delay. The wavefunction in the focal plane is the temporal inverse Fourier transform of $V_2(x, y, \nu)$, so that

$$U_2(x, y, t) \propto \int \nu G(\nu) P\left(\frac{\nu x}{cf}, \frac{\nu y}{cf}\right) \exp\left[j2\pi\nu\left(t - \frac{\rho^2}{2cf}\right)\right] d\nu. \tag{23.4-25}$$

The coupling of the temporal and spatial features of the pulsed beam is evident in (23.4-25). In addition to the space-dependent time delay $t - \rho^2/2cf$, the Fourier transform of the original spatial profile is scaled by the frequency-dependent factor cf/ν before it is averaged over the spectral distribution of the pulse.

As an example, for a Gaussian beam $p(x, y) = \exp\left(-\rho^2/W_0^2\right)$ modulated by a Gaussian pulse $g(t) = \exp(-t^2/\tau_0^2)\exp(j2\pi\nu_0 t)$, i.e., $G(\nu) \propto \exp[-\pi^2\tau_0^2(\nu - \nu_0)^2]$ and $P(\nu_x, \nu_y) = \exp\left[-\pi^2 W_0^2\left(\nu_x^2 + \nu_y^2\right)\right]$, (23.4-24) gives

$$V_2(x, y, \nu) \propto \nu \exp\left[-\pi^2\tau_0^2\left(\nu - \nu_0\right)^2\right] \exp\left[-\left(\frac{\pi W_0}{cf}\right)^2 \nu^2\rho^2\right] \exp\left(-j2\pi\nu\frac{\rho^2}{2cf}\right). \tag{23.4-26}$$

This expression is identical to that for the far-zone Gaussian beam (23.4-14), with $z = f$. Thus, the corresponding wavefunction $U_2(x,y,t)$ is given by (23.4-15)–(23.4-22) with $z = f$. The graphs in Fig. 23.4-4 are applicable here with $z = f$, z_0 being the diffraction length of the original (not the focused) beam, and

$$\rho_0 = \pi N W_0 f/z_0 = N\lambda_o f/W_0 = \pi N W_0', \tag{23.4-27}$$

where $W_0' = \lambda_o f/\pi W_0$ is the beam radius at the focal plane for a CW beam with wavelength λ_o [see (3.2-15) and (3.2-17)]. As before, $N = \nu_0\tau_0$ is the number of optical cycles within the initial pulse. The characteristic transverse radius ρ_0 is therefore πN times greater than W_0'. Figure 23.4-5 is an illustration of the spatiotemporal distribution of the pulse in the focal plane.

* Pulsed Beams in Dispersive Media

The process of diffraction of pulsed light in a dispersive medium can be complex. If the medium is linear and homogeneous, then the Helmholtz equation $[\nabla^2 + \beta^2(\nu)]V(\mathbf{r}, \nu) = 0$ describes this process for arbitrary dispersion properties, characterized by the propagation constant $\beta(\nu)$, and for a pulse with arbitrary spatial-spectral profile $V(\mathbf{r}, \nu)$. Once $V(\mathbf{r}, \nu)$ is determined by solving this equation, the corresponding wavefunction $U(\mathbf{r}, t)$ may be readily determined by an inverse Fourier transform. This approach is, in principle, valid no matter how dispersive the medium or how narrow the pulse.

Approximations similar to those that led independently to the paraxial Helmholtz equation, which describes beam diffraction, and the SVE equation, which describes pulse dispersion (see Table 23.3-2), may be combined to derive a partial differential equation for the envelope $\mathcal{A}(\mathbf{r}, t)$ of a pulse with a narrow spectral distribution. An

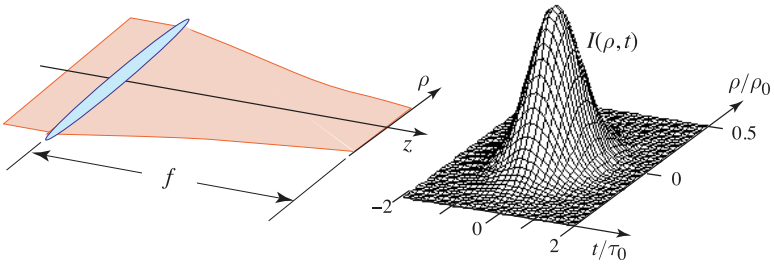


Figure 23.4-5 Focal-plane spatiotemporal profile of the intensity of a Gaussian beam modulated by a Gaussian pulse and focused by a lens of focal length f. In this plot the initial pulse has $N = 5$ optical cycles and the initial beam has a diffraction length $z_0 \gg f$. The difference between this and the spatiotemporal profile in Fig. 23.4-4 is attributed to the fact that here the time delay $\rho^2/2cf = \tau_0(\pi N/2)(f/z_0)(\rho^2/\rho_0^2)$ is negligible for $f \ll z_0$ at off-axis points with $\rho < \rho_0$.

approach following the same steps described in Sec. 23.2C results in the generalized paraxial wave equation:

$$-\lambda_o \nabla_T^2 \mathcal{A} + D_\nu \frac{\partial^2 \mathcal{A}}{\partial t^2} + j\left(\frac{\partial \mathcal{A}}{\partial z} + \frac{1}{v}\frac{\partial \mathcal{A}}{\partial t}\right) = 0. \tag{23.4-28}$$

Generalized Paraxial Wave Equation

This equation generalizes (23.1-24), which is applicable for nondispersive media ($D_\nu = 0$), as well as (2.2-23), which is applicable for the CW case for which $\partial^2 A/\partial t^2 = \partial A/\partial t = 0$.

□ **Proof of the Generalized Paraxial Wave Equation.** The wavefunction and its Fourier transform are related to the envelope and its Fourier transform by $U(\mathbf{r}, t) = \mathcal{A}(\mathbf{r}, t)\exp(-j\beta_0 z)\exp(j2\pi\nu_0 t)$ and $V(\mathbf{r}, \nu) = A(\mathbf{r}, \nu - \nu_0)\exp(-j\beta_0 z)$. The paraxial approximation, $(d^2/dz^2)[A\exp(-j\beta_0 z)] \approx [-j2\beta_0 dA/dz - \beta_0^2 A]\exp(-j\beta_0 z)$, can be used to convert the Helmholtz equation to

$$\left[\nabla_T^2 - j2\beta_0 d/dz\right] A + \left[\beta^2(\nu_0 + f) - \beta_0^2\right] A = 0. \tag{23.4-29}$$

For weak dispersion, we use the approximation $\beta^2(\nu_0 + f) - \beta_0^2 \approx 2\beta_0[\beta(\nu_0 + f) - \beta_0]$ together with a 3-term Taylor-series expansion $\beta(\nu_0 + f) = \beta_0 + 2\pi\beta' f + 2\pi^2\beta'' f^2$. The Helmholtz equation then becomes

$$\nabla_T^2 A - j2\beta_0 \partial A/\partial z + 2\beta_0\left[2\pi f\beta' + 2\pi^2 f^2 \beta''\right] A = 0. \tag{23.4-30}$$

Performing an inverse Fourier transform and noting that the multipliers $j2\pi f$ and $-4\pi^2 f^2$ are equivalent to the derivatives $\partial/\partial t$ and $\partial^2/\partial t^2$, respectively, we obtain

$$\nabla_T^2 A - 2\beta_0\left[j\partial A/\partial z + j\beta'\partial A/\partial t + \frac{1}{2}\beta''\partial^2 A/\partial t^2\right] = 0. \tag{23.4-31}$$

Finally, substituting $\beta' = 1/v$ and $\beta'' = D_\nu/2\pi$ and $\beta_0 = 2\pi/\lambda_o$, we obtain (23.4-28). ■

The paraxial SVE equation admits a space–time Gaussian solution

$$A(x, y, z, t) = A_0\sqrt{\frac{-jz_0'}{z - jz_0'}}\,\exp\left[-j\frac{\pi}{D_\nu}\frac{t - z/v}{z - jz_0'}\right]\cdot\frac{jz_0}{z + jz_0}\,\exp\left(-j\frac{\pi}{\lambda}\frac{\rho^2}{z + jz_0}\right), \tag{23.4-32}$$

that has a spatiotemporal Gaussian initial envelope $\mathcal{A}(x,y,0,t) = A_0 \exp(-t^2/\tau_0^2)$ $\exp(-\rho^2/W_0^2)$, where $z_0' = \pi\tau_0^2/D_\nu$ and $z_0 = \pi W_0^2/\lambda$ are, respectively, the dispersion length associated with the initial pulse width τ_0 and the diffraction length associated with the initial beam radius W_0. This solution combines the diffraction of a Gaussian beam (Chapter 3) and the dispersion of a Gaussian pulse (Sec. 23.3) in a space–time separable fashion, as illustrated in Fig. 23.4-6.

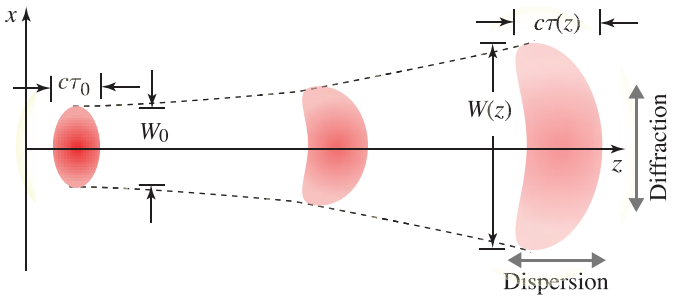


Figure 23.4-6 Three snapshots of the spatial distribution of a pulse as it travels through a linear dispersive medium. Because of diffraction, the pulse spreads in the transverse direction x. Because of dispersion, it spreads in time (which is shown here as spatial spread in the direction of propagation z).

Since (23.4-28) and (23.4-32) are separable in time and space, we conclude that the approximations to which these equations are subject are in effect tantamount to the quasi-CW approximation described in Sec. 23.1C.

*Envelope Equation for Ultranarrow Pulsed Beam

When conditions for the SVE approximation are not met (i.e., the pulse is very narrow and the beam is very thin), then the space–time dependence is no longer separable. The differential equation that governs the pulse envelope takes a more complex form, although the very concept of envelope is then less meaningful. Beginning with the Helmholtz equation $[\nabla^2 + \beta^2(\nu)]V(\mathbf{r},\nu) = 0$ and substituting $V(\mathbf{r},\nu) = A(\mathbf{r},\nu-\nu_0)\exp(-j\beta_0 z)$ and $\nu = \nu_0 + f$, we obtain $\left[\nabla_T^2 + \partial^2/\partial z^2 - j2\beta_0\partial/\partial z\right] A + \left[\beta^2(\nu_0+f) - \beta_0^2\right] A = 0$. Expanding the function $[\beta^2(\nu_0+f)-\beta_0^2]$ in a Taylor-series expansion up to the second order, we have $[\beta^2(\nu_0+f)-\beta_0^2] \approx (2\beta_0\beta')2\pi f + \frac{1}{2}(2\beta'^2 + 2\beta_0\beta'')(2\pi f)^2$. Transforming back to the time domain and reordering terms, we obtain

$$-\lambda_0\nabla_T^2\mathcal{A} + D_\nu\frac{\partial^2\mathcal{A}}{\partial t^2} + j4\pi\left(\frac{\partial}{\partial z} + \frac{1}{v}\frac{\partial}{\partial t}\right)\mathcal{A} - \lambda_0\left(\frac{\partial^2}{\partial z^2} - \frac{1}{v}\frac{\partial^2}{\partial t^2}\right)\mathcal{A} = 0, \quad (23.4\text{-}33)$$

where $v = 1/\beta'$ and $D_\nu = 2\pi\beta''$. Equation (23.4-33) is more general than (23.4-28) since the paraxial approximation and the weak dispersion approximation have not been used. If $\beta'^2 \ll \beta_0\beta''$ (or $\lambda_o/v^2 \ll D_\nu$) and $\partial^2\mathcal{A}/\partial z^2 \ll (4\pi/\lambda_o)\partial\mathcal{A}/\partial z$, then the fourth term in (23.4-33) is negligible and (23.4-33) reproduces (23.4-28).

Equation (23.4-33) may be expressed in a coordinate system moving at the pulse velocity v by using the transformation $t' = t - z/v$ and $z' = z$. The result is the differential equation

$$-\lambda_o\nabla_T^2\mathcal{A} + D_\nu\frac{\partial^2\mathcal{A}}{\partial t'^2} + j4\pi\frac{\partial\mathcal{A}}{\partial z'} - \lambda_o\left(\frac{\partial^2\mathcal{A}}{\partial z'^2} - \frac{2}{v}\frac{\partial^2\mathcal{A}}{\partial t'\partial z'}\right) = 0, \quad (23.4\text{-}34)$$

which clearly exhibits spatiotemporal coupling.

23.5 ULTRAFAST NONLINEAR OPTICS

The previous sections of this chapter have considered the propagation of optical pulses in *linear* media, with an emphasis on the role of group velocity dispersion (GVD) in the reshaping of short pulses. In this section, we consider the propagation of optical pulses in *nonlinear* media. Nonlinear effects are more frequently encountered with ultrashort pulses because of their higher intensity. Nonlinear optical phenomena were introduced in Chapter 22; in particular, three-wave mixing in media with second-order nonlinearity, and two- and four-wave mixing in media with third-order nonlinearity, were considered. In this section, some of these phenomena are revisited in the context of pulsed optical waves. Section 23.5A deals with pulsed parametric processes, including three-wave mixing, optical rectification, and self-phase modulation; Sec. 23.5B details the theory of optical solitons; Sec. 23.5C is devoted to supercontinuum generation; and Sec. 23.5D considers high-harmonic generation and attosecond optics.

A. Pulsed Parametric Processes

Three-wave mixing in a medium with second-order nonlinearity was discussed in Sec. 22.2C for continuous waves (CW), and a coupled-wave theory was developed in Sec. 22.4. The principal conditions for wave mixing are dictated by conservation of energy and momentum. For pulsed waves with central angular frequencies ω_1, ω_2 and ω_3, and central wavevectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$, these conditions are: $\omega_1 + \omega_2 = \omega_3$ and $\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3$. If dispersion effects are neglected, the CW theory is applicable to the pulsed case; i.e., the pulse is regarded as "quasi-CW" at any time during its course, and the envelopes of the three waves obey the same coupled-wave equations (22.4-20).

The Walk-Off Effect

If the medium exhibits first-order dispersion, but not second-order (GVD) or higher-order dispersion, then the three pulsed waves travel at their group velocities without altering their shapes (only their amplitudes are altered by the mixing process). Since these velocities are generally different, the pulses eventually separate and the parametric process responsible for wave mixing ceases, a phenomenon known as the **walk-off effect**. Therefore, for efficient pulsed-wave mixing, an additional condition is the equality of the group velocities, $v_1 = v_2 = v_3$. The walk-off effect is illustrated in Fig. 23.5-1 in the degenerate case of collinear second-harmonic generation ($\omega_1 = \omega_2 = \omega$ and $\omega_3 = 2\omega$).

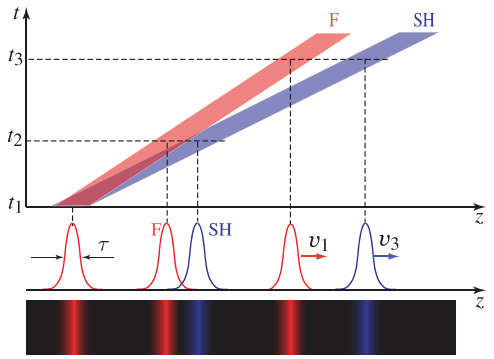


Figure 23.5-1 A pulsed wave at the fundamental frequency (F) and its associated second-harmonic wave (SH) separate as they travel at different velocities (in this example, the SH wave is faster). The upper graph is a space–time diagram for pulses of duration τ. The lower schematic shows three snapshots of the traveling pulses at times $t_1 < t_2 < t_3$.

It is difficult to satisfy both phase matching and group-velocity matching simultaneously. It was shown in Sec. 22.2D and Sec. 22.4A that for a phase matching error

Δk, second-harmonic generation diminishes significantly at a distance $L_c = 2\pi/|\Delta k|$, called the coherence length [see (22.2-28)]. For a group-velocity matching error $\Delta\beta' = 1/v_3 - 1/v_1$, the pulses separate by a time delay $\Delta\beta' z = z/v_3 - z/v_1$ after traveling a distance z. When this delay equals the pulse width τ, the pulses no longer overlap and the nonlinear coupling ceases. This occurs at a distance

$$\boxed{L_g = \tau/|\Delta\beta'|} \qquad \begin{matrix}(23.5\text{-}1)\\ \text{Walk-Off Length}\end{matrix}$$

called the walk-off length. The shorter of the distances L_c and L_g dictates which of the two effects, phase-velocity mismatch or group-velocity mismatch, dominates.

As an example, for a KDP crystal using an ordinary fundamental wave at λ_1 = 1.06 μm and an extraordinary second-harmonic wave at λ_3 = 0.53 μm in the Type-II o-e-o configuration, the group velocity mismatch $\Delta\beta' = 2(1/v_3 - 1/v_1) \approx 5.2 \times 10^{-10}$ s/m. For a 100-fs pulse, the walk-off length $L_g = \tau/|\Delta\beta'| \approx 0.2$ mm.

* Coupled-Wave Equations for Pulsed Three-Wave Mixing

The coupled-wave equations that were derived in Sec. 22.4 for CW waves may be readily generalized to pulsed waves. For collinear plane waves traveling in the z direction, the electric fields are expressed in terms of the complex envelopes as $\mathcal{E}_q = \text{Re}\{\sqrt{2\eta\hbar\omega_q}\ \mathfrak{a}_q(z,t)\exp[j(\omega_q t - \beta_q z)]\}$, q = 1,2,3, where $\mathfrak{a}_1$, $\mathfrak{a}_2$, and $\mathfrak{a}_3$ are normalized complex envelopes of the three pulses, and β_1, β_2 and β_3 are the propagation constants at the central frequencies ω_1, ω_2, and ω_3. Using the slowly varying envelope approximation and a two-term Taylor-series expansion of the propagation constant $\beta(\omega)$ near each of the central frequencies $\beta(\omega_q + \Omega) \approx \beta_q + \Omega\beta'_q$, where β'_q is the derivative $\partial\beta/\partial\omega$ at ω_q, we obtain the coupled equations:

$$\begin{aligned}
\left(\frac{\partial}{\partial z} + \frac{1}{v_1}\frac{\partial}{\partial t}\right)\mathfrak{a}_1 &= -j\mathsf{g}\mathfrak{a}_3\mathfrak{a}_2^*\\
\left(\frac{\partial}{\partial z} + \frac{1}{v_2}\frac{\partial}{\partial t}\right)\mathfrak{a}_2 &= -j\mathsf{g}\mathfrak{a}_3\mathfrak{a}_1^*\\
\left(\frac{\partial}{\partial z} + \frac{1}{v_3}\frac{\partial}{\partial t}\right)\mathfrak{a}_3 &= -j\mathsf{g}\mathfrak{a}_1\mathfrak{a}_2,
\end{aligned} \qquad (23.5\text{-}2)$$

where $v_q = 1/\beta'_q$ is the group velocity of the ω_q wave, and g is a constant given by (22.4-21). These equations are similar to the CW coupled equations (22.4-20). If the group velocities are equal, i.e., $v_1 = v_2 = v_3 = v$, then by use of a coordinate system moving with a velocity v, the pulsed coupled equations (23.5-2) become identical to the CW coupled equations (22.4-20), and the solutions presented in Sec. 22.4 are applicable with the variable z replaced by $z - vt$. If the group velocities are not equal, the solution of (23.5-2) becomes more complex.

When the medium also exhibits GVD (see Prob. 23.5-2), a three-term Taylor-series expansion $\beta(\omega_q + \Omega) \approx \beta_q + \Omega\beta'_q + \frac{1}{2}\Omega^2\beta''_q$ leads to the coupled-wave equations:

$$\begin{aligned}
\left(\frac{\partial}{\partial z} + \frac{1}{v_1}\frac{\partial}{\partial t} - j\frac{\beta''_1}{2}\frac{\partial^2}{\partial t^2}\right)\mathfrak{a}_1 &= -j\mathsf{g}\mathfrak{a}_3\mathfrak{a}_2^*\\
\left(\frac{\partial}{\partial z} + \frac{1}{v_2}\frac{\partial}{\partial t} - j\frac{\beta''_2}{2}\frac{\partial^2}{\partial t^2}\right)\mathfrak{a}_2 &= -j\mathsf{g}\mathfrak{a}_3\mathfrak{a}_1^*
\end{aligned} \qquad (23.5\text{-}3)$$

$$\left(\frac{\partial}{\partial z}+\frac{1}{v_3}\frac{\partial}{\partial t}-j\frac{\beta_3''}{2}\frac{\partial^2}{\partial t^2}\right)\mathrm{a}_3=-j\mathrm{g}\mathrm{a}_1\mathrm{a}_2.$$

Pulsed Optical Rectification: THz Pulse Generation

A pulsed wave with central frequency in the optical band and spectral width in the THz range may be downconverted into a pulse of THz radiation. In essence, the pulse is frequency shifted from the optical band to the THz band, as if it were rectified. Figure 23.5-2 is a schematic illustration of the process.

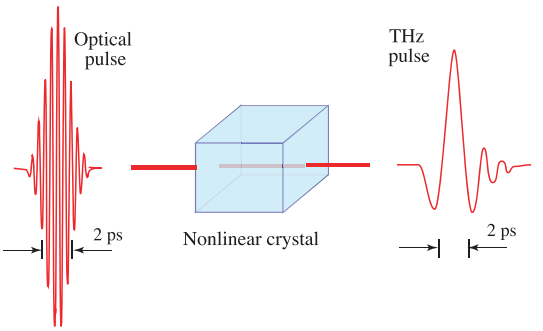


Figure 23.5-2 Generation of a THz pulse by downconversion of an optical wave.

When an optical pulse $\mathcal{E}(t)=\mathrm{Re}\{\mathcal{A}(t)\exp(j\omega_0 t)\}$ with slowly varying envelope $\mathcal{A}(t)$ travels through a medium with second-order nonlinear optical coefficient d, it induces a polarization density $2\mathrm{d}\mathcal{E}^2(t)$, which has a term at $2\omega_0$, responsible for second-harmonic generation, and another,

$$\mathcal{P}_{\mathrm{THz}}=\mathrm{d}|\mathcal{A}(t)|^2, \tag{23.5-4}$$

representing optical rectification (see Secs. 22.2A, 22.2C, and 22.4B).

In order to determine the appropriate phase matching conditions for this parametric process, we resort to a Fourier approach. The pulsed optical wave can be regarded as a sum of monochromatic waves with frequencies occupying a spectral band surrounding the central frequency ω_0. Upon passage through the nonlinear medium, these monochromatic components are mixed in pairs, each generating a downconverted monochromatic wave at the frequency difference. In accordance with (22.2-13e), a pair of waves at the angular frequencies $\omega_1=\omega$ and $\omega_2=\omega+\Omega$ generates a nonlinear polarization density $P_{\mathrm{THz}}(\Omega)=2\mathrm{d}E^*(\omega)E(\omega+\Omega)$ at the THz frequency Ω so that the sum for all the pairs is

$$P_{\mathrm{THz}}(\Omega)=\int 2\mathrm{d}E^*(\omega)E(\omega+\Omega)d\omega. \tag{23.5-5}$$

In the time domain, this is equivalent to (23.5-4). To include nonlinear dispersion effects, the nonlinear optical coefficient d in (23.5-5) must be replaced by a frequency-dependent version $\mathrm{d}(\Omega,\omega,\omega+\Omega)$ (see Sec. 22.7).

This downconversion process must satisfy the phase matching condition at all frequencies ω and Ω. This condition cannot be met exactly, and an error

$$\Delta k=k(\omega+\Omega)-k(\omega)-k(\Omega) \tag{23.5-6}$$

will arise. If $\Omega\ll\omega$, this relation may be written in the approximate form

$$\Delta k\approx\Omega dk/d\omega-k(\Omega)=\Omega[1/v(\omega)-1/c(\Omega)]=[N(\omega)-n(\Omega)]\Omega/c_o, \tag{23.5-7}$$

where $v(\omega) = (dk/d\omega)^{-1}$ is the group velocity and $N(\omega)$ is the group index at the optical frequency ω, and $c(\Omega)$ and $n(\Omega)$ are the phase velocity and refractive index at the THz frequency Ω. The device must therefore be designed in such a way that the group index at optical frequencies is equal to the phase index at THz frequencies.

As was shown in Sec. 22.2D for a crystal of length L, this phase-matching error is small if $L < L_c$, where $L_c = 2\pi/|\Delta k|$ is the coherence length [see (22.2-28)]. To account for this effect, the factor $\int_0^L \exp(j\Delta kz)dz = [\exp(j\Delta kL) - 1]/j\Delta k$ must be included within the integral of (23.5-5).

Pulse Self-Phase Modulation

Self-phase modulation (SPM) occurs in nonlinear media that exhibit the optical Kerr effect (see Sec. 22.3B). The phase $\Delta\varphi$ introduced by this effect for a wave traveling a distance z in a medium with optical Kerr coefficient n_2 is $\Delta\varphi = -n_2 I k_0 z$, where I is the optical intensity and k_0 is the wavenumber. For an optical pulse, the intensity is a function of time $I(t)$ so that the phase is time varying:

$$\Delta\varphi(t) = -n_2 I(t) k_0 z. \tag{23.5-8}$$

This corresponds to a change of the instantaneous frequency [see (23.1-4)]

$$\Delta\omega_{\mathrm{i}} = -n_2 \frac{dI}{dt} k_0 z. \tag{23.5-9}$$

For a pulse with a simple shape, such as that illustrated in Fig. 23.5-3, if n_2 is positive, the frequency of the trailing half of the pulse (the right half) is increased (blue shifted) since $dI/dt < 0$, whereas the frequency of the leading half (the left half) is reduced (red shifted) since $dI/dt > 0$. The pulse is therefore up-chirped (i.e., its instantaneous frequency is increasing) near its center. It follows that SPM may be used to introduce chirp, and may therefore be employed for pulse shaping (see Sec. 23.2D).

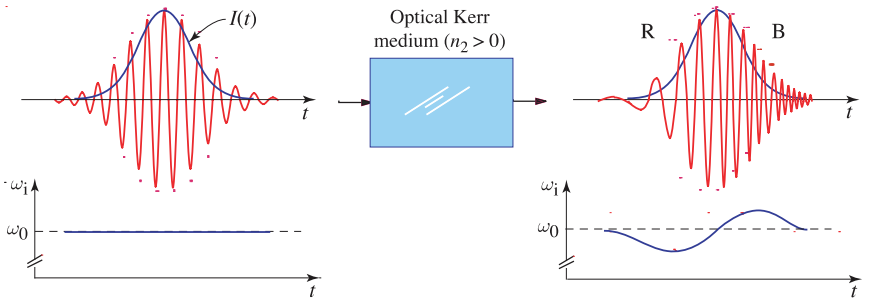


Figure 23.5-3 Chirping of an optical pulse by propagation through a nonlinear optical Kerr medium.

For example, a Gaussian pulse may be approximated near its center by a parabolic function, $I(t) = I_0 \exp(-2t^2/\tau^2) \approx I_0[1 - 2t^2/\tau^2]$, so that the time-varying component of the phase is approximately a quadratic function of time $\Delta\varphi = 2n_2 I_0 k_0 z t^2/\tau^2$, corresponding to a linear chirp with chirp coefficient $a = 2n_2 I_0 k_0 z$ of the same sign as the Kerr coefficient n_2. Self-phase modulation therefore introduces a quadratic phase modulation factor $\exp(jat^2/\tau^2) = \exp(j\zeta t^2)$, where

$$\zeta = 2n_2 I_0 k_0 z/\tau^2. \tag{23.5-10}$$

It is convenient to write the chirp parameter introduced by SPM in the form

$$a = z/z_{\mathrm{NL}}, \quad z_{\mathrm{NL}} = (2n_2 I_0 k_0)^{-1}, \tag{23.5-11}$$

SPM Chirp Parameter

where $|z_{\mathrm{NL}}|$ is called the **nonlinear characteristic length** of the Kerr medium. The phase introduced by traveling through the nonlinear material a distance $2|z_{\mathrm{NL}}|$ at the peak intensity I_0 is unity, i.e., $n_2 I_0 k_0 2|z_{\mathrm{NL}}| = 1$.

It has been implicitly assumed in the preceding analysis that the medium is weakly dispersive so that pulse broadening is negligible; i.e., GVD is negligible in comparison with SPM. This condition obtains if $|z_0| \gg |z_{\mathrm{NL}}|$. Analysis of pulse propagation in materials exhibiting both SPM and GVD is complex, as will be seen in the next section.

The quadratic phase modulation introduced by nonlinear SPM may be used in conjunction with a linear dispersive device, such as a diffraction grating or prism module, to implement pulse compression, as described in Sec. 23.2C and illustrated in Example 23.5-1. The combination results in pulse compression by a factor $\sqrt{1+a^2} = \sqrt{1+(z/z_{\mathrm{NL}})^2}$.

EXAMPLE 23.5-1. ***Pulse Compression Using Fiber SPM and Grating GVD.*** A 65-fs pulse of peak power P_0 = 300 kW at a central wavelength λ_o = 620 nm is chirped by a 9-mm long silica-glass optical fiber of cross-sectional area A = 100 μm^2, as illustrated in Fig. 23.5-4. At this wavelength, $n_2 \approx 3.2 \times 10^{-20}$ m^2/W so that the nonlinear characteristic length is $|z_{\mathrm{NL}}| = |2n_2 I_0 k_0|^{-1} = \lambda_o A/4\pi|n_2|P_0 \approx 0.5$ mm. Since the fiber length $z = 9$ mm, the chirp parameter introduced by the SPM is $a = z/z_{\mathrm{NL}} = 18$. This corresponds to a maximum pulse compression factor $\sqrt{1+a^2} \approx 18$, or a compressed pulse of width 3.6 fs. The fiber also introduces GVD. At 620 nm, $\beta'' = 6 \times 10^{-26}$ s^2/m, so that the dispersion length for a pulse of width $\tau_0 = 65$ fs is $z_0 = \tau_0^2/2\beta'' = 3.5$ cm. Since $z_0 \gg z_{\mathrm{NL}}$, SPM dominates GVD. To achieve maximum compression, the grating must introduce a chirp coefficient $b = [a/(1+a^2)]\tau_0^2 \approx 2.35 \times 10^{-28}$ s^2 = (3.6 fs)^2.

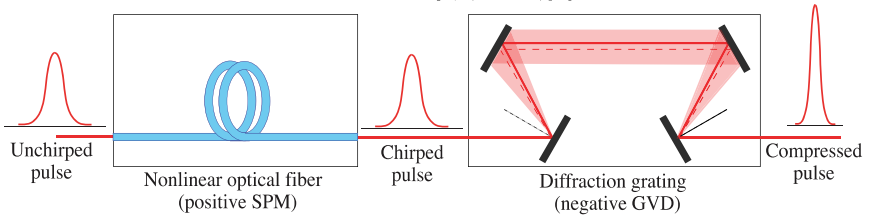


Figure 23.5-4 Pulse compression by a combination of a quadratic phase modulation (QPM) (introduced by SPM) and a chirp filter. The phase modulator is implemented using an optical fiber exhibiting SPM, via the optical Kerr effect. The chirp filter is implemented using the GVD introduced by a diffraction grating.

B. Optical Solitons

The interplay between self-phase modulation (SPM) and group velocity dispersion (GVD) in a medium exhibiting both the nonlinear optical Kerr effect and linear dispersion can result in a net pulse spreading or pulse compression, depending on the magnitudes and signs of these two effects. Under certain conditions, an optical pulse of prescribed shape and intensity can travel in such a *nonlinear dispersive* medium without ever altering its shape, as if it were traveling in an ideal *linear nondispersive* medium. This occurs when GVD fully compensates the effect of SPM, as illustrated in

Fig. 23.5-5(c). Such pulse-like stationary waves are called **solitary waves**. Optical **solitons** are special solitary waves that are orthogonal, in the sense that when two of these waves cross one another in the medium their intensity profiles are not altered (only phase shifts are imparted as a result of the interaction), so that each wave continues to travel as an independent entity.

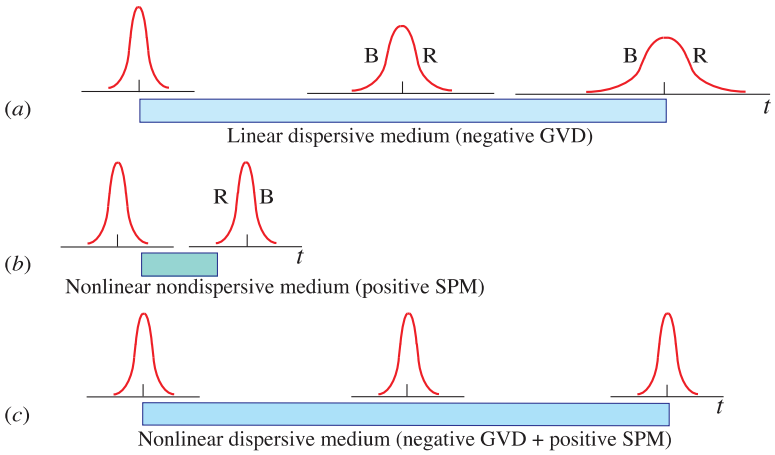


Figure 23.5-5 (a) In a linear medium with negative GVD (anomalous dispersion), the shorter-wavelength component B has a larger group velocity and therefore travels faster than the longer-wavelength component R; this results in pulse spreading. (b) In a nonlinear medium with positive optical Kerr effect ($n_2 > 0$), SPM introduces a negative frequency shift in the leading half of the pulse (denoted R) and a positive-frequency shift in the trailing half (denoted B). The pulse is chirped, but its shape is not altered. If the chirped wave in (b) travels in the linear dispersive medium depicted in (a), the pulse will be compressed since the blue-shifted half catches up with the red-shifted half. (c) If the medium is both nonlinear and dispersive, the pulse can be compressed, expanded, or maintained (creating a solitary wave), depending on the magnitudes and signs of the GVD and SPM. This illustration shows a solitary wave created by a balance between negative GVD and positive SPM.

The soliton process may be visualized via the mechanical analog illustrated by the cartoon in Fig. 23.5-6. Here, the heavy car represents the central portion of the optical pulse. It alters the surface of the ground, assumed to be elastic, much like the intense pulse peak alters the refractive index of the medium. The fast sports car, which is analogous to the trailing side of the pulse, is slowed down by the inclination created in the surface. The slow bicycle, which is analogous to the leading side of the pulse, is accelerated by the down-sloped surface. The result of this self-sustained process is that the three members of the team travel at the same velocity, and maintain the distances that separate them.

Figure 23.5-6 Transportation analog of the soliton.

Solitons have a characteristic pulse profile and level of intensity for which the effects of SPM and GVD are balanced. For these pulses, the chirping effect of SPM perfectly compensates the natural pulse expansion caused by the GVD. Any slight spreading of the pulse enhances the compression process, and any pulse narrowing reduces the

compression process, so that the pulse shape and width are maintained. Solitons can be thought of as the modes (eigenfunctions) of the nonlinear dispersive system. A mathematical analysis of this phenomenon is based on solutions of the nonlinear wave equation that governs the propagation of the pulse envelope, as described subsequently. However, we first present a simple derivation of the soliton condition.

Soliton Condition

An expression for the soliton condition is obtained by equating the sum of the phases introduced by SPM and GVD to zero, within an incremental distance Δz. As described earlier in this section, a pulse traveling through a nonlinear medium exhibiting the optical Kerr effect undergoes SPM, which introduces a quadratic phase modulation $\exp(j\zeta t^2)$, with $\zeta = 2n_2 I_0 k_0 \Delta z/\tau_0^2$, where I_0 and τ_0 are the pulse peak intensity and width, respectively, and n_2 is the optical Kerr coefficient. Also, as described in Sec. 23.3, GVD in a linear dispersive medium introduces a phase shift at^2/τ_0^2, where the chirp parameter $a = \Delta z/z_0 = 2\beta''\Delta z/\tau_0^2$, and where β'' is the material dispersion coefficient and $|z_0|$ is the dispersion length (see Table 23.3-1).

The condition that the pulse travel as a soliton is that the two phase factors are equal in magnitude and opposite in sign, so that

$$\zeta = -\frac{a}{\tau_0^2}\,, \tag{23.5-12}$$

or, equivalently,

$$\boxed{k_0 n_2 I_0 = -\frac{\beta''}{\tau_0^2}\,.} \tag{23.5-13}$$

Soliton Condition (Phase)

This expression in turn is equivalent to

$$\boxed{z_{\mathrm{NL}} = -z_0\,,} \tag{23.5-14}$$

Soliton Condition (Length)

which indicates that the GVD dispersion length must equal the nonlinear characteristic length. Stated differently, the phase shift introduced by SPM over a propagation distance equal to twice the GVD dispersion length $|z_0|$ is unity ($-k_0 n_2 I_0 2z_0 = 1$).

We may alternatively derive this condition by thinking of the medium as a periodic sequence of localized SPM elements separated by pulse-spreading elements (GVD) of widths Δz, as illustrated in Fig. 23.5-7. The scheme is identical to the pulse-relaying system described in Exercise 23.3-2. In fact (23.5-12) may be derived from (23.3-23) in the limit as $\Delta z \to 0$.

An expression for the soliton condition can also be cast in terms of the pulse amplitude A_0, where $I_0 = |A_0|^2/2\eta$ and η is the electromagnetic impedance of the medium. The result is written in terms of the product of the peak pulse amplitude A_0 and the temporal width τ_0,

$$\boxed{A_0\tau_0 = \sqrt{-\beta''/\gamma},} \tag{23.5-15}$$

Soliton Condition (Area)

where

$$\gamma = k_0 n_2/2\eta = \pi n_2/\lambda_o \eta \tag{23.5-16}$$

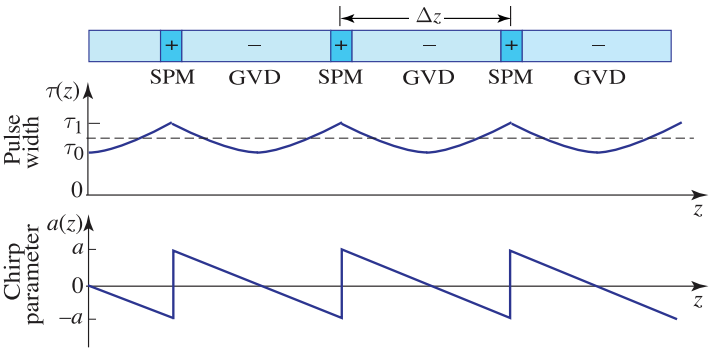


Figure 23.5-7 Simple model for a medium with negative GVD and positive SPM.

is a nonlinear coefficient that serves as another material parameter. Note that γ and β'' are assumed to have opposite signs. Thus, the product $A_0\tau_0$ of the peak amplitude and width is a constant determined by the ratio of the parameter β'', which describes GVD, and the parameter γ, which describes SPM. For a given material, the product $A_0\tau_0$ is fixed, which has the following implications:

- The pulse peak amplitude A_0 is inversely proportional to the pulse width τ_0.
- The pulse peak power is inversely proportional to τ_0^2.
- The pulse energy density $\int I(t)dt$ is inversely proportional to τ_0, so that a soliton of shorter duration must carry greater energy.

By solving the nonlinear wave equation that governs pulse propagation in a medium exhibiting both SPM and GVD, it will be shown subsequently that one of the solutions is the **soliton pulse**

$$|\mathcal{A}(t)| = |A_0|\,\mathrm{sech}(t/\tau_0), \tag{23.5-17}$$

Soliton Envelope

where $\mathrm{sech}(\cdot) = 1/\cosh(\cdot)$ is the hyperbolic-secant function illustrated in Fig. 23.5-8. This symmetric bell-shaped function has the following characteristics:

- Peak amplitude = A_0
- FWHM width of amplitude profile $= 2.63\,\tau_0$
- Area under amplitude profile = $2\pi A_0\tau_0$
- Intensity $I(t) \propto |A_0|^2\,\mathrm{sech}^2(t/\tau_0)$; width $\tau_{\mathrm{FWHM}} = 1.76\,\tau_0$

The Nonlinear Slowly Varying Envelope Wave Equation

To describe the propagation of an optical pulse in a nonlinear dispersive medium exhibiting both GVD and SPM, we begin with the wave equation in (5.2-25) and (22.1-3),

$$\left[\nabla^2 - \frac{1}{c_o^2}\frac{\partial^2}{\partial t^2}\right]\mathcal{E} = \mu_o\frac{\partial^2}{\partial t^2}\left(\mathcal{P}_{\mathrm{L}} + \mathcal{P}_{\mathrm{NL}}\right), \tag{23.5-18}$$

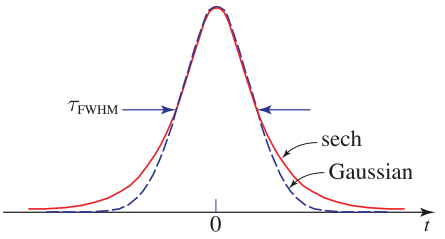


Figure 23.5-8 Comparison of a sech function and a Gaussian function of the same height and width (FWHM).

where $\mathcal{E}(\mathbf{r},t)$ is the electric field, $\mathcal{P}_{\text{L}}(\mathbf{r},t)$ is the linear component of the polarization density, which is governed by the medium dispersion, and $\mathcal{P}_{\text{NL}} = 4\chi^{(3)}\mathcal{E}^3$ is the nonlinear component of the polarization density, which is assumed to be nondispersive. Bringing the linear term from the right-hand side to the left-hand side of (23.5-18) and rewriting the equation in the Fourier domain, we obtain

$$\left[\nabla^2 + \beta^2(\omega)\right] E = -\mu_o \omega^2 P_{\text{NL}} \tag{23.5-19}$$

where $\beta(\omega)$ is the propagation constant in the linear medium and $E = E(\mathbf{r},\omega)$ and $P_{\text{NL}} = P_{\text{NL}}(\mathbf{r},\omega)$ are Fourier transforms of $\mathcal{E}(\mathbf{r},t)$ and $\mathcal{P}_{\text{NL}}(\mathbf{r},t)$, respectively. In the absence of nonlinearity, (23.5-19) reproduces the Helmholtz equation (5.3-16).

We consider a plane-wave optical pulse traveling in the z direction with central angular frequency ω_0 and central wavenumber $\beta_0 = \beta(\omega_0) = \omega_0/c$,

$$\mathcal{E} = \text{Re}\{\mathcal{A}(z,t)\exp[j(\omega_0 t - \beta_0 z)]\}, \tag{23.5-20}$$

and assume that the complex envelope $\mathcal{A}$ is a slowly varying function of t and z (in comparison with the period $2\pi/\omega_o$ and the wavelength $2\pi/\beta_0$, respectively).

Using three assumptions: (1) slowly varying envelope, (2) weak dispersion, and (3) small nonlinear effect, it will be shown below that the envelope $\mathcal{A}(z,t)$ satisfies the differential equation

$$\boxed{\frac{D_\nu}{4\pi}\frac{\partial^2 \mathcal{A}}{\partial t^2} + \gamma|\mathcal{A}|^2\mathcal{A} + j\left(\frac{\partial}{\partial z} + \frac{1}{v}\frac{\partial}{\partial t}\right)\mathcal{A} = 0\,,} \tag{23.5-21}$$

Nonlinear SVE Wave Equation

where $v = 1/\beta'$ is the group velocity, $D_\nu = 2\pi\beta''$ is the dispersion coefficient, β' and β'' are the first and second derivatives of $\beta(\omega)$ with respect to ω at $\omega = \omega_0$, and γ is given by (23.5-16). For a linear medium $\gamma = 0$, and the linear SVE wave equation (23.3-24) is reproduced.

☐ ***Derivation of the Nonlinear SVE Wave Equation.** We begin with the nonlinear Helmholtz equation (23.5-19). Substituting $E = A(z,\omega-\omega_0)\exp(-j\beta_0 z)$ as well as $P_{\text{NL}} = A_{\text{NL}}(z,\omega-\omega_0)\exp(-j\beta_0 z)$, and defining $\Omega = \omega - \omega_0$, we obtain

$$\left[\frac{\partial^2}{\partial z^2} + \beta^2(\omega)\right][A(z,\omega)\exp(-j\beta_0 z)] = -\mu_o\omega^2 A_{\text{NL}}(z,\Omega)\exp(-j\beta_0 z). \tag{23.5-22}$$

We now simplify (23.5-22) using a number of approximations:

- Since $\omega \approx \omega_0$, the ω^2 factor on the right-hand side of (23.5-22) can be approximated by ω_0^2.
- When the SVE approximation $(d^2/dz^2)[A\exp(-j\beta_0 z)] \approx [-j2\beta_0 dA/dz - \beta_0^2 A]\exp(-j\beta_0 z)$ is applied, (23.5-22) becomes

$$[-j2\beta_0 d/dz]\,A + \left[\beta^2(\omega_0+\Omega) - \beta_0^2\right]A = -\mu_o\omega_0^2 A_{\text{NL}}. \tag{23.5-23}$$

- Assuming weak dispersion, $\beta^2(\omega_0+\Omega)-\beta_0^2 \approx 2\beta_0\left[\beta(\omega_0+\Omega)-\beta_0\right]$. Further assuming a three-term Taylor-series expansion, $\beta(\omega_0+\Omega)=\beta_0+\beta'\Omega+\frac{1}{2}\beta''\Omega^2$, (23.5-23) becomes

$$-j2\beta_0\frac{dA}{dz}+2\beta_0\left(\Omega\beta'+\tfrac{1}{2}\Omega^2\beta''\right)A=-\mu_o\omega_0^2A_{\mathrm{NL}}. \tag{23.5-24}$$

- Since $\mathcal{P}_{\mathrm{NL}}=4\chi^{(3)}\mathcal{E}^3$, $\mathcal{P}_{\mathrm{NL}}$ contains components near the frequencies ω_0 and $3\omega_0$. Retaining only the term near ω_0, we write $\mathcal{P}_{\mathrm{NL}}=\mathrm{Re}\{\mathcal{A}_{\mathrm{NL}}(z,t)\exp[j(\omega_0t-\beta_0z)]\}$, where $\mathcal{A}_{\mathrm{NL}}(z,t)$ is a slowly varying envelope. Using (23.5-20) and (22.3-3a), it follows that

$$A_{\mathrm{NL}}=3\chi^{(3)}\left|A\right|^2A. \tag{23.5-25}$$

We then transform (23.5-24) back to the time domain, using the fact that $j\Omega A(z,\Omega)$ and $-\Omega^2A(z,\Omega)$ are equivalent to $\partial A/\partial t$ and $\partial^2A/\partial t^2$. Finally, using (23.5-25), we obtain the nonlinear SVE wave equation (23.5-21).

This result may also be obtained if we assume that the nonlinear medium is approximately linear with a propagation constant $\beta(\omega)+\Delta\beta$, where $\Delta\beta=(\omega_0/c_o)n_2I$. We take the intensity $I=\left|A\right|^2/2\eta$ to be sufficiently slowly varying so that it may be regarded as time-independent. The Fourier analysis that led to the differential equation (23.3-24) for the linear medium is then simply modified by adding a term proportional to $\Delta\beta A$. This contribution produces the additional term $\gamma\left|A\right|^2A$, whereupon (23.5-21) emerges. ■

The Nonlinear Schrödinger Equation

Equation (23.5-21) must be satisfied by the complex envelope $\mathcal{A}(z,t)$ of a plane-wave optical pulse traveling in the z direction in an extended nonlinear dispersive medium, with group velocity v, dispersion parameter β'', and nonlinear coefficient γ. As previously noted, a solitary-wave solution is possible if $\beta''<0$ (i.e., the medium exhibits negative GVD) and $\gamma>0$ (i.e., the optical Kerr coefficient $n_2>0$).

It is convenient to rewrite (23.5-21) in terms of dimensionless variables by normalizing the time, distance, and amplitude to the scales τ_0, $2z_0$, and A_0, respectively:

- τ_0 is the pulse width
- $z_0=\tau_0^2/2\beta''$ is the dispersion length of the linear dispersive medium for this pulse width
- $A_0=(-\beta''/\gamma)^{1/2}/\tau_0$ is the pulse peak amplitude that satisfies the soliton condition (23.5-15).

Using a retarded frame of reference, and defining the dimensionless variables

$$\mathrm{t}=\frac{t-z/v}{\tau_0},\quad \mathrm{z}=\frac{z}{2z_0},\quad \psi=\frac{A}{A_0}, \tag{23.5-26}$$

the nonlinear SVE wave equation in (23.5-21) becomes

$$\boxed{\frac{1}{2}\frac{\partial^2\psi}{\partial\mathrm{t}^2}+\left|\psi\right|^2\psi+j\frac{\partial\psi}{\partial\mathrm{z}}=0,} \tag{23.5-27}$$

Nonlinear Schrödinger Equation

which is recognized as the nonlinear Schrödinger equation.

Fundamental Soliton

The simplest solitary-wave solution of (23.5-27) is obtained by assuming a space–time separable function of the form $\psi(\mathrm{z},\mathrm{t})=\mathcal{T}(\mathrm{t})\exp[j\mathcal{Z}(\mathrm{z})]$, where $\mathcal{T}(\mathrm{t})$ and $\mathcal{Z}(\mathrm{z})$ are

real functions. By direct substitution in (23.5-27), and using a separation-of-variables approach, this leads to two differential equations: $\mathcal{Z}'(\mathrm{z}) = \vartheta$ and $\mathcal{T}''(\mathrm{t}) = 2(\vartheta - \mathcal{T}^2)\mathcal{T}$, where ϑ is a constant. Assuming that $\mathcal{T} = \mathcal{T}' = 0$ at $|\mathrm{t}| \to \infty$, and $\mathcal{T} = 1$ and $\mathcal{T}' = 0$ at $\mathrm{t} = 0$ (the pulse peak), these ordinary differential equations may be solved by direct integration to yield $\mathcal{T}(\mathrm{t}) = \mathrm{sech}(\mathrm{t})$ and $\mathcal{Z}(\mathrm{z}) = \frac{1}{2}\mathrm{z}$. The normalized amplitude is then given by

$$\psi(\mathrm{z}, \mathrm{t}) = \mathrm{sech}(\mathrm{t})\ \exp(j\mathrm{z}/2)\,, \tag{23.5-28}$$

and this solution is called the **fundamental soliton**. It corresponds to an envelope

$$\mathcal{A}(z,t) = A_0\,\mathrm{sech}\left(\frac{t - z/v}{\tau_0}\right)\,\exp\left(jz/4z_0\right) \tag{23.5-29}$$

Fundamental Soliton

that travels at velocity v without altering its shape. This solution is achieved if the incident pulse at $z = 0$ is

$$\mathcal{A}(0,t) = A_0\,\mathrm{sech}\,(t/\tau_0). \tag{23.5-30}$$

Higher-Order Soliton

The fundamental soliton is but one of a family of solutions of the nonlinear Schrödinger equation with solitary properties. Consistent with the initial pulse $\psi(0, \mathrm{t}) = N\mathrm{sech}(\mathrm{t})$, where N is an integer, is a solution known as the ***N*-soliton wave**. Such a wave propagates as a periodic function of z with period $z_p = \pi/2$, called the **soliton period**, which corresponds to a physical distance $z_p = \pi|z_0| = (\pi/2)\tau_0^2/|\beta''|$. At $z = 0$ the envelope $\mathcal{A}(0,t)$ is then a hyperbolic-secant function with peak amplitude NA_0, which is N times greater than that of the fundamental soliton. As the pulse travels, it contracts initially, then splits into distinct pulses that subsequently merge, and eventually it reproduces the initial pulse at $z = z_p$ and multiples thereof.

The $N = 2$ soliton serves as an example. It has a normalized amplitude given by

$$\psi(\mathrm{z}, \mathrm{t}) = 4\,\frac{\cosh 3\mathrm{t} + 3e^{4j\mathrm{z}}\cosh \mathrm{t}}{\cosh 4\mathrm{t} + 4\cosh 2\mathrm{t} + 3\cos 4\mathrm{z}}\,e^{j\mathrm{z}/2}, \tag{23.5-31}$$

whose magnitude is illustrated in Fig. 23.5-9.

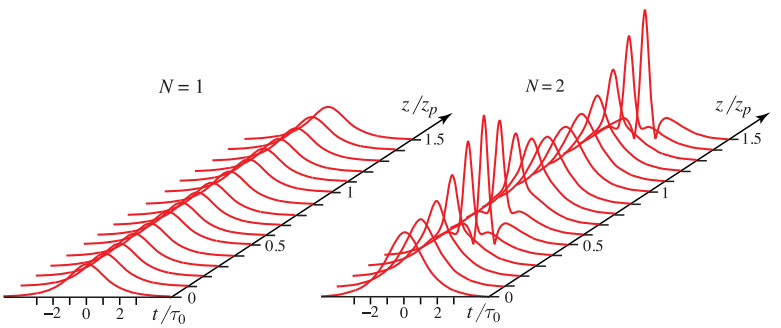


Figure 23.5-9 Propagation of the fundamental ($N = 1$) soliton and the $N = 2$ soliton.

The periodic compression and expansion of the multi-soliton wave is accommodated by a periodic imbalance between the pulse compression, which results from the chirping introduced by self-phase modulation, and the pulse spreading caused by group velocity dispersion. The initial compression has been used for the generation of subpicosecond pulses.

Soliton-Soliton Interaction

When two solitons separated by some time delay are launched into the nonlinear medium, their shape and time separation are altered as if they experience attractive or repulsive forces pulling them together or separating them. For example, two identical separated fundamental solitons are initially attracted as they travel through the medium and their time separation is reduced until they collapse into a single pulse, whereupon they experience repulsive forces that separate them again into two pulses. The process is repeated periodically with a period

$$L_p = \pi \exp(T/2\tau_0)\, z_0, \tag{23.5-32}$$

where T is the initial center-to-center separation, τ_0 is the width of the individual soliton, and z_0 is the GVD dispersion length. This result can be obtained by solving the nonlinear Schrödinger equation with the appropriate boundary condition. As an example, if $T = 10\tau_0$, so that the pulses are well separated and only their tails interact, $L_p \approx 466 z_0$, which is quite large. However, this effect can be significant in long optical fibers since it can impose restrictions on optical fiber communication systems that use solitons to represent bits, as described in Sec. 25.2E.

EXAMPLE 23.5-2. ***Solitons in Optical Fibers.*** Ultrashort solitons have been generated in glass fibers at wavelengths in the anomalous dispersion region ($\lambda_o > 1.3\ \mu$m), where the GVD is negative. They were first observed in a 700-m single-mode silica-glass fiber using pulses from a mode-locked laser operating at a wavelength $\lambda_o = 1.55\ \mu$m. The pulse shape closely approximated a hyperbolic-secant function with $\tau_0 = 4$ ps (corresponding to $\tau_{\text{FWHM}} = 1.76\,\tau_0 = 7$ ps). At this wavelength the dispersion coefficient is $D_\lambda = 16$ ps/km-nm (see Fig. 10.3-5), corresponding to $\beta'' = D_\nu/2\pi = (-\lambda_o^2/c_o)D_\lambda/2\pi \approx -20$ ps^2/km. The refractive index n = 1.45 and the nonlinear coefficient $n_2 = 3.19 \times 10^{-20}$ m^2/W, corresponding to $\gamma = (\pi/\lambda_o)n_2/\eta = 2.48 \times 10^{-16}$ m/V^2 (where $\eta = \eta_o/n = 260\ \Omega$). The amplitude $A_0 = (|\beta''|/\gamma)^{1/2}/\tau_0 \approx 2.25 \times 10^6$ V/m, corresponding to an intensity $I_0 = A_0^2/2\eta \approx 10^6$ W/cm^2. If the fiber area is $10\ \mu$m^2, this corresponds to a power of about 100 mW. The soliton period is $z_p = \pi z_0 = \pi\tau_0^2/2\,|\beta''|$ = 1.26 km.

Soliton Generation and Maintenance

To excite the fundamental soliton, the input pulse must have the hyperbolic-secant profile with the appropriate amplitude-width product $A_0\tau_0$, as specified in (23.5-15). A lower value of this product will excite an ordinary optical pulse, while a higher value will excite the fundamental soliton, or possibly a higher-order soliton, with the remaining energy diverted into a spurious ordinary pulse. When the initial pulse has a different profile or is chirped, the resulting pulse can, under certain conditions, evolve into a fundamental or higher-order soliton after a distance equal to a few soliton periods.

If the medium is lossy, the pulse power is gradually dissipated so that the nonlinear effect becomes weaker and dispersive effects dominate, which leads to pulse broadening and loss of the soliton nature of the pulse. In optical fibers, this problem may be mitigated by the use of distributed Raman amplification (see Secs. 15.3D and 22.3B) to overcome absorption and scattering losses. Lumped amplification can also be effective if the amplifier spacing is well within the soliton period z_p.

Because of their unique ability to maintain their shape and width over long propagation distances, optical solitons have the potential for transmitting digital data through optical fibers at higher rates, and for longer distances, than is currently possible with linear optics (see Sec. 25.2E). Optical solitons with durations of a few tens of picoseconds have been successfully transmitted through many thousands of kilometers of optical fiber.

Soliton Lasers

Optical-fiber lasers have also been used to generate picosecond solitons. One version is a single-mode fiber in a-ring resonator configuration (Fig. 23.5-10). The fiber is a combination of an erbium-doped fiber amplifier (see Sec. 15.3C) and an undoped optical fiber that provides the pulse shaping and soliton action. Pulses are generated by making use of a phase modulator to achieve mode locking. An integrated version makes use of a laser-diode pump, such as InGaAsP, and an integrated-photonic phase modulator.

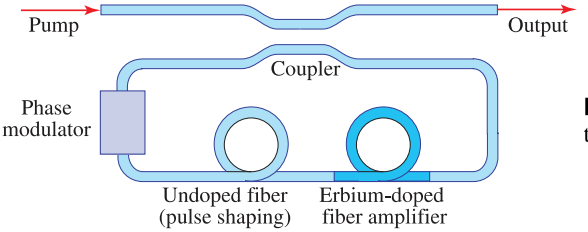


Figure 23.5-10 An optical-fiber soliton laser.

Dark Solitons

A dark soliton is a short-duration dip in the intensity of an otherwise continuous wave of light. Dark solitons have properties similar to the "bright" solitons described earlier, but can be generated in the normal dispersion region ($\lambda_o < 1.3\ \mu$m in silica optical fibers). They exhibit robust features that can be garnered for optical switching.

Analogy Between Temporal and Spatial Solitons

The optical solitons described in Sec. 23.5B are analogous to the spatial solitons (self-guided beams) described in Sec. 22.3B. Spatial solitons are monochromatic waves that are localized spatially in the transverse plane. They travel in a nonlinear medium without altering their spatial distribution, as a result of a balance between diffraction and spatial self-phase modulation in accordance with the nonlinear Schrödinger equation,

$$-\frac{\lambda}{4\pi}\frac{\partial^2 \mathcal{A}}{\partial x^2} + \gamma|\mathcal{A}|^2\mathcal{A} + j\frac{\partial \mathcal{A}}{\partial z} = 0, \tag{23.5-33}$$

Nonlinear Beam Diffraction

where $\gamma = \pi n_2/\lambda\eta_o$ and n_2 is the optical Kerr coefficient. Equation (23.5-33) is equivalent to (22.3-11).

The nonlinear Schrödinger equation that describes temporal solitons in nonlinear dispersive media (23.5-21) may be rewritten in the moving frame ($t' = t - z/v$, $z' = z$) as

$$\frac{D_\nu}{4\pi}\frac{\partial^2 \mathcal{A}}{\partial t^2} + \gamma|\mathcal{A}|^2\mathcal{A} + j\frac{\partial \mathcal{A}}{\partial z} = 0, \tag{23.5-34}$$

Nonlinear Pulse Dispersion

where $\gamma = \pi n_2/\lambda\eta_o$. This is identical to (23.5-33) with time t playing the role of the transverse spatial coordinate x, and the dispersion coefficient $-D_\nu$ (which governs pulse dispersion) playing the role of the wavelength λ (which governs beam diffraction). It is therefore evident that temporal solitons are formal analogs of spatial solitons. In fact the term *soliton* refers to generic solutions of the nonlinear Schrödinger equation, describing pulses that propagate without change; they may be temporal or spatial.

Spatiotemporal Solitons and Light Bullets

A spatiotemporal soliton is a combined temporal and spatial soliton, i.e., a pulsed beam that maintains its spatial *and* temporal profiles as it travels through a nonlinear medium exhibiting the optical Kerr effect (see Fig. 23.5-11). In this case, the temporal broadening associated with negative (anomalous) dispersion and the spatial spreading resulting from diffraction are simultaneously compensated for by self-phase modulation and self-focusing that ensue from a positive nonlinear optical Kerr effect. The partial differential equation describing these three phenomena is a combination of (23.5-33) and (23.5-34),

$$-\frac{\lambda}{4\pi}\nabla_T^2\mathcal{A} + \frac{D_\nu}{4\pi}\frac{\partial^2\mathcal{A}}{\partial x^2} + \gamma|\mathcal{A}|^2\mathcal{A} + j\frac{\partial\mathcal{A}}{\partial z} = 0. \tag{23.5-35}$$

Nonlinear Diffraction & Dispersion

A necessary condition for generating spatiotemporal solitons is the equality of the dispersion length $|z_0| = \pi\tau_0^2/|D_\nu|$ and the diffraction length $z_0 = \pi W_0^2/\lambda$ so that $\tau_0/W_0 = (\lambda/|D_\nu|)^{1/2}$.

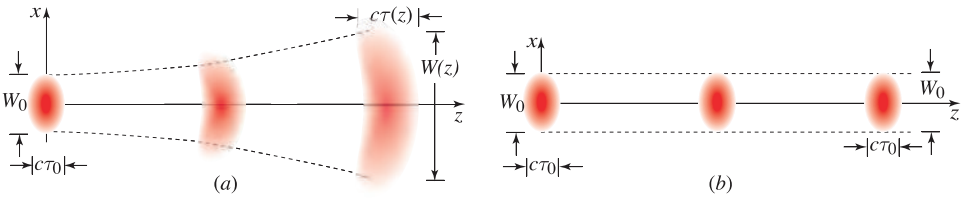


Figure 23.5-11 (*a*) Spatial and temporal spreading of a pulsed beam as a result of propagation in a linear dispersive medium. (*b*) A spatiotemporal soliton is a pulsed beam that maintains its spatial and temporal profiles as it propagates in a nonlinear medium.

*C. Supercontinuum Light

Supercontinuum light has an ultrabroad continuous spectrum and is of high brightness. **Supercontinuum generation (SCG)** is implemented by transmitting an ultrashort optical pulse of high peak power (a pump) through a nonlinear medium with special dispersive properties; examples of such media are dispersion-shifted, dispersion-flattened microstructured, and photonic-crystal optical fibers (PCFs). Supercontinuum light sources with spectra stretching from the mid infrared to the extreme ultraviolet have been demonstrated.

Depending on the details of the source, several nonlinear mechanisms, including self-phase modulation (SPM), stimulated Raman scattering (SRS), four-wave mixing

(FWM), and the soliton self-frequency shift (SSFS), may contribute individually or jointly to SCG. These nonlinear effects are sensitive to the sign of the medium dispersion at the central wavelength λ_0 of the pump pulse and to the relative location of the zero-dispersion wavelength λ_{ZD} of the medium. The widest SCG spectra are obtained when λ_0 is close to λ_{ZD}. It was the availability of nonlinear PCFs with λ_{ZD} close to the wavelength of the Ti:sapphire laser that first made SCG practical.

A brief description of the principal nonlinear mechanisms that contribute to SCG follows; Fig. 23.5-12 provides schematic illustrations of these processes.

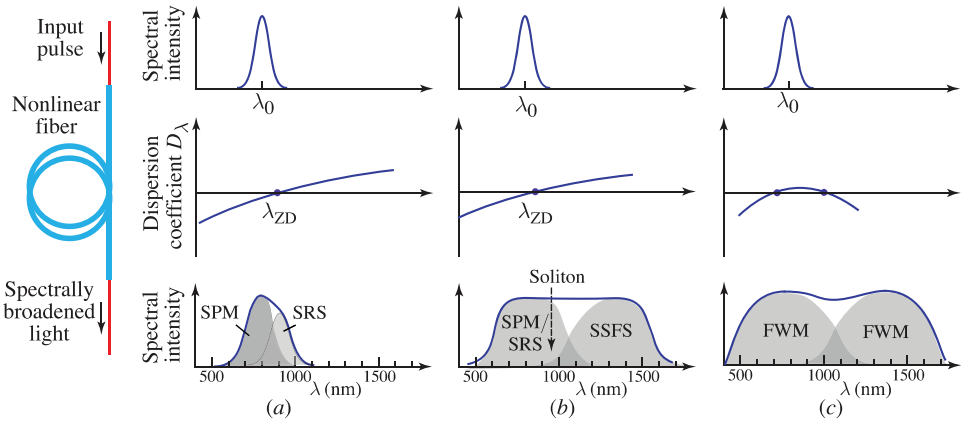


Figure 23.5-12 Principal nonlinear mechanisms for supercontinuum generation (SCG) via spectral broadening of an ultrashort pulse transmitted through a nonlinear dispersive fiber. (*a*) Self-phase modulation (SPM) combined with stimulated Raman scattering (SRS). (*b*) Soliton self-frequency shift (SSFS). (*c*) Four-wave mixing (FWM).

- **Self-phase modulation (SPM)** is the principal mechanism for producing SCG in nonlinear optical fibers with normal dispersion ($D_\lambda < 0$) at the pump central wavelength λ_0, since in this case solitons cannot be formed. As discussed in Sec. 23.5A, SPM results in pulse chirping, which in turn causes spectral broadening. A chirp coefficient a corresponds to spectral broadening by the factor $\sqrt{1+a^2}$. For a medium of length L and optical Kerr coefficient n_2, the chirp parameter is $a = L/z_{NL}$, where $z_{NL} = (2n_2 I_0 k_0)^{-1}$ is the nonlinear characteristic length of the Kerr medium and I_0 is the peak pulse intensity.
- **Stimulated Raman scattering (SRS)** broadens the spectral distribution further, toward longer wavelengths, since it results in a frequency downshift (see Secs. 14.5C, 15.3D, and 16.3C).
- The **soliton self-frequency shift (SSFS)** originates from intrapulse stimulated Raman scattering (SRS). When λ_0 is close to λ_{ZD} the combined SPM/SRS broadens the spectrum into the anomalous region, creating conditions suitable for soliton formation. Optical solitons generally experience a downshift of their carrier frequency, toward longer wavelengths, which increases with pump power.
- **Four-wave mixing (FWM)** can also contribute to SCG. In a microstructured fiber that has two widely separated, zero-dispersion wavelengths, with λ_0 lying between them, the dominant nonlinear mechanisms for spectral broadening are SPM and FWM. The SPM process broadens the pump pulse, enabling the phase-matching conditions for FWM to be met. This generates new light at both lower and higher frequencies, yielding SCG with double-peaked spectra. With sufficient broadening, the two FWM peaks may merge into a single flat distribution.

EXAMPLE 23.5-3. ***High-Energy Solitons Generated in a Photonic-Crystal Rod.*** The soliton pulse energy $E = A_{\text{eff}} \int I(t)\, dt$ can be maximized by making use of a photonic-crystal rod with large effective mode area A_{eff}. By virtue of the soliton condition specified in (23.5-15) and (23.5-16), the soliton intensity is given by $I(t) = (|A_0|^2/2\eta)\,\text{sech}^2(t/\tau_0)$, where $|A_0|^2/2\eta = |\beta''|\,\lambda_0/2\pi n_2 \tau_0^2$. Since $\int_{-\infty}^{\infty} \text{sech}^2(t/\tau_0)\, dt = 2\tau_0$, the soliton pulse energy can be written as

$$E = \frac{A_{\text{eff}}\,|\beta''|\,\lambda_0}{\pi n_2 \tau_0}. \tag{23.5-36}$$

Soliton Energy

A 36-cm-long photonic-crystal rod with an effective mode area $A_{\text{eff}} = 2300\ \mu\text{m}^2$, driven by a mode-locked, Er^{3+}-doped fiber laser that provides a sequence of 360-fs, 500-nJ pulses at a wavelength of $\lambda_o = 1550$ nm and at a repetition rate of 1 MHz, generates optical solitons via SSFS. The output of the rod is a train of 65-fs, 67-nJ optical solitons with a center wavelength of $\lambda_0 = 1675$ nm, a repetition rate of 1 MHz, and an average power of tens of mW. This source is useful for three-photon fluorescence microscopy (Sec. 14.5B).

*D. High-Harmonic Generation and Attosecond Optics

The previous sections of this chapter have considered ultrafast nonlinear phenomena in the context of a dielectric medium in which the motion of bound electrons is characterized by a weakly nonlinear $\mathcal{P}$–$\mathcal{E}$ relation (Fig. 22.1-1). Equation (22.1-1) provides a Taylor-series expansion for the polarization density $\mathcal{P}$ in terms of increasing powers of the electric field $\mathcal{E}$ that are responsible for second- and third-harmonic generation, as well as a host of other parametric processes. When the optical field exceeds a certain strength, however, this expansion is no longer viable and the light–matter interaction exhibits new physical manifestations, such as atomic ionization (see Sec. 14.1A). This domain is known as **extreme nonlinear optics**.

In **high-harmonic generation** (**HHG**), an infrared ultrafast pulse from a focused laser beam serves to ionize a gas atom, creating a free electron in the process. The electron is accelerated by the exciting laser field, which imparts kinetic energy to it before it is recaptured by its parent atom. The ensuing atom–electron recombination generates a high-energy photon in the form of a burst of radiation with sub-femtosecond structure. This same process takes place in a collection of atoms, which radiate coherently since they are entrained by the waveform of the exciting laser. The discrete spectrum of the emitted light comprises a frequency comb (Sec. 16.4E) that can contain hundreds of harmonics of the exciting laser frequency, reaching into the extreme-ultraviolet (EUV) region. The result is a tightly collimated beam of light that contains wavelengths far shorter than that of the exciting laser. High-harmonic generation serves to convert infrared (IR) light into extreme-ultraviolet (EUV) radiation and to generate **attosecond light pulses**, as illustrated schematically in Fig. 23.5-13. At high gas pressures, the HHG spectrum can contain thousands of harmonics of the exciting laser frequency and extend well into the soft-X-ray (SXR) region.

Emissions from a single atom. A simplified semiclassical model that describes the generation of HHG, known as the **recollisional model**, is illustrated in Fig. 23.5-14. A gas atom is modeled as a single electron in a potential well [Fig. 23.5-14(*a*)]. The exciting laser field at the location of the atom is taken to be monochromatic with angular frequency ω_0 and period $T = 2\pi/\omega_0$, and to be linearly polarized in the x direction. The photon energy of the laser $\hbar\omega_0$ is much smaller than the ionization energy of the atom W so ordinary photon absorption does not take place.
The HHG generation process follows three steps, in sequence:

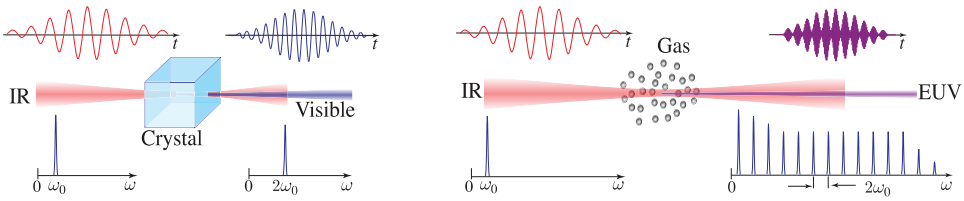


Figure 23.5-13 Comparison between (a) second-harmonic generation (SHG) and (b) high-harmonic generation (HHG), a process that generates pulses of EUV light on an attosecond timescale.

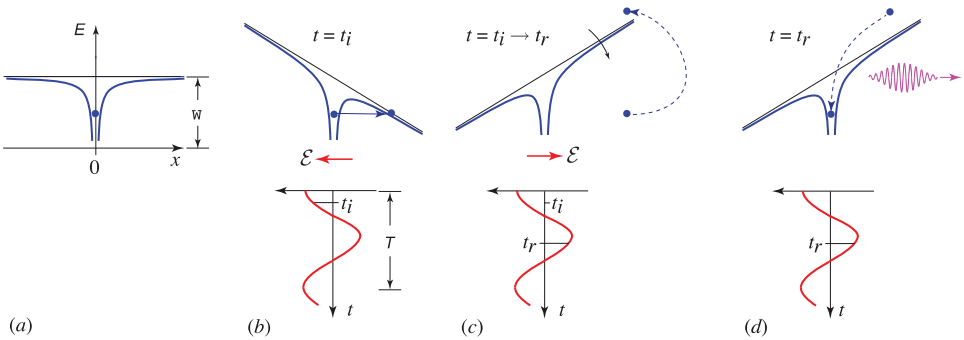


Figure 23.5-14 Simplified three-step recollisional model for HHG. (a) A gas atom is modeled as a single electron in the ground state of a potential well with ionization energy $\mathtt{W}$. (b) An applied optical field $\mathcal{E}(t)$ from an exciting laser alters the potential well and causes the electron to tunnel into free space at time t_i. (c) The free electron is accelerated in the $+x$ direction by the negative optical field but reverses direction when the optical field becomes positive, returning back to the ionized atom with increased kinetic energy E_k at time t_r. (d) The electron recombines with the ion and radiates an EUV photon of energy $\mathtt{W} + E_k$ that takes the form of a chirped pulse of radiation with sub-femtosecond structure, as illustrated. Emissions from all atoms illuminated by the exciting laser pulse add coherently.

- *Step 1.* During the first quarter of the optical cycle [Fig. 23.5-14(b)], the optical field tilts the atomic potential well, thereby converting it into a potential barrier. The electron tunnels through the barrier at time t_i. Different atoms tunnel through at different times.
- *Step 2.* The liberated electron is accelerated in free space by the exciting field, in the $+x$ direction away from the ionized atom [Fig. 23.5-14(c)]. When the field reverses direction during the second and third quarters of the excitation cycle, the electron reverses direction and is accelerated in the $-x$ direction, back toward the ionized atom. On its arrival at the atom, the atomic potential is tilted in the opposite direction and the electron is endowed with increased kinetic energy E_k as a consequence of its acceleration via the ponderomotive force.
- *Step 3.* When the electron collides with the ionized atom at time t_r, recombination occurs and the electron falls into the potential well [Fig. 23.5-14(d)]. In the process, a photon is emitted with energy $\hbar\omega = \mathtt{W} + E_k \gg \hbar\omega_0$, in the form of a burst of radiation with sub-femtosecond structure. The emitted EUV light has a frequency far greater than that of the exciting IR light.

The preceding three steps are repeated for each cycle of the exciting laser field, resulting in a periodic stream of radiation bursts. However, the process can also be initiated on the half optical cycle opposite to that portrayed above, in which case the

liberated electron initially accelerates in the $-x$ direction before turning around and reuniting with the ionized atom. Symmetry dictates that both alternatives have equal probability of occurrence in the collection of gas atoms, so that radiation pulses with attosecond structure are generated twice per cycle, i.e., with period $T/2 = \pi/\omega_0$. Such a periodic sequence of events exhibits a discrete spectrum whose frequencies are spaced by $2\pi/(T/2) = 2\omega_0$ and whose amplitudes are determined from a Fourier-series expansion (see Appendix A). The symmetry of the atom and the electric field gives rise to components at odd harmonics of the exciting laser frequency that are separated by $2\omega_0$, as schematically illustrated in Fig. 23.5-13(b).

Field to electron energy transfer. The energy imparted to the electron by the exciting field during its foray away from its parent ion determines the characteristics of the HHG. The kinetic energy acquired by an electron in an ionization–recombination cycle may be calculated by determining its free-space trajectory. The optical field of the exciting laser, $\mathcal{E}(t) = \mathcal{E}_0 \cos\omega_0 t$ at $x = 0$, exerts a force $e\mathcal{E} = m_0 \mathrm{a}$ on the electron, resulting in an acceleration in the $+x$ direction given by $\mathrm{a}(t) = (e\mathcal{E}_0/m_0)\cos\omega_0 t$, where $-e$ and m_0 are the charge and mass of the electron, respectively. If the ionization occurs at time t_i, and the electron emerges with zero velocity [$v(t_i) = 0$] at position $x = 0$, the following equations govern the electron's velocity $v(t) = \int \mathrm{a}(t)\,dt$, kinetic energy $E_k(t)$, and position $x(t) = \int v(t)\,dt$ at time t:

$$v(t) = v_0[\sin\omega_0 t - \sin\omega_0 t_i], \qquad E_k(t) = \tfrac{1}{2} m_0 v^2(t), \tag{23.5-37}$$

$$x(t) = \tfrac{1}{2} x_0 \left[\cos\omega_0 t_i - \cos\omega_0 t - \omega_0 (t - t_i)\sin\omega_0 t_i\right], \tag{23.5-38}$$

where $v_0 = e\mathcal{E}_0/m_0\omega_0$ and $x_0 = 2e\mathcal{E}_0/m_0\omega_0^2$. These equations establish the trajectory of the electron and the dependence of its kinetic energy $E_k = E_k(t_r)$ on the ionization time t_i and the recombination time t_r.

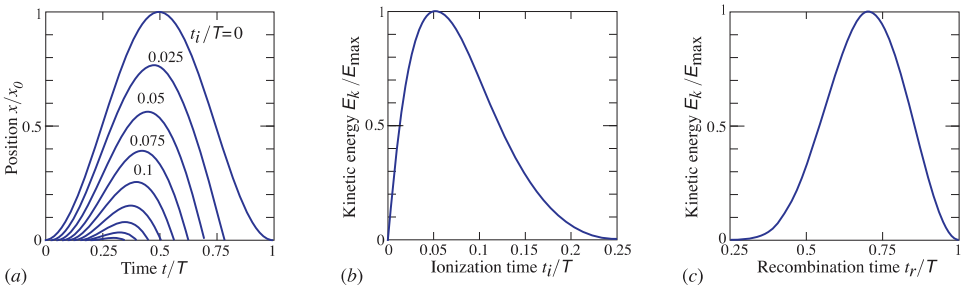


Figure 23.5-15 (a) Normalized trajectories $x(t)/x_0$ of liberated electrons for various normalized ionization times t_i/T within the first quarter of the laser cycle. Each trajectory begins at time t_i/T and terminates at the corresponding arrival time t_r/T. (b) Dependence of the normalized kinetic energy on the normalized ionization time t_i/T. (c) Dependence of the normalized kinetic energy on the normalized recombination time t_r/T.

Based on numerical solutions to these equations, which are displayed in Fig. 23.5-15, the following observations emerge:

- From Fig. 23.5-15(a) it is clear that electrons freed at times t_i in the first quarter of the laser cycle ($0 < t_i < T/4$) arrive at the parent atom at times t_r in the last three quarters of the cycle ($T/4 < t_r < T$). The earlier an electron is freed in the first quarter, the later it arrives in the last three quarters. In particular, an electron freed at the very beginning of the cycle ($t_i = 0$) arrives at the very end of the

cycle ($t_r = T$), and its excursion away from the parent atom is the greatest (a distance x_0). Electrons freed at times t_i in the second quarter of the exciting laser cycle ($T/4 < t_i < T/2$) drift away and never return to the parent atom.

- From Fig. 23.5-15(b) it is evident that an electron ionized at $t_i = 0$ returns with zero net kinetic energy ($E_k = 0$). As t_i increases within the first quarter of the laser cycle ($0 < t_i < T/4$), the kinetic energy of the returning electron increases monotonically, reaching a peak value $E_{\max} = 0.794\, m_0 v_0^2$ (corresponding to a velocity $1.26\, v_0$) for $t_i = 0.05T$, and then decreases to zero at $t_i = T/4$. The maximum kinetic energy $E_{\max}$ depends on the laser intensity $I = \mathcal{E}_0^2/2\eta_o$ and laser wavelength λ_o in accordance with

$$E_{\max} = \tau_0 \lambda_o^2 I, \qquad \tau_0 = 1.588 \frac{\eta_o e^2}{4\pi^2 m_0 c_o^2} \approx 4.739 \times 10^{-24}\ \text{s}, \tag{23.5-39}$$

  where η_o is the free-space impedance. The maximum kinetic energy is proportional to the laser intensity and to the square of its wavelength.
- Figure 23.5-15(c) demonstrates that the electron's kinetic energy E_k, as a function of the recombination time t_r, which is also the photon emission time, is greater than zero over the last three quarters of the exciting-laser cycle and exhibits a maximum value at $t_r \approx 0.7T$. The full width of this curve is therefore $\approx \frac{3}{4}T$, corresponding to ≈ 2 fs for an exciting wavelength of 800 nm.

Emissions from a collection of atoms. The characteristics of the generated HHG beam are determined by the collective recombinations of many electrons with their parent ionized atoms over a range of times within the same laser cycle. Figure 23.5-15(c) demonstrates that the kinetic energy E_k of an arriving electron lies between 0 and $E_{\max}$, indicating that the corresponding energy of its companion emitted photon lies in the range between $\mathbb{W}$ and $\mathbb{W} + E_{\max}$. If the ionization probability were constant and independent of the ionization time t_i within the first quarter cycle, each generated pulse would comprise a coherent superposition of recombination emissions from the various atoms. Each such pulse would have an overall temporal duration $< T/2$ and would be chirped by virtue of the dependence of the kinetic energy $E_k(t_r)$ on the recombination time t_r, as depicted in Fig. 23.5-15(c). The frequency of the emitted pulse, $\omega(t_r) = [\mathbb{W} + E_k(t_r)]/\hbar$, is up-chirped for $t_r < 0.7\,T$ and down-chirped for $t_r > 0.7\,T$. The corresponding spectral width is $\Delta\nu = E_{\max}/h = M/T$, where $M = E_{\max}/h\nu_0$ is the number of harmonics within the spectral band.

Pulse-compression techniques could be used to eliminate the attendant chirp and render the pulse transform-limited (Sec. 23.2), in which case its temporal width would be compressed to a value on the order of T/M. As an example, (23.1-10) reveals that a transform-limited Gaussian pulse of spectral width $\Delta\nu$ (FWHM) has a corresponding temporal width $\tau_{\text{FWHM}} = 0.44/\Delta\nu = 0.44T/M$, which represents a small fraction of the laser period T. The technology of attosecond optics offers methods for unchirping HHG pulses, as well as for using suitable optical gating to isolate individual attosecond pulses from such a pulse train.

EXAMPLE 23.5-4. ***HHG Attosecond Pulses in Ar Gas.*** An ultrafast Ti:sapphire laser operated at a wavelength of $\lambda_o = 800$ nm emits femtosecond pulses that are amplified and focused to an intensity $I = 5 \times 10^{14}$ W/cm^2. The infrared pulses impinge on a gas of argon atoms (ionization energy $\mathbb{W} = 15.78$ eV) to produce HHG. The simplified three-step HHG model yields a maximum electron excursion $x_0 = 2e\mathcal{E}_0/m_0\omega_0^2 = 1.94$ nm and a maximum electron kinetic energy $E_{\max} = \tau_0\lambda_o^2 I \approx 94.8$ eV [see (23.5-39)]. The maximum energy of an emitted photon is thus $\mathbb{W} + E_{\max} = 110.6$ eV, which corresponds to a wavelength of ≈ 11.2 nm in the EUV. Since the exciting photon energy is $\hbar\omega_0 = 1.55$ eV, the number of harmonics generated is $M = E_{\max}/\hbar\omega_0 \approx 61$. The period of

the laser cycle is $T = 2.67$ fs so optimal pulse compression would yield a train of pulses of duration $T/61 \approx 43.8$ attoseconds.

For very large values of the laser intensity, the probability of ionization is not constant and independent of the ionization time t_i within the first quarter cycle, as assumed above. Rather, a subset of ionization times, smaller than the full quarter cycle and corresponding to larger field values, dominates. This in turn causes the duration of the emissions to shrink below $T/2$ and gives rise to the generation of sub-femtosecond pulses without the use of compression. However, compression techniques serve to further shrink the temporal width of the pulses to their transform-limited value, which is established by the overall spectral width.

To foster the constructive growth of the HHG light along the length of the gas region, phase matching must be maintained, i.e., the phase velocities of the exciting laser light and the HHG light must match (Sec. 22.2D). While the phase velocity of the high-frequency HHG light is essentially equal to the free-space velocity c_o, the phase velocity of the laser light is lower and depends on the gas pressure. One phase-matching scheme involves guiding the laser light through a hollow-core waveguide filled with the gas. While the guided laser light travels at the velocity of the guided mode, the high-frequency HHG light is unguided and travels with phase velocity c_o. In the absence of the gas, the velocity of the laser guided mode is greater than c_o (as the propagation constant is lower), but it can be reduced to c_o by increasing the gas pressure, thereby permitting phase matching to be achieved.

Aside from Ti^{3+}:sapphire lasers, ultrafast Yb^{3+}:silica-fiber lasers with high repetition rates are often used as sources to produce HHG. Coherent attosecond beams, with hundreds of μW in individual harmonic components and pulse durations in the attosecond domain, can be generated. Attosecond optics has myriad applications, particularly in spectroscopy and imaging. It is also useful for understanding chemical reactions since the electrons that participate in such reactions move on that timescale. Light travels at a speed of 0.3 nm/as, so it traverses the width of a water molecule (0.3 nm) in 1 attosecond. The generation of HHG using lasers operating in the mid infrared can, in principle, allow multiple re-encounters of the electron with its parent ion, thereby promising the generation of sub-attosecond (zeptosecond) optical pulses. HHG can also be elicited from solids. The character of such emissions from crystalline solids differs somewhat from those elicited from gases because of the periodic crystal structure of the material. In particular, the HHG depends on the polarization of the incident radiation and the frequency combs generated contain even harmonics as well as odd ones for band structures that effectively lack inversion symmetry. The sum of a collection of EUV or SXR frequency combs with slightly different spectral spacings, generated by sweeping the exciting-laser wavelength and averaging, can serve as a quasi-supercontinuum source in these spectral domains. Attosecond radiation may also be generated by free-electron lasers (Sec. 16.3F), but FELs are currently available only at large-scale facilities.

Limitations of the model. Though highly oversimplified, the semiclassical analysis presented here captures many of the essential features of HHG production and attosecond pulse generation. Many simplifications and approximations have been made: 1) the atomic model incorporates only a single active electron; 2) though the electron is assumed to tunnel through the potential barrier, it is treated as a classical particle rather than as a quantum-mechanical wavepacket; 3) the ionization is taken to be instantaneous and the effect of the Coulomb field associated with the resulting ionized atom is ignored; 4) the analysis is one-dimensional; and 5) the field is assumed to be linearly polarized. Advances in the field of attosecond optics have enabled the development of more sophisticated models that reveal more subtle features of HHG.

23.6 PULSE DETECTION

The measurement of an ultranarrow optical pulse is challenging since the fastest available photodetector is almost always too slow to carry out the task. Methods of effecting such measurements rely primarily on the use of an ultrafast optical shutter (gate) controlled by another, shorter reference pulse and a mechanism for introducing a controllable time delay between the two pulses. The light transmitted through the gate is measured as the process is repeated for different delays, thereby providing an estimate of the pulse-intensity profile $I(t)$. To measure the pulse phase $\varphi(t)$, interferometric approaches can be cleverly combined with nonlinear optical techniques.

In the spectral domain, the pulse is fully characterized by its spectral intensity $S(\nu)$ and spectral phase $\psi(\nu)$. These functions may be measured by the use of an optical spectrum analyzer in conjunction with interferometric measurements. A challenging aspect of ultrashort pulse detection is the fact that the optical components employed in the measurement system unavoidably alter the optical pulse before measurement. Such effects must be minimized by careful system design, or compensated by appropriate post-detection signal processing.

A. Measurement of Intensity

Direct Photodetection

Ideally, the intensity profile of a short optical pulse may be directly measured by making use of a photodetector whose response time is much shorter than the pulse. In accordance with (19.1-4), the measured photocurrent $i_p(t)$ is then proportional to the pulse intensity $I(t)$ via

$$i_p(t) = \mathsf{R}AI(t), \qquad \text{(23.6-1)}$$
Fast Detector

where R is the responsivity (A/W) of the detector and A is its active area (which is assumed to be sufficiently small so that the optical intensity is sampled at the position of the detector).

If the response time of the photodetector is not small in comparison with the pulse duration, which is the usual case for ultrashort optical pulses, the photocurrent response is a broadened and distorted version of the optical pulse. If $h_D(t)$ is the impulse response function of the detector, where $\int h_D(t)\,dt = \mathsf{R}$, then the photocurrent response is given by the convolution integral

$$i_p(t) = A\int I(\tau)\,h_D(t-\tau)\,d\tau, \qquad \text{(23.6-2)}$$
Arbitrary Detector

which has greater duration than the optical pulse. Other measures must then be used to determine its true intensity waveshape $I(t)$. It is apparent that (23.6-2) reduces to (23.6-1) in the limiting case when the photodetector response time [the width of $h_D(t)$] is much shorter than the duration of the optical pulse [the width of $I(t)$].

At the opposite extreme, when the optical pulse duration is much shorter than the response time of the photodetector, (23.6-2) becomes

$$i_p(t) \approx h_D(t)\,A\int I(\tau)\,d\tau, \qquad \text{(23.6-3)}$$

in which case the photocurrent takes on the temporal profile of the detector's impulse response function, rather than that of the optical pulse. These three cases are illustrated schematically in Fig. 23.6-1.

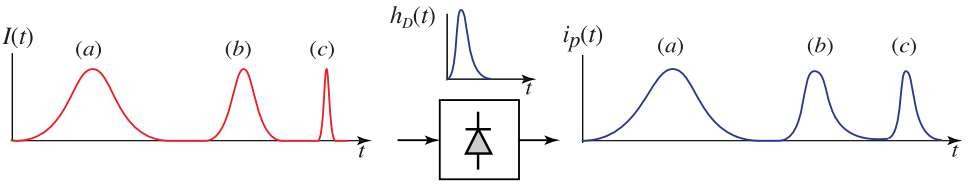


Figure 23.6-1 Response of a photodetector with impulse response function $h_D(t)$ to optical pulses of (a) long; (b) intermediate; and (c) short duration. For short optical pulses, as illustrated in (c), the photocurrent $i_p(t)$ follows $h_D(t)$ rather than $I(t)$.

Furthermore, if the photoreceiver circuitry has a time constant τ_R that is longer than the response time of the photodetector, with an impulse response function that is rectangular with $h_R(t) = 1/\tau_R$, the photocurrent response is $i_p(t) \approx (1/\tau_R) \int h_D(t)\, dt \cdot A \int I(\tau)\, d\tau$. This is a rectangular function of duration τ_R that can be written as

$$i_p(t) \approx \frac{1}{\tau_R} \mathsf{R}A \int I(\tau)\, d\tau. \qquad (23.6\text{-}4)$$

Slow Detector

The receiver then measures the area multiplied by the time integral of the intensity, which is the energy of the optical pulse. The receiver then altogether lacks temporal resolution and may be simply modeled as an integrator.

So, how might one measure the temporal profile of an ultrashort optical pulse, whose duration is in the picosecond or femtosecond regime, by making use of a "slow" photodetector, whose response time is a few tenths of a nanosecond at best?

Measurement of Short Pulse with Slow Detector and Fast Shutter

The temporal profile of a short optical pulse may be measured with a slow detector by making use of a fast shutter (switch or gate). As illustrated in Fig. 23.6-2, the gate opens for only a short time window during the course of the pulse, allowing a sample of the pulse to be detected by the slow detector. The measurement is repeated by opening the gate at different times, and a set of measured samples are used to estimate the pulse intensity profile. Since electronically operated gates are not viable at speeds in the picosecond or femtosecond range, we consider instead an optical gate controlled by a reference optical pulse whose duration is much shorter than that of the measured pulse (see Sec. 24.3C for a discussion of all-optical switches).

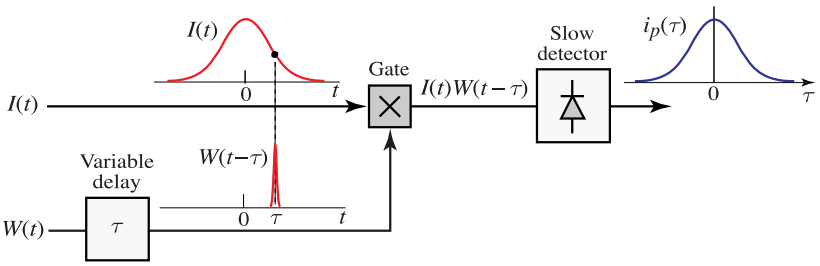


Figure 23.6-2 Measurement of an optical pulse $I(t)$ by use of an optical gate controlled by a far shorter optical gating pulse $W(t)$.

Two examples of optical gates used for the measurement of ultranarrow pulses are displayed in Fig. 23.6-3.

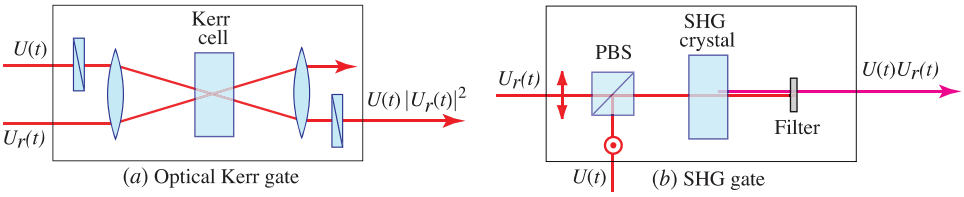


Figure 23.6-3 (*a*) An anisotropic nonlinear Kerr gate. The reference pulse intensity $I_r(t) = |U_r(t)|^2$ alters the Kerr-medium phase retardation, as explained in Sec. 22.3A. Since the test pulse $U(t)$ is transmitted through a pair of crossed polarizers, with the Kerr medium sandwiched between, it is modulated by the gating function $W(t) \propto |U_r(t)|^2$. (*b*) A second-harmonic generation (SHG) gate. The test pulse $U(t)$ and the gating pulse $U_r(t)$, which have orthogonal polarizations, combine in a collinear Type-II configuration (see Prob. 22.2-3), generating a pulse at the second-harmonic frequency with amplitude $\propto U(t)U_r(t)$, so that the gating function $W(t) \propto U_r(t)$.

Finally, we assess the effect of the finite switching time on the measurement resolution. If $W(t)$ is the transmittance of the gate when initiated by a gating pulse at $t = 0$, then when the gating action is delayed by time τ the transmitted optical pulse intensity is $I(t)W(t - \tau)$. The delay τ may be imparted either to the gating function $W(t)$ or to the optical pulse itself, $I(t)$. When detected by the slow photodetector, the resultant photocurrent is proportional to the area under the transmitted pulse,

$$i_p(t) \propto \int I(\tau)W(t - \tau)d\tau. \tag{23.6-5}$$

The measured photocurrent is thus proportional to the convolution of the optical pulse and the window function. Hence, the temporal resolution of the measurement is equal to the width of the window function $W(t)$, which is governed by the gate/shutter speed. Were the window function $W(t)$ a delta function $\delta(t)$, the photocurrent would be proportional to $I(\tau)$ and there would be no loss of resolution.

Single-Shot Pulse versus Pulse Train

The preceding method for measuring the shape of a short pulse with a slow detector can be easily implemented if a periodic train of identical pulses is available. The shutter is set at a different time delay τ for each of a sequence of pulses, as illustrated in Fig. 23.6-4, and the readings of the detector are recorded sequentially. The pulse repetition rate must, of course, be sufficiently low so that the slow detector can recover before encountering a subsequent pulse.

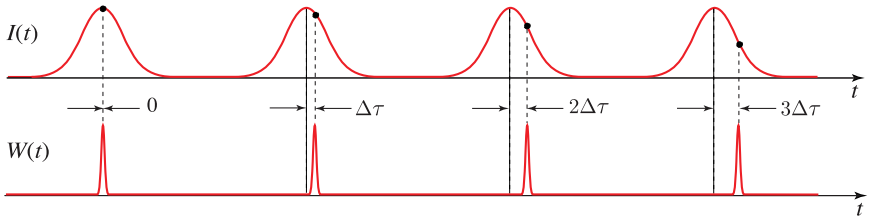


Figure 23.6-4 Measurement of a pulse intensity profile $I(t)$ by sampling individual pulses of a pulse train at time delays $\tau = m\Delta\tau$, $m = 0, 1, 2, \ldots$.

But what if a single-shot pulse is to be measured? This may be accomplished by generating multiple copies of the pulse via a fan-out optical element (see Fig. 24.1-4). As shown in Fig. 23.6-5, each copy is then subjected to a different time delay before

transmission through a gate controlled by a gating pulse $W(t)$. The use of an array of detectors then permits $I(t)$ to be recovered for a single-shot pulse.

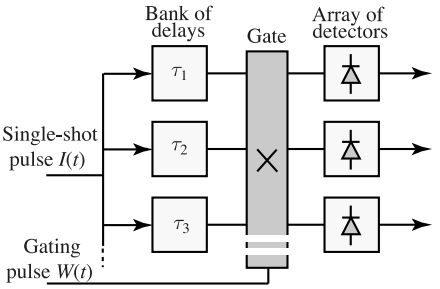


Figure 23.6-5 Measurement of the intensity profile $I(t)$ of a brief single-shot optical pulse by means of an optical fan-out device, followed by a bank of optical delays, an optical gate, and an array of slow photodetectors.

Temporal-to-Spatial Transformation: Streak-Camera Principle

The fan-out and multiple-delay concept depicted in Fig. 23.6-5 may be implemented optically by using an extended beam intercepted at an angle by a planar spatial detector (e.g., an array detector or CCD camera), as illustrated in Fig. 23.6-6. A pulsed plane wave traveling in the z direction has intensity $I(t - z/c)$, so that a wave traveling at an angle θ with respect to the z axis has an intensity $I(t - [x \sin\theta + z\cos\theta]/c)$. If this beam is intercepted by a spatial detector at the plane $z = 0$, it detects the intensity $I(t - [x \sin\theta]/c)$; hence, at position x the pulse is delayed by time $\tau_x = [x \sin\theta]/c$. Every detector element therefore sees its own delay, which implements the scheme displayed in Fig. 23.6-5. If a shutter is used to take a snapshot at time $t = 0$, the detector reading at location x is proportional to $I(-[x \sin\theta]/c)$. The result is that the pulse shape $I(t)$ is spatially recorded with a mirror-image profile, scaled such that an incident pulse of duration τ_0 (spatial width $c\tau_0$) creates an image of transverse width $c\tau_0/\sin\theta$ at the plane of the spatial detector, as sketched in Fig. 23.6-6. As an example, a pulse of duration $\tau_0 = 10$ ps extends over a spatial width $c\tau_0$ = 3 mm along the direction of propagation; at an angle $\theta = 30°$ this yields an image width of 6 mm at the detector plane.

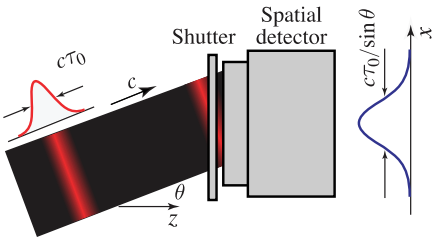


Figure 23.6-6 Temporal-to-spatial transformation of an optical pulse by use of an oblique wave and a spatial detector such as a CCD camera. The streak camera makes use of this principle.

This is the principle underlying the **streak camera**. The pulsed light is "streaked" in such a way that rays hitting different points on the extended spatial detector travel different distances and therefore experience different time delays. Such a position-dependent time delay may alternatively be introduced by transmitting the beam through a glass wedge.

The shutter used in the system portrayed in Fig. 23.6-6 may be an optical Kerr gate or a SHG gate, as schematized in Figs. 23.6-3(a) and (b), respectively. A particularly convenient implementation is the non-collinear Type-II SHG gate illustrated in Fig. 23.6-7. The test and gating pulses take the form of orthogonally polarized oblique

waves at angles θ and $-\theta$ with respect to the z axis. Their wavefunctions are thus $U(t - [x\sin\theta + z\cos\theta]/c)$ and $U_r(t - [-x\sin\theta + z\cos\theta]/c)$, respectively, so that the relative time delay is $\tau_x = [2x\sin\theta]/c$ at position x. The wave generated at the second-harmonic frequency has a wavefunction proportional to the product $U \cdot U_r$, so that the measured intensity is proportional to $I \cdot I_r$. As a result, the detected signal is proportional to the intensity autocorrelation function $G_I(\tau_x)$, as explained below.

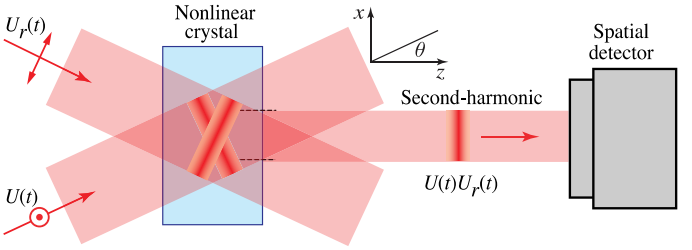


Figure 23.6-7 Measurement of a single-shot pulse by use of non-collinear Type-II SHG (Exercise 22.2-1) and time-to-space transformation (streaking). The detected signal is proportional to the intensity correlation function $G_I(\tau_x)$.

Measurement of Intensity Autocorrelation

As indicated above, the measurement of an ultrashort optical pulse $I(t)$ with a slow detector is achieved by making use of a shorter optical pulse $W(t)$ to control an optical gate. When no such external gating pulse exists, the test pulse itself may be compressed and used for this purpose. Or, a squared version of the test pulse obtained via SHG, for example, can also serve this purpose since $I^2(t)$ is narrower than $I(t)$. Higher-order harmonic generation could be used to generate an even narrower pulse, although it would be of lower intensity.

In circumstances when neither compression nor harmonic generation of the test pulse is feasible, the test pulse itself may be used as the gating pulse. As illustrated in Fig. 23.6-8, the photocurrent will then be proportional to the intensity autocorrelation function, given by

$$G_I(\tau) = \int I(t)I(t-\tau)dt = \int I(t)I(t+\tau)dt. \qquad (23.6\text{-}6)$$

Intensity Autocorrelation

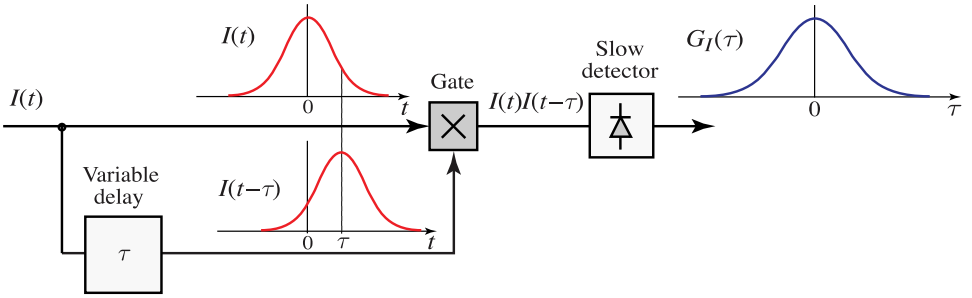


Figure 23.6-8 Measurement of the intensity autocorrelation function $G_I(\tau)$.

Since $I(t)$ is a real function of finite duration, $G_I(\tau)$ is a symmetric function that decreases from a peak value $G_I(0)$ at $\tau = 0$, to zero at $\tau = \infty$. The autocorrelation function of a pulse of arbitrary shape is a broader symmetric function. For example, a Gaussian pulse of intensity $I(t) = \exp[-2(\tau/\tau_0)^2)$, with width τ_0, has a (Gaussian) autocorrelation function $G_I(\tau) \propto \exp[-(\tau/\tau_0)^2]$, which may be written as $\exp[-2(\tau/\sqrt{2}\,\tau_0)^2]$, revealing that the width has broadened from τ_0 to $\sqrt{2}\,\tau_0$.

Knowledge of the autocorrelation function is generally not sufficient to determine the underlying function itself. This can be seen by noting that the Fourier transform of $G_I(\tau)$ is $|\mathcal{I}(\nu)|^2$, where $\mathcal{I}(\nu)$ is the Fourier transform of $I(t)$ (see Sec. A.1 in Appendix A). Measurement of $G_I(\tau)$ thus permits the magnitude $|\mathcal{I}(\nu)|$ to be determined but provides no information about the phase and hence cannot be used to completely recover the complex envelope. An exception is the symmetric pulse, for which $I(-t) = I(t)$, since $\mathcal{I}(\nu)$ is then real and therefore has zero phase. However, if the mathematical profile of a nonsymmetric function is known, the measurement of its autocorrelation function does allow parameters such as its width to be estimated.

B. Measurement of Spectral Intensity

Optical Spectrum Analyzer

The spectral intensity $S(\nu) = |A(\nu)|^2$ of an optical pulse of complex envelope $\mathcal{A}(t)$ may be measured by use of an **optical spectrum analyzer**, which is simply a bank of spectral filters tuned to an appropriate set of frequencies/wavelengths. If a bank of "slow" detectors is used to detect the energy in each of the spectral components, the result of the measurement is the spectral intensity $S(\nu)$. An optical implementation is sketched in Fig. 23.6-9.

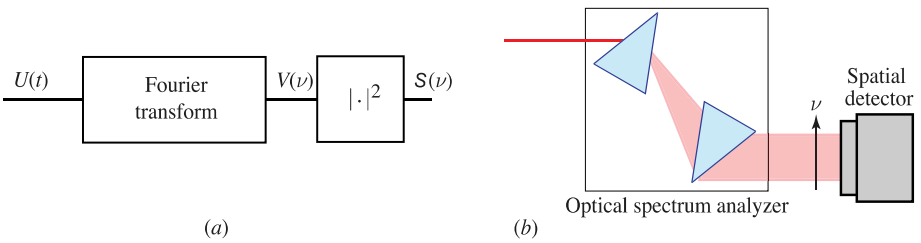


Figure 23.6-9 Measurement of spectral intensity with an optical spectrum analyzer. (*a*) System. (*b*) Optical implementation using prisms.

It is generally not possible to retrieve the complex function $\mathcal{A}(t)$ from the magnitude of its Fourier transform $|A(\nu)|$ in the absence of phase information. An exception is the case of a symmetric pulse, whose Fourier transform is real.

Interferometric Spectrum Analyzer

The spectral intensity $S(\nu)$ of an optical pulse may also be measured by use of an **interferometric spectrum analyzer**, as portrayed in Fig. 23.6-10. Recall from Sec. 12.2B that a Michelson interferometer may be used as a Fourier-transform spectrometer. When a pulsed optical beam of complex wavefunction $U(t)$ is split into two beams by a 50/50 beamsplitter, and one beam is delayed by time τ with respect to the other, the combined beam has optical field $(1/\sqrt{2})[U(t)+U(t-\tau)]$ and intensity $(1/2)|U(t)+U(t-\tau)|^2$.

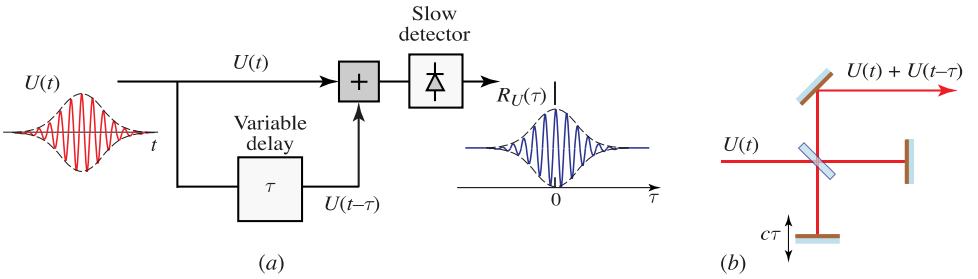


Figure 23.6-10 Interferometric measurement of the pulse spectral intensity. The interferogram is used to determine the autocorrelation function of the pulse envelope $G_{\mathcal{A}}(\tau)$, whose Fourier transform is the spectral intensity.

When detected with a "slow detector," the result is a function of the optical delay,

$$R_U(\tau) = \tfrac{1}{2}\int |U(t)+U(t-\tau)|^2 dt$$
$$= \tfrac{1}{2}\int |U(t)|^2 dt + \tfrac{1}{2}\int |U(t-\tau)|^2\, dt + \text{Re}\int U^*(t)U(t-\tau)dt. \quad (23.6\text{-}7)$$

Substituting $U(t) = \mathcal{A}(t)\exp(j2\pi\nu_0 t)$ leads to

$$R_U(\tau) = G_{\mathcal{A}}(0) + \text{Re}\left\{G_{\mathcal{A}}(\tau)\exp(-j2\pi\nu_0\tau)\right\}$$
$$= G_{\mathcal{A}}(0) + |G_{\mathcal{A}}(\tau)|\cos\left[2\pi\nu_0\tau - \arg\left\{G_{\mathcal{A}}(\tau)\right\}\right], \quad (23.6\text{-}8)$$

where

$$G_{\mathcal{A}}(\tau) = \int \mathcal{A}^*(t)\mathcal{A}(t-\tau)dt \quad (23.6\text{-}9)$$

is the autocorrelation function of the complex envelope, which is the inverse Fourier transform of the spectral intensity $S(\nu) = |A(\nu)|^2$. The measurement $R_U(\tau)$ is a fringe pattern of visibility $|G_{\mathcal{A}}(\tau)|/G_{\mathcal{A}}(0)$. This scheme permits us to determine $G_{\mathcal{A}}(\tau)$ through careful analysis of the visibility and location of the fringes. The interferometer therefore provides the same information as does the conventional spectrum analyzer.

C. Measurement of Phase

The full characterization of an optical pulse involves knowledge of the complex envelope, i.e., the magnitude and phase of the wavefunction $U(t) = \sqrt{I(t)}\exp[j2\pi\nu_0 t + \varphi(t)]$, or equivalently the magnitude and phase of its Fourier transform $V(\nu) = \sqrt{S(\nu)}\exp[j\psi(\nu)]$. The techniques presented in Sec. 23.6A provide means for determining the intensity $I(t)$, but not the phase $\varphi(t)$. The approaches set forth in Sec. 23.6B, on the other hand, offer ways of measuring the spectral intensity $S(\nu)$, but provide no information about the spectral phase $\psi(\nu)$. Only under special conditions can knowledge of the magnitudes of a function and its Fourier transform serve to fully characterize that function, absent knowledge of the associated phases.

We now consider measurements that are directly sensitive to the phase $\varphi(t)$ and the spectral phase $\psi(\nu)$. Techniques for phase measurement are often based on interferometry since the intensity at the interferometer output is highly sensitive to the difference between the phases of the interfering waves.

A conventional method for measuring phase is **optical heterodyning**, a form of **temporal interferometry** (see Sec. 2.6B). In this approach the complex wavefunction of the test pulse $U(t) = \sqrt{I(t)}\exp[j2\pi\nu_0 t + \varphi(t)]$ is mixed with a known reference pulse $U_r(t) = \sqrt{I_r(t)}\exp[j2\pi\nu_r t + \varphi_r(t)]$, which has a central frequency $\nu_r = \nu_0 + f$. The intensity of the sum is given by

$$|U(t) + U_r(t)|^2 = I(t) + I_r(t) + 2\sqrt{I(t)I_r(t)}\cos\left[2\pi f t + \varphi_r(t) - \varphi(t)\right]. \quad (23.6\text{-}10)$$

Equation (23.6-10) is a generalization of (2.6-12), in which the phases $\varphi(t)$ and $\varphi_r(t)$ were arbitrarily set to zero. Equation (23.6-10) represents an interferogram whose beat frequency f (fringes per second) is the difference between the central frequencies, and whose time-varying phase is $[\varphi_r(t) - \varphi(t)]$; both the beat frequency and phase may be readily extracted from the interferogram. For ultrashort pulses, however, the detector is always "slow" so that the temporal features of the interferogram are unfortunately averaged out. Temporal interferometry, including heterodyning, therefore cannot be marshaled for the measurement of $\varphi(t)$.

Spectral Interferometry

Interferometry can yield useful results, however, if it is carried out in the Fourier domain. In an approach known as **spectral interferometry**, the pulse $U(t)$ is delayed by a fixed time τ and added to a known reference pulse $U_r(t)$ of the same frequency. The Fourier transform of the sum, $U(t-\tau) + U_r(t)$, is then measured with a slow detector to create an interferogram, as illustrated in Fig. 23.6-11. If the Fourier transforms of $U(t)$ and $U_r(t)$ are $V(\nu) = \sqrt{S(\nu)}\exp[j\psi(\nu)]$ and $V_r(\nu) = \sqrt{S_r(\nu)}\exp[j\psi_r(\nu)]$, respectively, the spectral interferometer measures the interferogram

$$\left|V(\nu)e^{-j2\pi\tau\nu} + V_r(\nu)\right|^2 = S(\nu) + S_r(\nu) + 2\sqrt{S(\nu)S_r(\nu)}\cos\left[2\pi\tau\nu + \psi_r(\nu) - \psi(\nu)\right]. \quad (23.6\text{-}11)$$

This is a fringe pattern (in frequency) whose visibility is determined by the spectral intensity $S(\nu)$ and whose fringe locations are governed by the phase difference $\psi_r(\nu) - \psi(\nu)$. The measurement therefore yields full information about $V(\nu)$, and hence about $U(t)$.

The duality between temporal and spectral interferometry may be understood by noting that (23.6-10) and (23.6-11) are identical in form, with t and ν playing dual roles, while the delay τ plays the role of the frequency difference f. The principal challenge of spectral interferometry is the necessity for a known reference pulse.

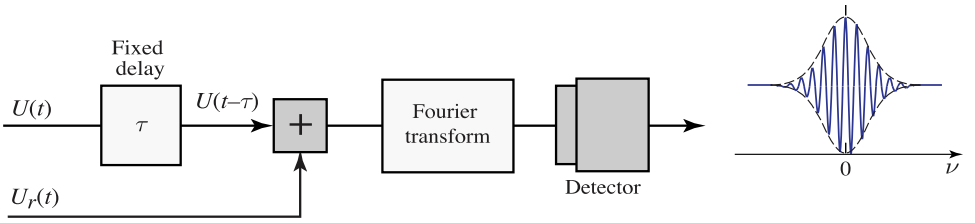


Figure 23.6-11 A spectral interferometer generates an interferogram in the Fourier domain.

Self-Referenced Spectral Interferometry

The test pulse cannot be used as its own reference since the phase term in (23.6-11) vanishes if $\psi_r(\nu) = \psi(\nu)$. One method of working around this problem is to use a frequency-shifted version of the test pulse as a reference, i.e., by choosing $V_r(\nu) = V(\nu + f)$. A block diagram of the configuration is presented in Fig. 23.6-12. The resultant interferogram is

$$\left|V(\nu)e^{-j2\pi\tau\nu} + V(\nu+f)\right|^2 = S(\nu) + S(\nu+f) + 2\sqrt{S(\nu)S(\nu+f)}\cos\left[2\pi\tau\nu + \psi(\nu+f) - \psi(\nu)\right], \tag{23.6-12}$$

from which the phase difference $\psi(\nu + f) - \psi(\nu)$ may be estimated. If the frequency shift f is small, this phase difference may be used as an approximation of the derivative $d\psi/d\nu$, which may then be integrated to provide the phase $\psi(\nu)$.

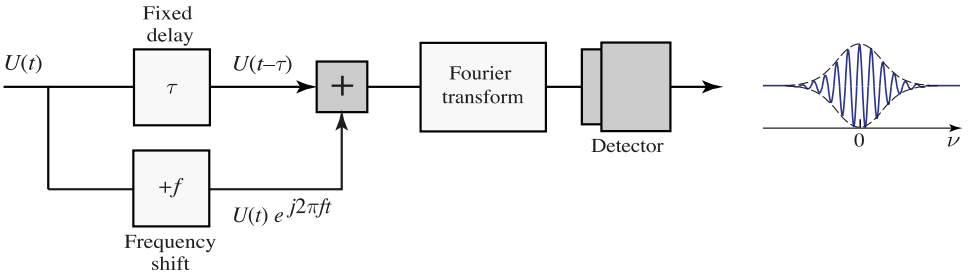


Figure 23.6-12 Self-referenced spectral interferometer.

Nonlinear Interferometry

As discussed in Sec. 23.6B, an interferometric spectrum analyzer can provide full information about the spectral intensity $S(\nu)$ of a pulse, although it provides no information about the spectral phase. The block diagram in Fig. 23.6-10 demonstrated that conventional time-domain interferometry achieves this by measuring of the area under the function $|U(t) + U(t-\tau)|^2$ [see (23.6-7)].

One way to extract phase information is to make use of a nonlinear interferometer. If the integrand in (23.6-7) is squared prior to detection, the interferometer extracts the area under the function $|[U(t) + U(t-\tau)]^2|^2$, as illustrated by the block diagram in Fig. 23.6-13. The squaring operation may be implemented by second-harmonic generation in a nonlinear optical crystal.

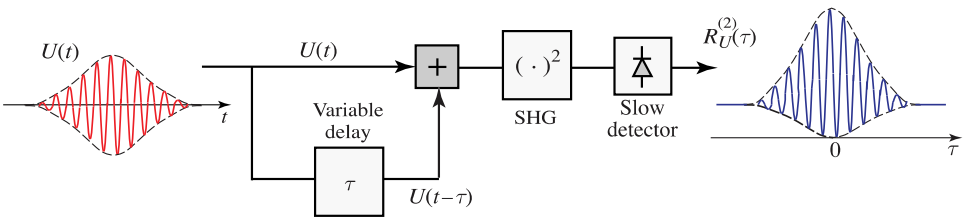


Figure 23.6-13 Nonlinear interferometer.

The nonlinear function analogous to $R_U(\tau)$ in (23.6-7), denoted $R_U^{(2)}(\tau)$, is then given by

$$R_U^{(2)}(\tau) = \int \left| [U(t) + U(t-\tau)]^2 \right|^2 dt. \tag{23.6-13}$$

To demonstrate that $R_U^{(2)}(\tau)$ contains the phase information we seek, we substitute $U(t) = \mathcal{A}(t)\exp(j2\pi\nu_0 t)$ into (23.6-13), and separate terms with frequencies 0 , ν_0, and $2\nu_0$. This leads to

$$R_U^{(2)}(\tau) \propto C_0(\tau) + 4\,\mathrm{Re}\left\{C_1(\tau)e^{j2\pi\nu_0\tau}\right\} + 2\,\mathrm{Re}\left\{C_2(\tau)e^{j4\pi\nu_0\tau}\right\} \tag{23.6-14}$$

with

$$\begin{aligned} C_0(\tau) &= \int I^2(t)dt + \int I^2(t-\tau)dt + 4\int I(t)I(t-\tau)dt \\ &= 2G_I(0) + 4G_I(\tau), \end{aligned} \tag{23.6-15}$$

$$C_1(\tau) = \int \mathcal{A}^*(t)\mathcal{A}(t-\tau)\left[I(t) + I(t-\tau)\right]dt, \tag{23.6-16}$$

$$C_2(\tau) = \int \left[\mathcal{A}^*(t)\mathcal{A}(t-\tau)\right]^2 dt, \tag{23.6-17}$$

where $G_I(\tau)$ is the intensity autocorrelation function defined in (23.6-6). The function $R_U^{(2)}(\tau)$ is seen to be the sum of three terms: a nonoscillatory term $C_0(\tau)$ and two oscillatory terms at frequencies ν_0 and $2\nu_0$, which may be separated by Fourier analyzing $R_U^{(2)}(\tau)$.

The first term depends on the intensity autocorrelation function $G_I(\tau)$ and has no phase dependence. The two other terms depend on both the pulse intensity and phase. The overall function is bounded by an upper envelope with maximum value $R_U^{(2)}(0) = 16\int I^2(t)dt = 16G_I(0)$ and a lower envelope with minimum value $R_U^{(2)}(0) = 0$. Its asymptotic value is $R_U^{(2)}(\infty) = C_0(\infty) = 2\int I^2(t)dt = 2G_I(0)$. The ratio $R_U^{(2)}(\tau)/R_U^{(2)}(\infty)$ therefore goes from a peak value of 8 at $\tau = 0$ to an asymptotic value of unity at $\tau = \infty$.

As a specific example, we examine a linearly chirped Gaussian pulse with time constant τ_0 and chirp parameter a, for which

$$C_0(\tau) = 2G_I(0)\left[1 + 2\exp(-\tau^2/\tau_0^2)\right], \tag{23.6-18}$$

$$C_1(\tau) = 2G_I(0)\exp\left[-(3+a^2)\tau^2/4\tau_0^2\right]\cos\left(a\tau^2/2\tau_0^2\right), \tag{23.6-19}$$

$$C_2(\tau) = G_I(0)\exp\left[-(1+a^2)\tau^2/\tau_0^2\right]. \tag{23.6-20}$$

The normalized function $R_U^{(2)}(\tau)/R_U^{(2)}(\infty)$ is plotted in Fig. 23.6-14 for three values of the chirp parameter a. It is evident that the profile of the interferogram, particularly the point at which the oscillatory terms vanish, is highly sensitive to a, and can therefore be used to estimate the chirp parameter from experimental data.

Though there is no general procedure for estimating the pulse phase from the measurement of $R_U^{(2)}(\tau)$, the measurement can nevertheless can be used to verify known models for pulse amplitude and phase as well as to estimate unknown parameters.

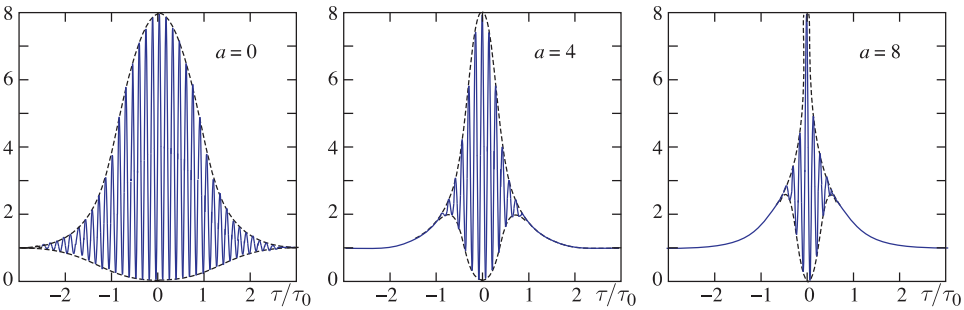


Figure 23.6-14 Normalized intensity autocorrelation function $R_U^{(2)}(\tau)/R_U^{(2)}(\infty)$, plotted against the normalized time delay τ/τ_0, for a chirped Gaussian pulse with three different values of the chirp parameter a.

Nonlinear Interferometry with Nonlinear Detectors. In an alternative implementation of the nonlinear interferometer depicted in Fig. 23.6-13, the squaring operation is carried out by the detector itself. This is accomplished by use of a detector based on two-photon absorption, e.g., a photodiode with a bandgap energy greater than the photon energy, but smaller than twice the photon energy (see Prob. 19.1-5). In such a detector, the photocurrent is proportional to the square of the intensity (since it absorbs pairs of photons). As a result, such a nonlinear interferometer would measure the function

$$R_U'^{(2)}(\tau) = \int |U(t) + U(t-\tau)|^4 dt, \tag{23.6-21}$$

which, like (23.6-13), contains information about the pulse distribution and width.

*D. Measurement of Spectrogram

As set forth in Sec. 23.1A, the spectrogram $S(\nu, \tau)$ of an optical pulse $U(t)$ is a time–frequency representation given by the squared magnitude of the Fourier transform of the pulse, as seen through a moving window (gating function) $W(t)$:

$$S(\nu,\tau) = |\Phi(\nu,\tau)|^2; \quad \Phi(\nu,\tau) = \int U(t)W(t-\tau)\exp(-j2\pi\nu t)dt. \tag{23.6-22}$$

Consequently, the spectrogram may be measured by transmitting the pulse $U(t)$ through an optical gate controlled by a time-delayed gating function $W(t-\tau)$, and measuring the spectrum of the product $U(t)W(t-\tau)$ with a spectrum analyzer at each time delay τ, as depicted schematically in Fig. 23.6-15. An optical implementation relies on a moving mirror to introduce the time delay, an appropriate optical gate, and an optical spectrum analyzer such as that shown in Fig. 23.6-9. The technique is known as **frequency-resolved optical gating** (**FROG**).

In the absence of a sufficiently short gating function $W(t)$, the pulse $U(t)$ itself, or a related pulse, may be used for this purpose. The relation between $W(t)$ and $U(t)$ depends on the nature of the optical gate, as illustrated by the following examples:

- For a second-harmonic generation (SHG) gate (see Fig. 23.6-7) with input waves $U(t)$ and $U(t-\tau)$ at the fundamental frequency, the wave at the second-harmonic

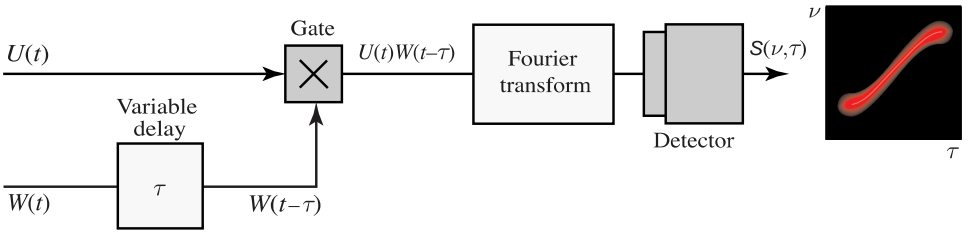


Figure 23.6-15 Measurement of the spectrogram $S(\nu, \tau)$ by frequency-resolved optical gating (FROG).

frequency is proportional to the product $U(t)U(t-\tau)$, so that $W(t) \propto U(t)$ and

$$\Phi(\nu, \tau) = \int U(t)U(t-\tau)\exp(-j2\pi\nu t)dt. \tag{23.6-23}$$

The time-frequency function in (23.6-23) is known as the **Wigner distribution function**. The overall optical system that implements the block diagram in Fig. 23.6-15 is depicted in Fig. 23.6-16(a) and the system is known as the **SHG-FROG**. This system is suitable for single-shot measurement, as discussed earlier.

- For a polarization-based optical Kerr gate [Fig. 23.6-3(a)], $W(t)$ is proportional to the pulse intensity $I(t)$ so that $W(t) \propto I(t) = |U(t)|^2$ and

$$\Phi(\nu, \tau) = \int U(t)|U(t-\tau)|^2\exp(-j2\pi\nu t)dt. \tag{23.6-24}$$

When this gate is used to implement the block diagram in Fig. 23.6-15, the system is called the polarization-gated FROG (**PG-FROG**), which is illustrated in Fig. 23.6-16(b).

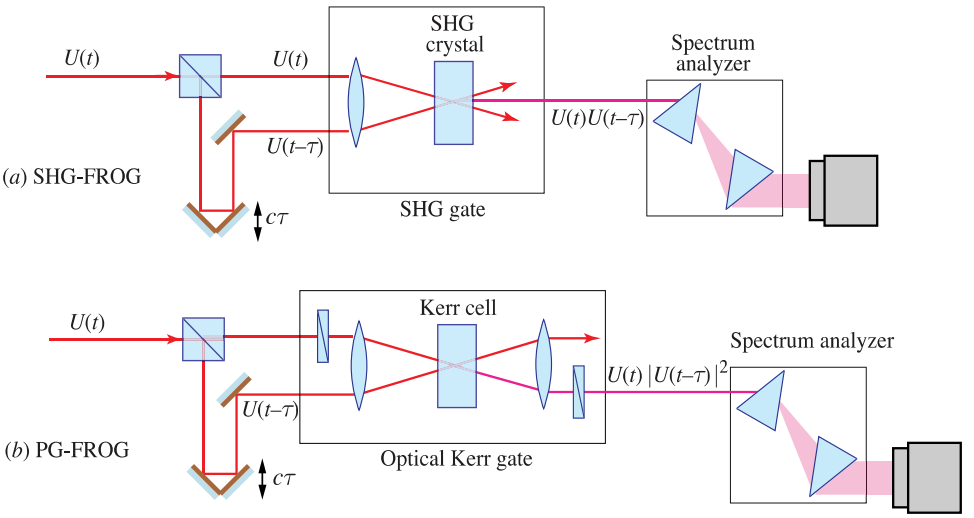


Figure 23.6-16 Two implementations of frequency-resolved optical gating (FROG): (a) Second-harmonic generation FROG (SHG-FROG); (b) Polarization-gated FROG (PG-FROG).

Other nonlinear optical configurations have also been devised, including a gate based on third-harmonic generation, which corresponds to the gating function $W(t) \propto U^2(t)$, and a gate based on self-diffraction, which corresponds to $W(t) \propto [U^*(t)]^2$.

Estimation of the Pulse Wavefunction from the Spectrogram

In its many variations, the spectrogram $S(\nu,\tau)$ provides a 2D "picture" that may be used to characterize the optical pulse by displaying signatures of its key features. With some difficulty, it may also be used to estimate the complex wavefunction (both magnitude and phase) of the pulse $U(t)$.

The estimation of $U(t)$ from the measured spectrogram $S(\nu,\tau)$ is not straightforward. A general expression for $S(\nu,\tau)$, as determined for any of the gating systems considered above, may be written in the form

$$S(\nu,\tau) = |\Phi(\nu,\tau)|^2; \quad \Phi(\nu,\tau) = \int g(t,\tau)\exp(-j2\pi\nu t)dt, \tag{23.6-25}$$

where $g(t,\tau) = U(t)W(t-\tau)$ and $W(t)$ is related to $U(t)$. As we have seen previously, $W(t) = U(t)$ for the SHG-FROG, and $W(t) = |U(t)|^2$ for the PG-FROG.

If the complex function $\Phi(\nu,\tau)$ were known, $U(t)$ could be readily estimated as follows. Form the inverse Fourier transform of $\Phi(\nu,\tau)$ with respect to ν at each τ, to obtain:

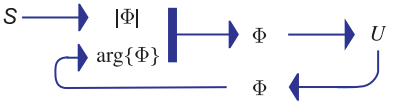
$$g(t,\tau) = \int \Phi(\nu,\tau)\exp(j2\pi\nu t)d\nu. \tag{23.6-26}$$

Given that $g(t,\tau) = U(t)W(t-\tau)$, compute the wavefunction $U(t)$ by integration over τ:

$$\int g(t,\tau)d\tau = \int U(t)W(t-\tau)d\tau = U(t)\int W(t-\tau)d\tau \propto U(t). \tag{23.6-27}$$

The proportionality constant is the area under the window function, which is unknown. Nevertheless, this analysis will prove useful in the following paragraph.

The problem of estimating $\Phi(\nu,\tau)$ from its measured absolute square, $S(\nu,\tau) = |\Phi(\nu,\tau)|^2$, is known as a *missing-phase problem*. Many algorithms have been devised for addressing this kind of problem. One iterative approach follows the steps illustrated by the following diagram:



1. Beginning with the measured spectrogram $S(\nu,\tau)$, determine the magnitude $|\Phi(\nu,\tau)| = [S(\nu,\tau)]^{1/2}$. With an initial guess for the missing phase, $\arg\{\Phi(\nu,\tau)\}$, use the procedure discussed in the previous paragraph [inverse Fourier transform $\Phi(\nu,\tau)$ with respect to ν and integrate over τ] to estimate $U(t)$ up to an unknown proportionality constant.
2. Using this value of $U(t)$, compute $\Phi(\nu,\tau)$ to obtain a revised estimate of the unknown phase $\arg\{\Phi(\nu,\tau)\}$; use this together with the measured magnitude $|\Phi(\nu,\tau)|$ to obtain a new and better estimate of $U(t)$.
3. Repeat the process until it converges to a pulse wavefunction $U(t)$ that is consistent with the measured spectrogram.

An example displaying the outcome of this procedure for SHG-FROG is provided in Fig. 23.6-17.

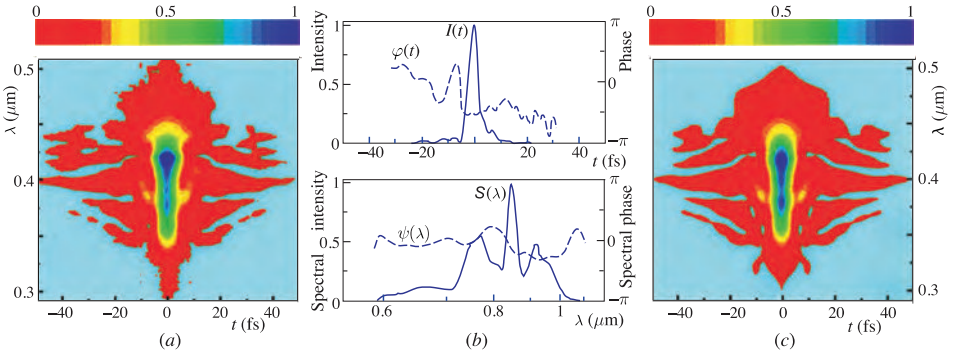


Figure 23.6-17 (a) Measured spectrogram $S_\lambda(\lambda, \tau)$ of a $2^1/_2$-cycle, $4^1/_2$-fs optical pulse with a central wavelength $\approx$ 0.85 μm obtained by SHG-FROG. (b) Estimated temporal and spectral characteristics of the pulse. (c) The SHG-FROG spectrogram computed from the pulse in (b) is approximately the same as the measurement in (a). Note that the wavelength regions of peak response in (a) and (c) are at the second harmonic. (Adapted from A. Baltuška, M. S. Pshenichnikov, and D. A. Wiersma, *IEEE Journal of Quantum Electronics*, vol. 35, pp. 459–478, Figs. 17(a), 17(b), and 18 ©1999 IEEE; R. Trebino, ed., *Frequency-Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses*, Kluwer, 2000, figure on associated CD-ROM.)

READING LIST

Ultrafast Optics

See also the reading lists in Chapters 5, 6, and 22.

D. T. Reid, C. M. Heyl, R. R. Thomson, R. Trebino, G. Steinmeyer, H. H. Fielding, R. Holzwarth, Z. Zhang, P. Del'Haye, T. Südmeyer, G. Mourou, T. Tajima, D. Faccio, F. J. M. Harren, and G. Cerullo, Roadmap on Ultrafast Optics, *Journal of Optics*, vol. 18, 093006, 2016.

A. M. Weiner, *Ultrafast Optics*, Wiley, 2009.

K. E. Oughstun, *Electromagnetic and Optical Pulse Propagation. 2: Temporal Pulse Dynamics in Dispersive, Attenuative Media*, Springer-Verlag, 2009.

K. E. Oughstun, *Electromagnetic and Optical Pulse Propagation. 1: Spectral Representations in Temporally Dispersive Media*, Springer-Verlag, 2007.

J.-C. Diels and W. Rudolph, *Ultrashort Laser Pulse Phenomena*, Elsevier, 2nd ed. 2006.

C. Rullière, ed., *Femtosecond Laser Pulses: Principles and Experiments*, Springer-Verlag, 2nd ed. 2005.

F. X. Kärtner, ed., *Few-Cycle Laser Pulse Generation and Its Applications*, Springer-Verlag, 2004.

R. Trebino, ed., *Frequency-Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses*, Springer-Verlag, 2000.

S. A. Akhmanov, V. A. Vysloukh, and A. S. Chirkin, *Optics of Femtosecond Laser Pulses*, American Institute of Physics, 1992.

T. R. Gosnell and A. J. Taylor, eds., *Selected Papers on Ultrafast Laser Technology*, SPIE Optical Engineering Press (Milestone Series Volume 44), 1991.

W. Rudolph and B. Wilhelmi, *Light Pulse Compression*, Harwood, 1989.

D. Strickland and G. Mourou, Compression of Amplified Chirped Optical Pulses, *Optics Communications*, vol. 56, pp. 219–221, 1985.

Optical Solitons

S. Boscolo and C. Finot, eds., *Shaping Light in Nonlinear Optical Fibers*, Wiley, 2017.

L. N. Binh, *Optical Multi-Bound Solitons*, CRC Press/Taylor & Francis, 2016.

G. Fibich, *The Nonlinear Schrödinger Equation: Singular Solutions and Optical Collapse*, Springer-Verlag, 2015.

G. Agrawal, *Nonlinear Fiber Optics*, Academic Press/Elsevier Press, 5th ed. 2013.

M. F. S. Ferreira, *Nonlinear Effects in Optical Fibers*, Wiley, 2011.

L. F. Mollenauer and J. P. Gordon, *Solitons in Optical Fibers: Fundamentals and Applications*, Academic Press/Elsevier, 2006.

T. Dauxois and M. Peyrard, *Physics of Solitons*, Cambridge University Press, 2006.

A. Hasegawa and M. Matsumoto, *Optical Solitons in Fibers*, Springer-Verlag, 3rd ed. 2003.

Y. S. Kivshar and G. P. Agrawal, *Optical Solitons: From Fibers to Photonic Crystals*, Academic Press, 2003.

N. N. Akhmediev and A. Ankiewicz, *Solitons: Nonlinear Pulses and Beams*, Chapman & Hall, 1997.

L. F. Mollenauer and R. H. Stolen, The Soliton Laser, *Optics Letters*, vol. 9, pp. 13–15, 1984.

L. F. Mollenauer, R. H. Stolen, and J. P. Gordon, Experimental Observation of Picosecond Pulse Narrowing and Solitons in Optical Fibers, *Physical Review Letters*, vol. 45, pp. 1095–1098, 1980.

Supercontinuum Light Generation

R. R. Alfano, ed., *Supercontinuum Laser Source: The Ultimate White Light*, Springer, 3rd ed. 2016.

G. Genty, A. T. Friberg, and J. Turunen, Coherence of Supercontinuum Light, in T. D. Visser, ed., *Progress in Optics*, Elsevier, 2016, vol. 61, pp. 71–112.

J. M. Dudley and J. R. Taylor, eds., *Supercontinuum Generation in Optical Fibers*, Cambridge University Press, 2010.

J. M. Dudley, G. Genty, and S. Coen, Supercontinuum Generation in Photonic Crystal Fiber, *Reviews of Modern Physics*, vol. 78, pp. 1135–1184, 2006.

High-Harmonic Generation and Attosecond Optics

C. D. Lin, A.-T. Le, C. Jin, and H. Wei, *Attosecond and Strong-Field Physics: Principles and Applications*, Cambridge University Press, 2018.

M. Sivis, M. Taucer, G. Vampa, K. Johnston, A. Staudte, A. Yu. Naumov, D. M. Villeneuve, C. Ropers, and P. B. Corkum, Tailored Semiconductors for High-Harmonic Optoelectronics, *Science*, vol. 357, pp. 303–306, 2017.

F. Calegari, G. Sansone, S. Stagira, C. Vozzi, and M. Nisoli, Advances in Attosecond Science, *Journal of Physics B: Atomic, Molecular and Optical Physics*, vol. 49, 062001, 2016.

U. Huttner, K. Schuh, J. V. Moloney, and S. W. Koch, Similarities and Differences Between High-Harmonic Generation in Atoms and Solids, *Journal of the Optical Society of America B*, vol. 33, pp. C22–C29, 2016.

P. H. Bucksbaum, Sources and Science of Attosecond Light, *Optics & Photonics News*, vol. 26, no. 5, pp. 28–35, 2015.

S. Hädrich, A. Klenke, J. Rothhardt, M. Krebs, A. Hoffmann, O. Pronin, V. Pervak, J. Limpert, and A. Tünnermann, High Photon Flux Table-Top Coherent Extreme-Ultraviolet Source, *Nature Photonics*, vol. 8, pp. 779–783, 2014.

T. Popmintchev, M.-C. Chen, D. Popmintchev, P. Arpin, S. Brown, S. Ališauskas, G. Andriukaitis, T. Balčiunas, O. D. Mücke, A. Pugzlys, A. Baltuška, B. Shim, S. E. Schrauth, A. Gaeta, C. Hernández-García, L. Plaja, A. Becker, A. Jaron-Becker, M. M. Murnane, and H. C. Kapteyn, Bright Coherent Ultrahigh Harmonics in the keV X-Ray Regime from Mid-Infrared Femtosecond Lasers, *Science*, vol. 336, pp. 1287–1291, 2012.

A. Cingöz, D. C. Yost, T. K. Allison, A. Ruehl, M. E. Fermann, I. Hartl, and J. Ye, Direct Frequency Comb Spectroscopy in the Extreme Ultraviolet, *Nature*, vol. 482, pp. 68–71, 2012.

Z. Chang, *Fundamentals of Attosecond Optics*, CRC Press/Taylor & Francis, 2011.

G. Sansone, L. Poletto, and M. Nisoli, High-Energy Attosecond Light Sources, *Nature Photonics*, vol. 5, pp. 655–663, 2011.

T. Popmintchev, M.-C. Chen, P. Arpin, M. M. Murnane and H. C. Kapteyn, The Attosecond Nonlinear Optics of Bright Coherent X-Ray Generation, *Nature Photonics*, vol. 4, pp. 822–832, 2010.

F. Krausz and M. Ivanov, Attosecond Physics, *Revs. of Modern Physics*, Vol. 81, pp. 163–234, 2009.

P. B. Corkum and Z. Chang, The Attosecond Revolution, *Optics & Photonics News*, vol. 19, no. 10, pp. 24–29, 2008.

P. B. Corkum and F. Krausz, Attosecond Science, *Nature Physics*, vol. 3, pp. 381–387, 2007.

P. Jaeglé, *Coherent Sources of XUV Radiation: Soft X-Ray Lasers and High-Order Harmonic Generation*, Springer-Verlag, 2006.

H. C. Kapteyn, M. M. Murnane, and I. P. Christov, Extreme Nonlinear Optics: Coherent X Rays from Lasers, *Physics Today*, vol. 58, no. 3, pp. 39–44, 2005.

M. Wegener, *Extreme Nonlinear Optics: An Introduction*, Springer-Verlag, 2005.

P. B. Corkum, Plasma Perspective on Strong-Field Multiphoton Ionization, *Physical Review Letters*, vol. 71, pp. 1994–1997, 1993.

A. McPherson, G. Gibson, H. Jara, U. Johann, T. S. Luk, I. A. McIntyre, K. Boyer, and C. K. Rhodes, Studies of Multiphoton Production of Vacuum-Ultraviolet Radiation in the Rare Gases, *Journal of the Optical Society of America B*, vol. 4, pp. 595–601, 1987.

N. H. Burnett, H. A. Baldis, M. C. Richardson, and G. D. Enright, Harmonic Generation in CO_2 Laser Target Interaction, *Applied Physics Letters*, vol. 31, pp. 172–174, 1977.

Generation of Ultrafast Petawatt Pulses

E. Cartlidge, The Light Fantastic, *Science*, vol. 359, pp. 382–385, 2018.

B. Le Garrec, D. N. Papadopoulos, C. Le Blanc, J. P. Zou, G. Chériaux, P. Georges, F. Druon, L. Martin, L. Fréneaux, A. Beluze, N. Lebas, F. Mathieu, and P. Audebert, Design Update and Recent Results of The Apollon 10 PW Facility, *SPIE Proceedings*, vol. 10238 (High-Power, High-Energy, and High-Intensity Laser Technology III), 2017.

X. Zeng, K. Zhou, Y. Zuo, Q. Zhu, J. Su, X. Wang, X. Wang, X. Huang, X. Jiang, D. Jiang, Y. Guo, N. Xie, S. Zhou, Z. Wu, J. Mu, H. Peng, and F. Jing, Multi-Petawatt Laser Facility Fully Based on Optical Parametric Chirped-Pulse Amplification, *Optics Letters*, vol. 42, pp. 2014–2017, 2017.

A. Heller, Lighting a New Era of Scientific Discovery, *Science & Technology Review*, pp. 4-11, January/February 2014.

PROBLEMS

23.1-1 **Superposition of Two Gaussian Pulses.** A transform-limited Gaussian pulse is added to a chirped Gaussian pulse with chirp parameter a but otherwise identical parameters. Derive expressions for the intensity, phase, spectral intensity, spectral phase, and chirp parameter of the superposition pulse.

23.1-2 **The Hyperbolic-Secant Pulse.** Consider a pulse with complex envelope $\text{sech}(t/\tau)$, where $\text{sech}(\cdot) = 1/\cosh(\cdot)$ and τ is a time constant.

(a) Show that the width of the intensity function is $\tau_{\text{FWHM}} = 1.76\,\tau$.

(b) Show that the spectral intensity $S(\nu) \propto \text{sech}^2(\pi^2\tau\nu)$ and that the FWHM spectral width is $\Delta\nu = 1.786/\tau$.

(c) Compare these results with those for the Gaussian pulse provided in Appendix A.

23.2-1 **Thick-Prism Chirp Filter.** A thick prism is used as a chirp filter. The angle of incidence is selected to satisfy the Brewster-angle condition in order to minimize reflection loss. The apex angle α is selected such that the incident ray and the central deflected ray are symmetric with respect to the prism. Under these two conditions, show that the angle of deflection θ_d satisfies the condition $d\theta_d/dn = -2$, and that the chirp coefficient can be expressed as $b \approx -4(n-N)^2\ell_0\lambda_o/\pi c^2$. Show that this chirp coefficient is greater than that for the thin-prism chirp filter (Example 23.2-4) by a factor of $4/\alpha^2$, given that all other parameters are the same.

23.2-2 **Bragg-Grating Chirp Filter.** Design a Bragg-grating chirp filter for Gaussian pulses of central frequency $\nu_0 = 300$ THz (corresponding to a free-space wavelength of 1 μm) and $\tau_{\text{FWHM}} = 0.44$ ps. The filter is to have a chirp coefficient $b = (2\text{ ps})^2$. Specify the dimensions of the grating and the maximum and minimum pitch of its periodic structure to ensure that all spectral components of the pulse are reflected by the grating.

23.3-3 **Propagation of a Rectangular Pulse Through an Optical Fiber.** A rectangular pulse of width τ travels through an optical fiber, which is modeled as a chirp filter with chirp parameter $b = D_\nu z/\pi$ [see (23.3-5)]. Show that after a sufficiently long distance z, the pulse alters its shape from a rectangular function to a sinc function. Derive an expression for the modified pulse width.

23.3-4 **Temporal Imaging with a Time Lens.** An optical pulse of width τ_1 and arbitrary shape travels a distance d_1 through a fiber with positive GVD, where it is modulated by a phase factor $\exp(j\zeta t^2)$; it subsequently travels a distance d_2 through a fiber of the same material. The width of the final pulse is τ_2. Assuming that d_1 and d_2 are much larger than the dispersion length z_0 of the fiber, show that the new pulse will be a delayed replica of the original pulse, with time magnification $\tau_2/\tau_1 = d_2/d_1$, if the condition $1/d_1+1/d_2 = 1/f$ is satisfied. Here, $f = -\pi/\zeta D_\nu$ is the focal length of the phase modulator for this medium (ζ is negative and f is positive). This result indicates that the system is equivalent to a temporal imaging system.

23.5-1 **Mixing of Pulsed Chirped Waves and Chirp Amplification.**

(a) Three pulsed collinear plane waves with central angular frequencies ω_1, ω_2, and $\omega_3 = \omega_1 + \omega_2$ are mixed in a second-order nonlinear medium with nonlinear optical coefficient d. The medium is dispersive and has indices of refraction n_1, n_2, and n_3 and group velocities v_1, v_2, and v_3 at the three central frequencies. The three pulses are chirped with chirp parameters a_1, a_2, and a_3. What should be the relation between a_1, a_2, and a_3 for efficient three-wave mixing. *Hint*: Assume that energy conservation and momentum conservation (phase matching) relations are satisfied at all instants of time.

(b) Demonstrate that the chirp parameter of the signal and/or the idler may be greater than that of the pump. Discuss possible applications of this "chirp amplification" process.

*23.5-2 **Pulsed Three-Wave Mixing in a Medium with Group Velocity Dispersion (GVD).** Derive the three-wave-mixing coupled-wave equations (23.5-3) for a medium with GVD. You may use the following procedure. Begin with the Helmholtz equation with a source equal to the Fourier transform of $\mathcal{S} = \mu_o \partial^2 \mathcal{P}_{\rm NL}/\partial t^2$, where $\mathcal{P}_{\rm NL} = 2d\mathcal{E}^2$. Express the field $\mathcal{E}$ as a superposition of three waves with distinct central frequencies and slowly varying envelopes (SVEs), and convert the Helmholtz equation into three separate equations at the three frequencies. Simplify these equations using the SVE approximation, weak dispersion, and a three-term Taylor-series expansion of the propagation coefficient. Use an inverse Fourier transform to convert the equations back to the time domain.

23.5-3 **Dependence of Soliton Characteristics on Group Velocity Dispersion.** Compare the characteristics of two fundamental solitons of equal energy traveling in two different extended media (e.g., optical fibers) with GVD coefficients $D_\lambda = 20$ and 10 ps/km-nm, but with otherwise identical optical properties (same refractive index and same Kerr coefficient n_2). Compare the soliton widths, peak amplitudes, areas under amplitude profiles, and soliton distances.

23.5-4 **Solitons in an Optical Fiber.** Show that the product of the peak intensity and dispersion length for the fundamental soliton is a constant: $I_0|z_0| = \lambda_o/4\pi n_2$. For a silica-glass fiber with Kerr coefficient $n_2 = 3.19 \times 10^{-20}$ m^2/W, determine the peak intensity I_0 for a dispersion distance $|z_0| = 30$ km.

23.6-1 **Measurement of a Gaussian Pulse.** A Gaussian transform-limited optical pulse with a 50-fs FWHM and a central frequency corresponding to a wavelength of 800 nm is measured with the help of the intensity correlator illustrated in Fig. 23.6-8.

(a) Determine the shape and FWHM of the measured autocorrelation function.

(b) It has been suggested that the measurement can be improved if one of the pulses, say the one traveling in the upper branch, is deliberately stretched by passage through a silica-glass fiber. If the pulse is to be stretched by a factor of 5, what length of fiber is required to do so, given that silica glass has a dispersion coefficient $D_\lambda = -110$ ps/km-nm at 800 nm? What would be the width of the modified correlation function after insertion of the fiber?

(c) Consider applying this notion to the nonlinear interferometer displayed in Fig. 23.6-13, with the fiber placed in the upper branch. Describe the possible merits and difficulties of using this approach as a tool for pulse measurement.

23.6-2 **Interferometer with a Two-Photon Absorbing Detector.** An interferometer that uses a two-photon absorber as a detector (Prob. 19.1-5) provides a measurement of the function $R_U'^{(2)}(\tau)$ specified in (23.6-21). Compare this interferometer with a nonlinear interferometer that makes use of a second-harmonic generator followed by a conventional detector, which provides a measurement of the function $R_U^{(2)}(\tau)$ given in (23.6-13). *Hint*: Expand (23.6-21) in a form similar to that of (23.6-14) and compare the different terms.

CHAPTER

24

OPTICAL INTERCONNECTS AND SWITCHES

24.1 OPTICAL INTERCONNECTS 1166
- A. Free-Space Refractive and Diffractive Interconnects
- B. Guided-Wave Interconnects
- C. Nonreciprocal Optical Interconnects
- D. Optical Interconnects in Microelectronics and Computer Systems

24.2 PASSIVE OPTICAL ROUTERS 1178
- A. Wavelength-Based Routers
- B. Polarization- , Phase- , and Intensity-Based Routers

24.3 PHOTONIC SWITCHES 1187
- A. Space-Switch Architectures
- B. Implementations of Photonic Space Switches
- C. All-Optical Space Switches
- D. Wavelength-Selective Switches
- E. Time-Domain Switches
- F. Packet Switches

24.4 PHOTONIC LOGIC GATES 1211
- A. Bistable Systems
- B. Principles of Optical Bistability
- C. Bistable Optical Devices

The development of optical interconnects and photonic switches began in earnest in the 1980s under the aegis of Bell Laboratories, an organization created by AT&T in 1925. Bell Laboratories became part of Lucent Technologies in 1996, part of Alcatel–Lucent in 2006, and then part of Nokia in 2016.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Interconnections and switches are essential components of distributed systems such as communication systems and networks (*telecom* and *datacom*) and computing systems (*computer-com*). As displayed in Fig. 24.0-1, the scale of the interconnection distances in these different venues extend over a broad range: 1–100 km for telecom; 10–1000 m for datacom; and 1–1000 cm for computer-com. The interconnection links may be passive (fixed), or they may be active (reconfigurable) and controllable by switches.

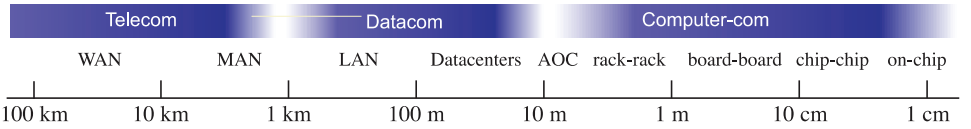


Figure 24.0-1 Typical interconnection distances for telecom [wide-area networks (WANs) and metropolitan-area networks (MANs)]; datacom [local-area networks (LANs), datacenters, and active optical cables (AOCs)]; and computer-com [rack-to-rack, board-to-board, chip-to-chip, and on-chip interconnects].

Optical fibers have emerged as the preferred medium for telecom and datacom interconnects, and indeed many types of photonic switches have been developed to support active links. The reach of optical-fiber technology has developed more slowly for short-distance applications, yet optical interconnects and switches have continued to make inroads in computer-com, including computer rack-to-rack, board-to-board, chip-to-chip, and even on-chip links. The principal challenges in implementing short-distance interconnects are the greater link density and the attendant higher dissipated power density. While several decades of research in digital optical computing have not yielded commercial products that are competitive with electronic computing, a byproduct of this effort has been the development of a number of technologies for optical interconnects and logic gates.

A generic interconnection system is displayed in Fig. 24.0-2. Light entering each of a set of M points at the input is directed to one or several of a set of N points at the output. If the connections are fixed and independent of the content of the incoming light, the system is referred to as an **interconnect**. On the other hand, if the incoming light is routed to different destinations that depend on its content, the system is called a **router**. If the connections are reconfigurable, in response to an external control signal, the system is called a **switch**.

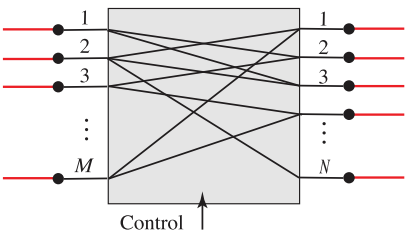


Figure 24.0-2 A generic $N \times M$ system may function as a passive interconnect, a router, or a switch.

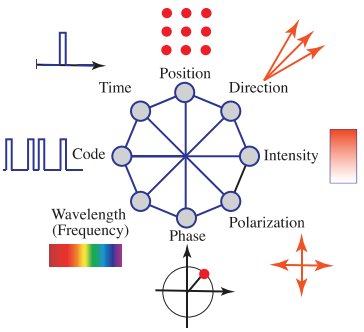


Figure 24.0-3 Attributes of an optical beam that can be used for modulation, multiplexing, routing, and switching.

Light has multiple attributes, as illustrated in Fig. 24.0-3: position (space); time (for optical pulses or specific sequences of optical pulses that form a code); wavelength (or frequency); polarization; direction; intensity; and phase (for coherent waves). These attributes serve as resources that can be used to enrich the interconnection and switching fabric. When the different interconnection points schematized in Fig. 24.0-2 refer to space, time, wavelength, or polarization, the system is called a **space-domain switch** (or simply a **space switch**), a **time-domain switch**, a **wavelength-domain switch**, or a **polarization switch**, respectively. For example, a time-domain switch may be engaged to transfer a signal from one time slot to another, while a wavelength-domain switch converts a signal from one wavelength to another. Switches controlled by an address coded in each packet of incoming data are known as **packet switches**.

An interconnection for which an attribute, such as wavelength, of the incoming light dictates the destination output ports is known as a **passive optical router**. Such routers enable **multiplexing**: multiple data streams can be transported on a single link by using one attribute as a distinguishing marker while modulating another attribute to encode information. For example, the optical intensities of different wavelengths in a single light beam may be modulated by different data streams that are ultimately separated (demultiplexed) at the destination by means of a wavelength-sensitive passive optical router. The introduction of **wavelength-division multiplexing** (WDM) has substantially advanced modern optical fiber communication systems, as discussed in Sec. 25.5B, and has motivated the development of special wavelength-based photonic switches.

This Chapter

This chapter introduces the basic principles associated with optical interconnects, passive optical routers, photonic switches, and photonic logic gates. Many of the fundamental principles of optics and photonics introduced in earlier chapters find use here, including Fourier optics and holography, guided-wave and fiber optics, semiconductor optics, acousto-optics, electro-optics, nonlinear optics, and ultrafast optics.

Optical interconnects via free-space, planar photonic circuits, and optical fibers are considered in Sec. 24.1. The input beams are directed to prescribed output ports regardless of their attributes and the information they carry.

Passive optical routers are described in Sec. 24.2. Each input optical beam is directed to one or more output ports based on beam attributes such as wavelength, polarization, or intensity. Different wavelength components in a single beam may, for example, be routed to separate output ports, in which case the device serves as a wavelength-division demultiplexer. The inverse of this operation, in which beams with different wavelengths are combined into a single optical beam, is implemented by a multiplexer.

Photonic switches are considered in Sec. 24.3. The simplest example is an ON–OFF switch that can be directed by a control signal to connect or disconnect two ports (i.e., transmit or block a beam of light), or a switch that selectively directs a beam to one of two possible locations, regardless of the data content or attributes of the beam. Following an introduction to the types and properties of these switches, we provide a brief overview of the different technologies used to implement them; these include mechano-optic, electro-optic, semiconductor-photonic, thermo-optic, and all-optical devices. Time-domain switches and packet switches are also described.

Photonic logic gates based on bistable optical devices are examined in Sec. 24.4. These switches have memory so that the output takes on one of two (or several) values, depending both on the current value and on the previous history of the input.

24.1 OPTICAL INTERCONNECTS

Digital signal-processing and computing systems contain large numbers of interconnected gates, switches, and memory elements. In electronic systems, the interconnections are effected by means of conducting wires, coaxial cables, or conducting channels within semiconductor integrated circuits. Photonic interconnections may be similarly realized by the use of optical waveguides with integrated-photonic or fiber-optic couplers. Free-space light beams may also be used for interconnections, in which microlenses or diffractive optical elements direct these beams. This latter option is not available in electronic systems since electron beams require vacuum and cannot cross one another without mutual repulsion.

Figure 24.1-1 illustrates a number of configurations for interconnects (also called **couplers**). Each input port is connected to one or many output ports, and vice-versa. For example, in the T-coupler or fan-out configuration, the input port is connected to each of the output ports. In the 3-dB or star coupler, each input port is connected to each and every output port.

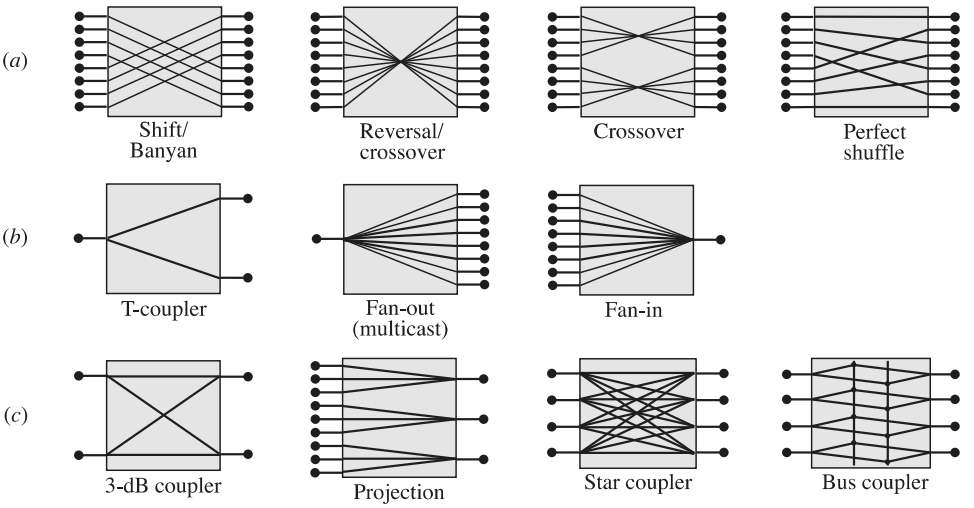


Figure 24.1-1 Representative examples of interconnect configurations. (*a*) One-to-one. (*b*) One-to-many or many-to-one. (*c*) Many-to-many.

Interconnection Matrix

The diagrams displayed in Fig. 24.1-1 are solely schematic connectivity diagrams; they do not specify quantitative relations among the optical fields or intensities at the connected ports. For linear coherent optical interconnects, the optical field $U_\ell^{(o)}$ at the ℓ th output port ($\ell = 1, 2, \ldots, N$) is related to the optical fields $U_m^{(i)}$ at the input ports, $m = 1, 2, \ldots, M$, via the superposition

$$U_\ell^{(o)} = \sum_{m=1}^{M} T_{\ell m} U_m^{(i)}, \tag{24.1-1}$$

where the weights $\{T_{\ell m}\}$ are complex numbers that define an **interconnection matrix** $\mathbf{T}$. For example, the 2×2 3-dB coupler shown in Fig. 24.1-1(*c*) is described by

$$\mathbf{T} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & j \\ j & 1 \end{bmatrix}, \tag{24.1-2}$$

and which harks back to the scattering matrix for an ideal beamsplitter (7.1-18) and the transmission matrix for phase-matched guided waves in coupled-mode theory (9.4-11). For this device, the optical power carried in by one beam (in the absence of the other) is divided equally between the two outgoing beams. Other interconnects may be similarly described. The interconnection matrix of a cascade of interconnects may be determined by making use of matrix multiplication, as described in Sec. 7.1A.

Since the light is assumed to be coherent, the phase relations between the incoming beams and the phases introduced by the elements of the interconnection device play important roles. Indeed, interferometric effects are often used to redistribute the incoming power among the output ports in a prescribed manner. If the light is incoherent, on the other hand, then the intensity (and hence the power) at each output port is a weighted superposition of the intensities (powers) at the input ports (see Sec. 12.3B):

$$P_\ell^{(o)} = \sum_{m=1}^{M} |T_{\ell m}|^2 P_m^{(i)}. \tag{24.1-3}$$

In this case, the powers at the output and input ports of a 3-dB coupler are related by an interconnection matrix whose elements are all equal to $1/2$.

Key performance specifications of practical couplers include the following power ratios, usually expressed in dB [i.e., $-10\log(1/\text{power ratio})$]:

- The **insertion loss** describes the port-to-port power transmittance, ideally 0 dB for a lossless path.
- For a coupler distributing power among multiple output ports, the **splitting ratio** is the ratio of the power at one output port to the power at all output ports. For example, for an ideal 3-dB coupler, the splitting ratio is -3 dB.
- The **crosstalk** is the ratio of the undesired power received at an output port to the input power directed to another output port(s).
- The **excess loss** is the ratio of the total output power to the total input power.

Nonreciprocal Interconnects: Isolators and Circulators

The designation of the ports of an interconnect as input or output ports implies a specific direction of transmission — from input to output (from left to right in the examples in Fig. 24.1-1). Certain interconnects are reciprocal, i.e., if the transmission is directed instead from the output ports to the input ports, the interconnection matrix remains the same. Otherwise, the interconnect is **nonreciprocal**.

Isolators. The simplest example of a nonreciprocal interconnect is a 1×1 unidirectional link that transmits in only one direction, as illustrated in Fig. 24.1-2(a). This is often implemented by using an optical isolator, much like a diode or a one-way valve (see Sec. 6.4B). The performance of an isolator is specified by the **insertion loss** (power transmittance in the forward direction in dB) and the **reverse isolation** (power transmittance in the reverse direction in dB).

Multiport Nonreciprocal Interconnects. The input/output designation is not applicable when a port plays a dual role, as transmitter and receiver. The interconnect is then designated simply by the number of ports. Figure 24.1-2(b) and (c) are example of 3-port interconnects using unidirectional links. These interconnects are used in duplex (two-way) communication systems, as depicted in the 4-port interconnect in Fig. 24.1-2(e). In another 4-port system, shown in Fig. 24.1-2(d), the connections between the left and right ports are in the parallel configuration in the forward direction (left to right), and in the cross configuration in the backward direction (right to left).

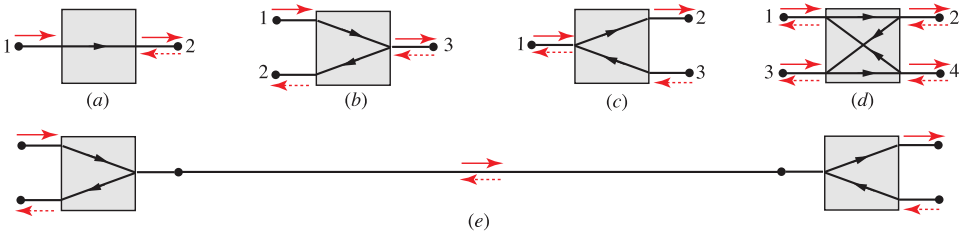


Figure 24.1-2 (a) A 2-port unidirectional link (isolator). (b), (c) A 3-port interconnect using two unidirectional links. (d) A 4-port nonreciprocal interconnect. (e) A 4-port interconnect for duplex (bidirectional) communications.

Circulators. Another example of nonreciprocal interconnects is the optical circulator. This is an interconnect with three or more ports connected by unidirectional links pointing in the same direction. As illustrated in Fig. 24.1-3, the 4-port circulator is equivalent to the interconnect in Fig. 24.1-2(d). Circulators find many applications in communication systems and networks. They are used, for example, in optical add–drop multiplexers (OADMs), as described in Sec. 24.2A (Fig. 24.2-3).

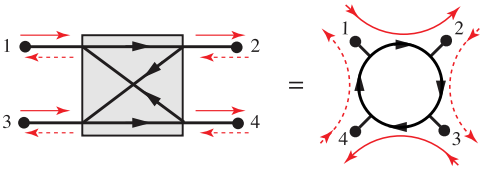


Figure 24.1-3 A 4-port circulator represented by two equivalent configurations.

A. Free-Space Refractive and Diffractive Interconnects

Refractive interconnects. Conventional optical components (mirrors, lenses, prisms, etc.) are routinely used as interconnects in optical systems. Consider, as an example, a simple imaging system in which a lens connects points in the object and image planes. To appreciate the enormous density of such interconnections in a well-designed imaging system, observe that as many as 1000×1000 independent points per mm^2 in the object plane are optically connected by means of the lens to a corresponding 1000×1000 points per mm^2 in the image plane. Implementing such connectivity electrically would require one million nonintersecting and suitably insulated conducting channels per mm^2.

Standard optical components may be used to implement special interconnects, such as shift, reversal, crossover, shuffle, fan-in, fan-out, star coupling, and projection, as illustrated in Fig. 24.1-4 (see also Fig. 24.1-1). Bulk-optical components can be miniaturized to a micro-optical scale with the help of miniature beamsplitters, lenses, graded-index rods, prisms, filters, and gratings, which are compatible with optical fibers for light transmission.

Diffractive interconnects. Arbitrary optical interconnection maps require the design of custom optical components that may be quite complex and impractical. However, computer-generated holograms comprising a large number of phase-grating segments with different spatial frequencies and orientations have been successfully used to create high-density optical interconnections. A phase grating is a thin optical element whose complex amplitude transmittance is a two-dimensional periodic function of unit amplitude. The simplest phase grating has complex amplitude transmittance $\mathrm{t}(x, y) =$

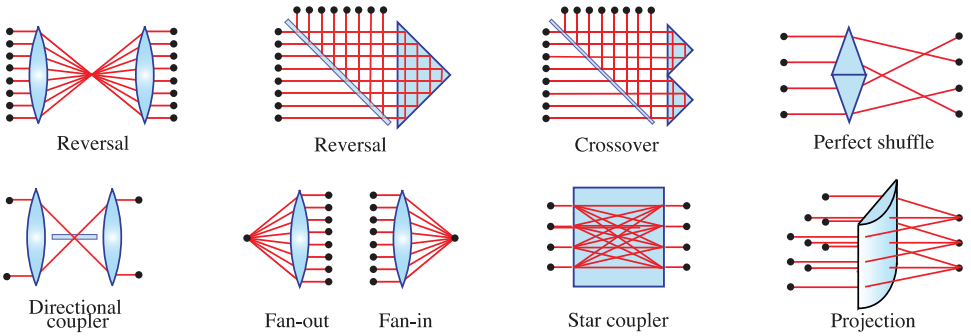


Figure 24.1-4 Examples of simple optical interconnects created by conventional refractive optical components: A prism bends parallel optical rays preferentially and establishes an ordered interconnection map corresponding to a reversal or crossover. Two appropriately oriented prisms perform a perfect-shuffle — an operation used in sorting algorithms and in the fast Fourier transform (FFT). A lens establishes a fan-in, a fan-out, or a reversal. A beamsplitter together with two lenses creates a directional coupler. A glass rod serves as a star coupler. An astigmatic optical system, such as a cylindrical lens, implements a projection by connecting points of each row at the input plane to one point at the output plane.

$\exp[-j2\pi(\nu_x x + \nu_y y)]$, where ν_x and ν_y are the spatial frequencies in the x and y directions, respectively; they determine the period and orientation of the grating. It was shown in Secs. 2.4B and 4.1A that when a coherent optical beam of wavelength λ is transmitted through such a grating, it undergoes a phase shift that causes the beam to tilt by the angles $\sin^{-1}\lambda\nu_x \approx \lambda\nu_x$ and $\sin^{-1}\lambda\nu_y \approx \lambda\nu_y$, when $\lambda\nu_x \ll 1$ and $\lambda\nu_y \ll 1$, as illustrated in Fig. 24.1-5. Varying the spatial frequencies ν_x and ν_y (i.e., the periodicity and orientation of the grating) alters the tilt angles.

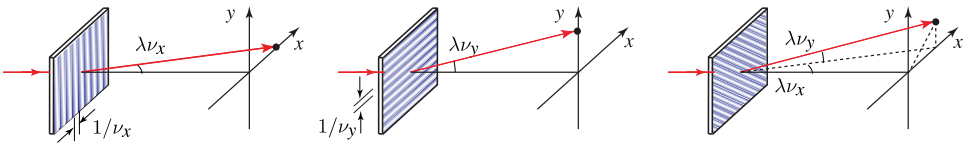


Figure 24.1-5 Bending of an optical wave as a result of transmission through a phase grating. The deflection angles, assumed to be small, depend on the spatial frequency and orientation of the grating.

As described in Sec. 4.1A and illustrated in Fig. 4.1-5, this principle may be used to implement an arbitrary interconnection map by constructing a phase grating comprising a collection of grating segments with different spatial frequencies. As displayed in Fig. 24.1-6, optical beams transmitted through the different segments undergo different tilts, in accordance with the desired interconnection map.

If the grating segment located at position (x, y) has spatial frequencies $\nu_x = \nu_x(x, y)$ and $\nu_y = \nu_y(x, y)$, the angles of tilt are approximately $\lambda\nu_x$ and $\lambda\nu_y$, respectively. The beam then impinges on the output plane at the position (x', y') that satisfies

$$\frac{x' - x}{d} \approx \lambda\nu_x, \qquad \frac{y' - y}{d} \approx \lambda\nu_y, \tag{24.1-4}$$

where d is the distance between the hologram and the output plane, and all angles are assumed to be small. Given the interconnection map, i.e., the desired relation between

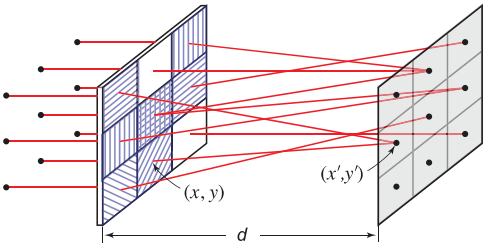


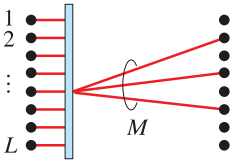
Figure 24.1-6 Holographic interconnection map created by an array of phase gratings of different periodicities and orientations.

(x', y') and (x, y), the requisite spatial frequencies ν_x and ν_y at each position are determined by using (24.1-4).

Holographic interconnection devices are capable of establishing one-to-many or many-to-one interconnections (i.e., connecting one point to many points, or *vice versa*). In Fig. 24.1-6, for example, the center grating element is seen to be a superposition of two harmonic functions so that its complex amplitude transmittance $\mathrm{t}(x, y) \propto \exp[-j2\pi(\nu_{x1}x + \nu_{y1}y)] + \exp[-j2\pi(\nu_{x2}x + \nu_{y2}y)]$; the incident beam is thus split equally into two components, one tilted at the angles $(\lambda\nu_{x1}, \lambda\nu_{y1})$ and the other at the angles $(\lambda\nu_{x2}, \lambda\nu_{y2})$, where all angles are small. Weighted interconnections may be realized by assigning different weights to the different gratings. Arbitrary interconnections may therefore be created by appropriate selection of the grating spatial frequencies at each point of the hologram.

EXERCISE 24.1-1

Interconnection Capacity. The space–bandwidth product of a square hologram of size $a \times a$ is the product $(Ba)^2$, where B is the highest spatial frequency (lines/mm) that may be printed on the hologram. Show that if the hologram is used to direct each of L incoming beams to M directions, the product ML cannot exceed $(Ba)^2$,



$$ML \le (Ba)^2.$$

Hint: Use an analysis similar to that presented in Sec. 20.2C in connection with acousto-optic interconnection devices [see (20.2-9)].

What is the maximum number of interconnections per mm^2 if the highest spatial frequency is 1000 lines/mm and if every point in the input plane is connected to every point in the output plane?

In the limit in which the grating elements have infinitesimal areas, the result is a continuous (instead of discrete) interconnection map: a geometric coordinate transformation rule that transforms each point (x, y) in the input plane into a corresponding point in the output plane (x', y'). If the desired transformation is defined by the two continuous functions

$$x' = \psi_x(x, y), \qquad y' = \psi_y(x, y), \tag{24.1-5}$$

the grating frequencies must vary continuously with x and y as in a frequency-modulated (FM) signal (see Fig. 24.1-7). Assuming that the grating has a transmittance

$\mathrm{t}(x, y) = \exp[-j\varphi(x, y)]$, the associated local (or instantaneous) spatial frequencies are

$$2\pi\nu_x = \frac{\partial\varphi}{\partial x}, \qquad 2\pi\nu_y = \frac{\partial\varphi}{\partial y}, \tag{24.1-6}$$

as provided in Sec. 4.1A. Substituting (24.1-6) into (24.1-4) then leads to a pair of partial differential equations,

$$\boxed{\frac{\psi_x(x, y) - x}{d} = \frac{\lambda}{2\pi}\frac{\partial\varphi}{\partial x}, \qquad \frac{\psi_y(x, y) - y}{d} = \frac{\lambda}{2\pi}\frac{\partial\varphi}{\partial y},} \tag{24.1-7}$$

which may be solved to determine the grating phase function $\varphi(x, y)$.

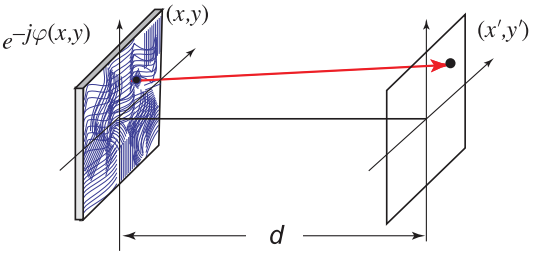


Figure 24.1-7 Diffraction from a phase hologram as a continuous interconnection system.

EXAMPLE 24.1-1. ***Fan-In Map.*** Suppose that all points (x, y) in the input plane are to be steered to the point $(x', y') = (0, 0)$ in the output plane, so that a fan-in interconnection map is created. Substituting $\psi_x(x, y) = \psi_y(x, y) = 0$ in (24.1-7) and solving the two partial differential equations, we obtain $\varphi(x, y) = -\pi(x^2 + y^2)/\lambda d$. Not surprisingly, this is nothing but the phase shift introduced by a lens of focal length d [see (2.4-9)].

EXERCISE 24.1-2

The Logarithmic Map. Show that the logarithmic coordinate transformation

$$x' = \psi_x(x, y) = \ln x, \qquad y' = \psi_y(x, y) = \ln y \tag{24.1-8}$$

is realized by a hologram with the phase function

$$\varphi(x, y) = \frac{2\pi}{\lambda d}\left(x \ln x - x - \tfrac{1}{2}x^2 + y \ln y - y - \tfrac{1}{2}y^2\right). \tag{24.1-9}$$

Once the appropriate phase $\varphi(x, y)$ is decided, the optical element is fabricated by using the techniques of **computer-generated holography**. This approach allows a complex function $\exp[-j\varphi(x, y)]$ to be encoded with the help of a binary function that takes on only two values, 1 and 0, or 1 and -1, for example. This is similar to encoding an image by making use of a collection of black dots with density or sizes that vary in proportionality to the local gray value of the image (an example is the halftone process

used for printing newspaper images). Software is used to print the binary image on a mask (a transparency) that plays the role of the hologram. The binary image may also be printed by etching grooves in a substrate, which modulate the phase of an incident coherent wave via a technology known as **surface-relief holography**.

Dynamic (reconfigurable) interconnections may be constructed by using acousto-optic or magneto-optic devices, although the number of interconnection points is far smaller than that achievable with holographic gratings. Dynamic holographic interconnections may also be implemented by making use of nonlinear optical processes such as four-wave mixing in photorefractive materials. In this approach, two waves interfere to create a grating, from which a third wave is reflected. The angle between the two waves determines the spatial frequency of the grating, which in turn determines the tilt of the reflected wave (see Secs. 21.4 and 22.3E).

B. Guided-Wave Interconnects

Optical interconnects are implemented in integrated photonics by patterning optical waveguides in silicon or $LiNbO_3$ substrates (see Sec. 9.4B), much as metal wires are implemented in integrated electronics. Examples are illustrated in Fig. 24.1-8; combinations and cascades of these basic interconnects can be used to create more complex interconnects.

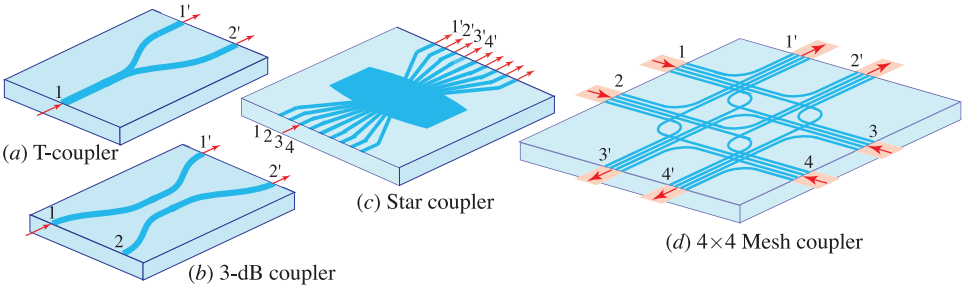


Figure 24.1-8 Integrated-photonic devices implementing some of the interconnects depicted in Fig. 24.1-1. The mesh coupler is another implementation of the star coupler.

Waveguide couplers are used to distribute optical power in prescribed proportions. The coupler portrayed in Fig. 24.1-8(b), for example, is described by an interconnection matrix that is identical to the transmission matrix of (9.4-11),

$$\mathbf{T} = \begin{bmatrix} \cos \mathcal{C}L & -j \sin \mathcal{C}L \\ -j \sin \mathcal{C}L & \cos \mathcal{C}L \end{bmatrix}, \tag{24.1-10}$$

where $\mathcal{C}$ is the coupling coefficient and L is the interaction length. The power arriving at input port 1 is therefore apportioned between output ports 1 and 2 in accordance with the factors $\cos^2 \mathcal{C}L$ and $\sin^2 \mathcal{C}L$, respectively. For $\mathcal{C}L = \pi/4$, the input power is equally divided between the output ports so that the device becomes a 3-dB coupler.

Fiber-optic interconnects are widely used in optical fiber technology, particularly in optical fiber communications, and the devices portrayed in Fig. 24.1-9 parallel those depicted in Fig. 24.1-8 for integrated photonics. Indeed, the coupler shown in Fig. 24.1-9(b) is described by the same interconnection matrix as that in Fig. 24.1-8(b), namely (24.1-10).

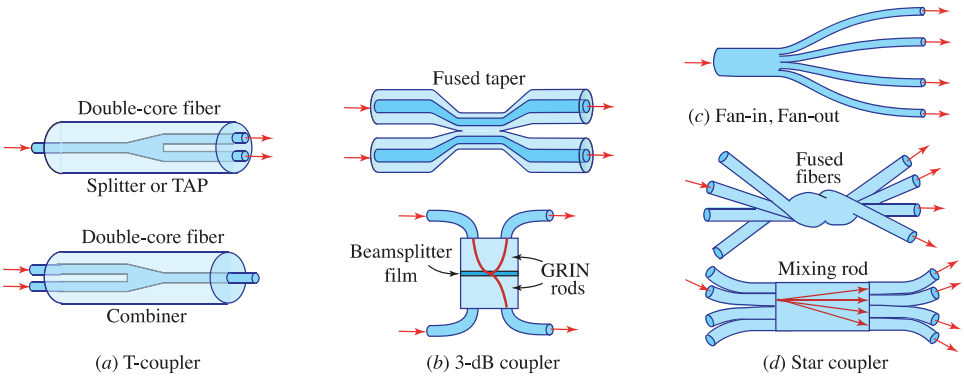


Figure 24.1-9 Fiber-optic couplers that implement some of the interconnects displayed in Fig. 24.1-1. (*a*) Double-core fiber used as a T-coupler, splitter, or combiner. (*b*) 3-dB coupler comprising two fused fibers; another version makes use of a pair of GRIN-rod lenses separated by a beamsplitter film. (*c*) Fan-in or fan-out. (*d*) Star coupler that relies on fused fibers; another version makes use of a mixing rod, a slab of glass through which light from one fiber is dispersed to reach all other fibers.

C. Nonreciprocal Optical Interconnects

Optical implementations of nonreciprocal interconnects are based primarily on nonreciprocal polarization devices. As explained in Sec. 6.6D, an **optical isolator** may be implemented by use of a $45°$ Faraday rotator sandwiched between a pair of polarizers oriented at $45°$ with respect to each another. Linearly polarized light is transmitted in the forward direction but blocked in the reverse direction.

A $45°$ Faraday rotator followed by a half-wave retarder is also a useful nonreciprocal device, as portrayed in Fig. 6.6-6. The polarization of a forward-traveling linearly polarized wave whose plane of polarization is oriented at $22.5°$ with respect to the fast axis of the retarder is not altered, while the plane of polarization of the backward-traveling wave is rotated by $90°$. This device may be used in conjunction with polarizing beamsplitters to implement nonreciprocal interconnects, such as 3-port and 4-port (circulator) devices, as illustrated in Figs. 24.1-10(*a*) and (*b*), respectively.

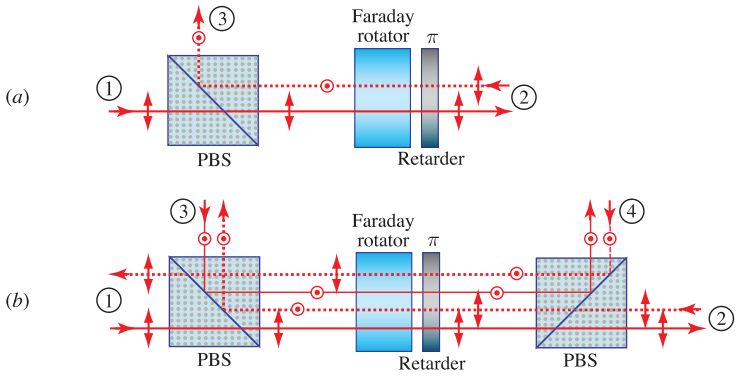


Figure 24.1-10 (*a*) Implementation of the 3-port nonreciprocal interconnect of Fig. 24.1-2(*c*) by means of a polarizing beamsplitter (PBS) together with the combination of a Faraday rotator and a half-wave (π) retarder. Light travels from port ① to port ②, and from ② to port ③. (*b*) Implementation of the 4-port nonreciprocal circulator of Fig. 24.1-3 by means of a pair of polarizing beamsplitters together with the combination of a Faraday rotator and a half-wave retarder. Light travels from port to port in accordance with ① → ② → ③ → ④ → ①.

D. Optical Interconnects in Microelectronics and Computer Systems

The prospect of using optical interconnects in place of conventional electrical interconnects in microelectronics and computer systems has engendered substantial research and development efforts over the past several decades. Successful implementations include backplane-to-backplane, board-to-board, chip-to-chip, and intrachip interconnects, as illustrated schematically in Fig. 24.1-11. Integrated photonics has been fueling the transition from electronic to optical interconnects in venues such as datacenters and high-performance computers. Some of the strictures that apply to components used in optical fiber communication systems (Sec. 25.1) are not restrictive in the context of optical interconnects because of their short reach.

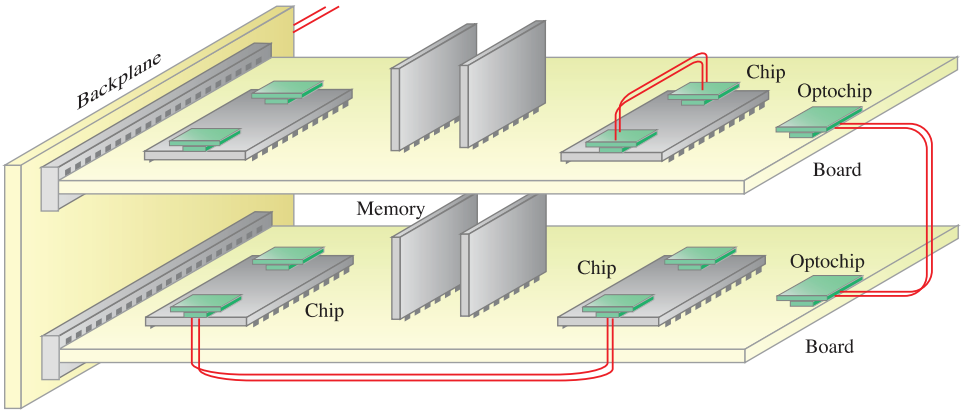


Figure 24.1-11 Schematic illustration of optical interconnects in microelectronics and computer systems: backplane-to-backplane, board-to-board, chip-to-chip, and intrachip. Optochips are chip-scale optical transceivers.

Inter-Board and Inter-Chip Optical Interconnects

With the successful implementation of fiber optics for computer-to-computer communications in local area networks and datacenters (Sec. 25.5), systems employing optical fibers for backplane-to-backplane and board-to-board communications have been developed and implemented in high-performance computers. Such short-reach optical-fiber links can operate at data rates in excess of 100 Gb/s, far greater than those of electrical links, and the technology is well established (Chapter 25).

Board-to-board interconnects may also be implemented by making use of integrated-optic waveguides placed on the backplane. In the example illustrated in Fig. 24.1-12(a), multiple boards are connected in a bus configuration that serves as a star coupler connecting each board to all others. Board-to-board free-space optical links using reflecting mirrors or holographic optical elements have also been considered, but they are cumbersome and inefficient. On-board chip-to-chip optical interconnects using on-board mirrors and planar waveguides, as depicted in Fig. 24.1-12(b), have also been developed.

Optochips. A principal component of optical interconnects for microelectronics and high-performance computing is the **optochip**, a chip-scale **optical transceiver** (transmitter/receiver) comprising an array of light sources (e.g., VCSELs) and an array of photodetectors (PDs), together with their associated transmitter (Tx) and receiver (Rx) electronic circuitry. A schematic example of an optochip coupled to an on-board planar

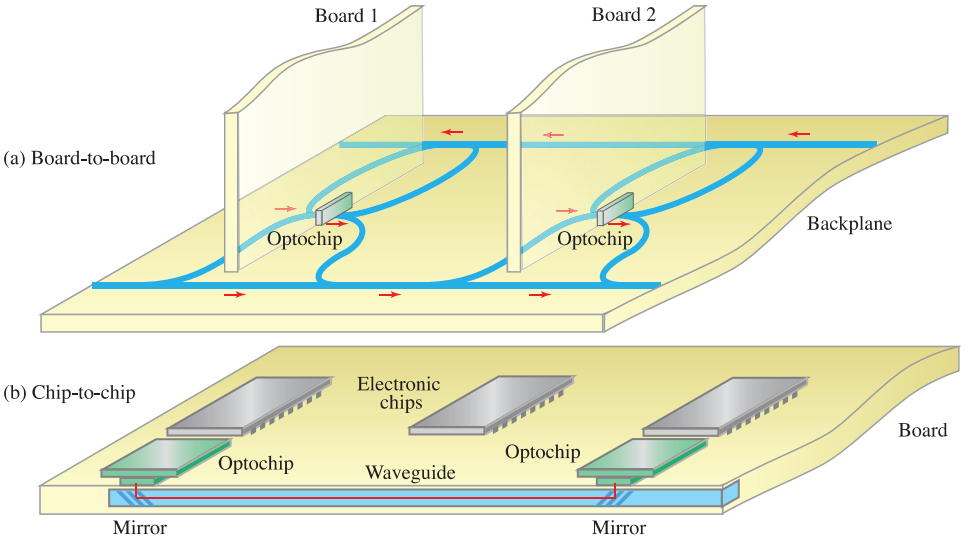


Figure 24.1-12 (*a*) Multiple boards interconnected via planar waveguides in a bus configuration on the backplane. (*b*) Chip-to-chip optical interconnect using on-board planar waveguides and mirrors. Optochips are chip-scale optical transceivers.

waveguide via mirrors is displayed in Fig. 24.1-13(*a*). Four chips (VCSEL, PD, Tx, and Rx) are mounted on a Si carrier; light is transmitted between the VCSEL/PD arrays and the waveguide via holes in the carrier and is focused by means of integrated lenses. In another example, portrayed in Fig. 24.1-13(*b*), an optochip is coupled to an array of multimode optical fibers (MMFs). In this case the VCSEL and PD arrays are supported by an organic carrier and are attached to an electronic integrated circuit (IC) that contains the Tx and Rx circuitry. Light to and from the fibers passes through holes (*vias*) in the IC and an array of integrated lenses serve as the focusing element. VCSELs can generate multimode emission with a numerical aperture compatible with the MMFs.

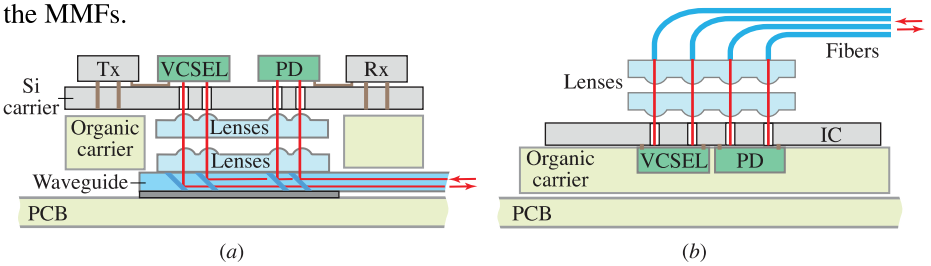


Figure 24.1-13 Schematic illustrations of optochips mounted on printed circuit boards (PCBs). Light is generated by a VCSEL array driven by a transmitter circuit (Tx), and detected by a photodetector (PD) array connected to a receiver circuit (Rx). (*a*) The four chips are mounted on a Si carrier; light from/to the VCSEL/PD array is directed through holes in the Si carrier, integrated lenses, and mirrors in the on-board planar waveguide. (*b*) The VCSEL and PD arrays are attached to an electronic IC that contains the Tx and Rx circuitry; they are linked to a fiber array via holes (vias) in the IC and an integrated lens array.

An optochip using an array of 24 VCSELs and 24 PDs, each operating at a data rate of 15 Gb/s, provides communication at an overall rate of 360 Gb/s. Transceivers that operate at data rates of 100 Gb/s are readily available for datacenter interconnects and high-performance computing. These compact modules incorporate lasers, modulators, splitters, wavelength multiplexers, wavelength demultiplexers, and photodetectors.

Intrachip Optical Interconnects

The use of ultra-short-reach optical links to connect points *within* a chip is a more challenging enterprise. The use of intrachip optical interconnects is motivated by advances in high-speed, high-density microelectronics circuitry and the parallel-processing architectures found in high-performance computers, which demand high-quality interconnectivity to avoid communications bottlenecks. In ultra-large-scale integrated circuits (ULSI), the interconnects occupy a substantial portion of the available chip area so considerable effort must be devoted to equalizing interconnect lengths in order to minimize interconnection time delays (which can exceed gate delays). Intrachip optical interconnects incorporating carefully designed optochips are commercially available.

An intrachip optical interconnect features three key components: 1) an electronic-to-optical (E/O) transducer (transmitter) modulated by the electrical signal at a point within the chip; 2) a point-to-point optical link that carries the signal to another point on the chip; and 3) an optical-to-electronic (O/E) transducer (receiver) that detects the signal at the destination point. Ideally, this device should take the form of a **photonic integrated circuit (PIC)** in which all three components are monolithically integrated within the silicon substrate of the chip and are compatible with CMOS (complementary metal-oxide-semiconductor) technology, which is the principal platform for integrated electronics. In fact, silicon photodiodes (Fig. 19.3-7) can be readily embedded in silicon chips, and silicon-on-insulator (SOI) optical waveguides (Sec. 9.3) can serve as high-efficiency conduits (although the on-chip real-estate for such guides may be scarce).

Si-based on-chip sources. The principal difficulty in creating a monolithic intrachip-interconnect PIC lies with the fabrication of the laser-diode transmitter inasmuch as Si is an indirect-bandgap material from which light sources cannot be efficiently fashioned (Fig. 17.2-7). Nevertheless, as discussed in Sec. 18.1D in the context of **silicon photonics**, silicon-based on-chip light sources can be implemented via three approaches: 1) *flip-chip integration* of III–V laser diodes onto a separately fabricated silicon platform with the help of solder bumps; 2) *heterogeneous integration* of III–V lasers into prepatterned silicon circuits (*hybrid approach*); and 3) *direct heteroepitaxial growth* of III–V lasers on Si substrates using intermediate buffer layers to minimize dislocations in the light-emitting region. Each of these approaches has its own limitations and merits. Hybrid integration is widely used (Fig. 24.1-14) but direct heteroepitaxy may be the most attractive alternative going forward.

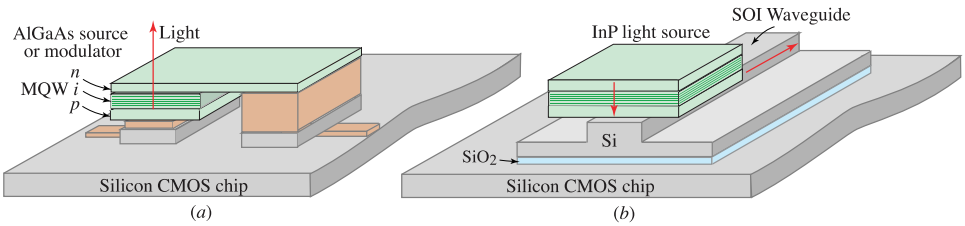


Figure 24.1-14 Light sources integrated with Si CMOS chips via heterogeneous integration (hybrid approach). The III–V source chips are fabricated separately and then integrated with the Si structure. (*a*) An AlGaAs optical source is bonded to a silicon chip in a surface normal architecture. (*b*) An InP light source is bonded to a silicon chip and the emitted light is coupled into an on-chip silicon-on-insulator (SOI) ridge waveguide.

Alternative on-chip sources. Group-IV materials other than Si can be used to fabricate light sources (Sec. 17.1B and Fig. 17.1-6); GeSn alloys, for example, are

compatible with silicon and CMOS technology. Various types of compact, low-threshold lasers are also good candidates for use in optical interconnects; these include VCSELs (Sec. 18.5A), microdisk and microring lasers (Sec. 18.5B), photonic-crystal lasers (Sec. 18.5C), and nanocavity lasers (Sec. 18.6). Also, directly modulated optical sources can be replaced with externally illuminated electro-optic (Sec. 21.1B) or electroabsorption (Sec. 21.5) modulators that derive their modulation signals directly from the local electrical signals within the chip. This approach, although inefficient and cumbersome, has the merit that it decouples the modulation and light-generation mechanisms, thereby allowing them to be independently optimized.

On-chip holographic interconnects. Free-space optical links using external devices such as holograms have also been considered for intrachip interconnects, as illustrated in Fig. 24.1-15(a). A configuration that does not require the use of on-chip transmitters is a one-way interconnect between one, or several, external point(s) and points on the chip, as illustrated in Fig. 24.1-15(b). A useful application for such a configuration is optical clock distribution, where a signal from an external clock modulates an external light source that broadcasts the signal to multiple photodetectors on the chip using a reflection hologram. This ensures accurate synchronization of high-speed synchronous circuits and mitigates the problem of clock skew that results from differential time delays. The hologram may, of course, be eliminated and the light "broadcast" directly to all points on the chip. While this creates a robust system that is insensitive to misalignment, the power efficiency is low since a large portion of the optical power is lost.

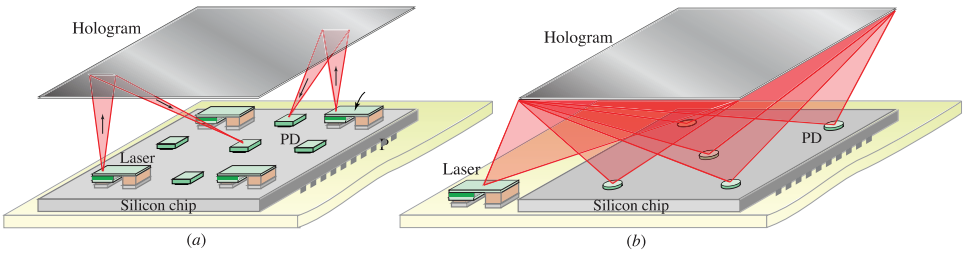


Figure 24.1-15 (a) Interconnects between on-chip sources and detectors via an external reflection hologram used as an interconnect element. (b) One-way interconnects directing clock pulses from an external light source to photodetectors on a silicon chip.

Rationale for Chip Optical Interconnects

Optical interconnects offer a number of advantages for inter- and intrachip interconnects that stem principally from the short wavelength of light and its corresponding high frequency (e.g., 20–50 THz), which is substantially greater than the bandwidth of the transmitted data. Electronic interconnects, in contrast, use baseband signals at far lower frequencies (e.g., in the GHz range). The principal advantages of optical interconnects, in terms of their larger bandwidths, shorter delays, higher densities, and lower power consumption, are set forth below.

- *Bandwidth.* The bandwidth of an electronic strip line of length ℓ and cross-sectional area A placed above a ground plane is proportional to the ratio A/ℓ^2. This can be understood for a line limited by RC effects, since the resistance $R \propto \ell/A$ while the capacitance $C \propto \ell$, so that the time constant $RC \propto \ell^2/A$. A similar argument applies to lines limited by LC effects. The bandwidth is thus determined by the aspect ratio $\ell/\sqrt{A}$ and cannot be altered by miniaturizing the device or by increasing its size. Optical interconnects do not suffer from

this *aspect ratio limitation* since the bandwidth is governed by other physical effects and is generally larger. Moreover, for optical interconnects, the maximum bandwidth of the data carried by each connection is not affected by the density of proximate interconnects. Said differently, the optical crosstalk among neighboring lines is not influenced by an increase in the data rate. This stems from the small ratio of the bandwidth to the carrier frequency of the modulated light. This is not the case for electronic interconnects, in which the density must be reduced sharply at high modulation frequencies to eliminate capacitive and inductive coupling among proximate interconnects. Optical interconnects therefore offer greater density–bandwidth products than do electronic interconnects.

- *Delay.* Photons travel at a speed $c_o = 0.3$ mm/ps in free space and at $c_o/n \approx 0.086$ mm/ps in silicon ($n = 3.5$). The corresponding propagation time delays are therefore $\approx$ 3.3 ps/mm and ≈ 11.7 ps/mm, respectively. Electrical-signal delays on strip lines fabricated from ceramics and polyimides are approximately 10.2 and 6.8 ps/mm, respectively. Propagation delay is therefore not an issue *per se*. However, the velocity of light is independent of the number of interconnections branching from an optical interconnect, whereas in electronic transmission lines the velocity is inversely proportional to the capacitance per unit length and therefore depends on the total capacitive "load"; hence the propagation delay time increases as fan-outs increase. Optics thus offers greater flexibility with respect to fan-out and fan-in interconnections, which are limited only by the available optical power.
- *Density.* The most dense set of interference-free interconnects makes use of unguided beams, each with a small width and a small divergence angle, limited only by diffraction (the product of the width and the angle of a narrow beam is of the order of a wavelength, which is small at optical frequencies). Moreover, since such beams can intersect (pass through one another) without mutual interference (assuming that the medium is linear), they can be used in three-dimensional configurations to create interconnects with densities unmatchable by electrical wires. Light may also be guided in planar or quasi-planar low-loss *dielectric* waveguides, with widths as small as a wavelength, that can be packed densely with minimal crosstalk. Electrical interconnects, on the other hand, use *metallic* conductors, such as strip lines, that serve as transmission lines or waveguides for the electromagnetic waves associated with the oscillating electric charges. Metallic conductors introduce losses and cannot be packed as tightly since they become susceptible to electromagnetic interference when in close proximity.
- *Power.* To avoid reflections, electrical interconnects must be terminated with a matched impedance, which generally requires an increase in the expenditure of power. For optical interconnects, on the other hand, reflections can be significantly reduced by making use of antireflection coatings. Optical power requirements are typically limited by photodetector sensitivities, the efficiencies of electrical-to-optical and optical-to-electrical conversion processes, and the power transmission efficiencies of the interconnect elements.

24.2 PASSIVE OPTICAL ROUTERS

A passive optical router redirects the data carried by a set of incoming optical beams to one or more of a set of outgoing optical beams, on the basis of the location of the beam and a physical attribute X associated with the data (e.g., wavelength, polarization, intensity, phase, or arrival time). A single incoming or outgoing beam may carry several components marked by different values of X. The system avoids contention, i.e., routing different data marked by the same attribute value to the same outgoing

beam. Three common routers are the demultiplexer, the multiplexer, and the add–drop multiplexer, as described below and illustrated in Fig. 24.2-1.

- A **demultiplexer** (**DMUX** or **DEMUX**) is a $1 \times N$ router that sorts the components with attribute values $X_1, X_2, \ldots, X_N$ in a single input beam and directs them to separate output ports, as shown in Fig. 24.2-1(a). The DMUX may be implemented by use of a **broadcast-and-select** operation: a $1 \times N$ fan-out interconnect broadcasts copies of the incoming beam to all output ports and is followed by a bank of filters that select components with the desired attribute values and reject all others.
- The **multiplexer** (**MUX**) is the inverse of the DMUX. As illustrated in Fig. 24.2-1(b), input beams with distinct values of the optical attributes $X_1, X_2, \ldots, X_N$ are combined into a single beam that can be subsequently separated by use of a demultiplexer. Multiplexing and demultiplexing based on wavelength, frequency, and time are used extensively in optical communication systems.
- The **add–drop multiplexer** (**ADM**), portrayed in Fig. 24.2-1(c), is another important routing device used in communication networks. Here, a demultiplexer sorts components with different attribute values, separates the component of a selected attribute value, say X_2, drops its data content and instead adds new data, and then subsequently combines all components into a single beam by use of a multiplexer. The optical ADM (**OADM**) is commonly used in optical networks.

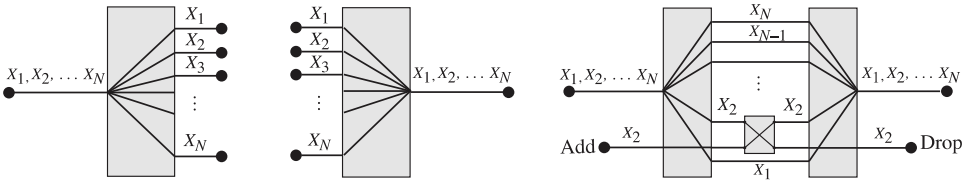


Figure 24.2-1 Attribute-based routers. (a) Demultiplexer (DMUX). (b) Multiplexer (MUX). (c) Add–drop multiplexer (ADM).

A. Wavelength-Based Routers

Wavelength-based routers are commonly used in wavelength-division multiplexing (WDM) optical fiber communication systems and networks. As described in Sec. 25.3C, these systems incorporate channels with multiple wavelengths in the same optical fiber. They employ routers that combine the channels at the fiber input and separate them at the fiber output using wavelength-based routers called **wavelength-division multiplexers** and **wavelength-division demultiplexers**, respectively.

Implementations of Wavelength-Division Multiplexers/Demultiplexers

A number of techniques, some of which are illustrated in Fig. 24.2-2, can be used for wavelength-division demultiplexing.

- An angularly dispersive optical device will separate the components of different wavelengths within a single optical beam into separate optical beams. The simplest devices that exhibit angular dispersion are the prism [Fig. 24.2-2(a)] and the diffraction grating [Fig. 24.2-2(b)]. The angular dispersion of a prism is limited by the rate of change of the refractive index with respect to the wavelength, $dn/d\lambda$, which is usually not sufficiently large to adequately separate components of slightly different wavelengths. Prisms made of photonic-crystal materials

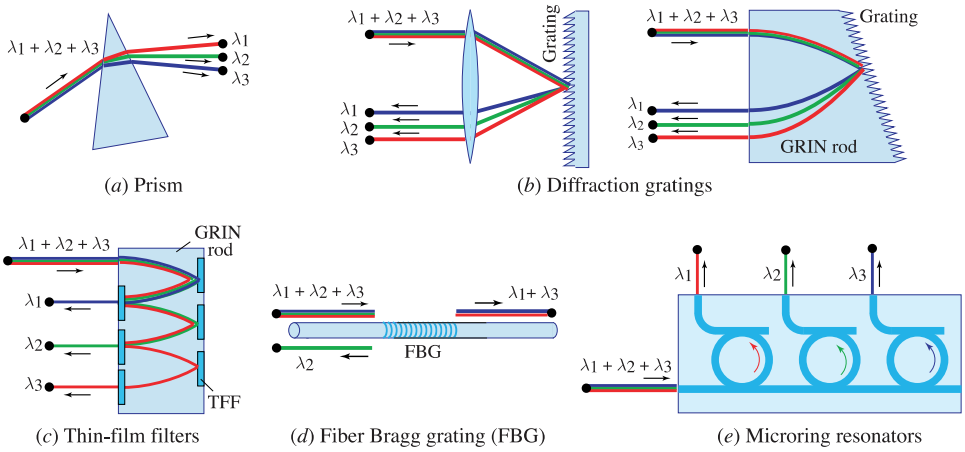


Figure 24.2-2 Wavelength-division demultiplexers. (*a*) Prism. (*b*) Diffraction grating with a lens or graded-index (GRIN) rod. (*c*) Dielectric thin-film interference filters (TFFs). (*d*) Fiber Bragg gratings (FBGs). (*e*) Microring resonator filters.

(Chapter 7), called **superprisms**, can exhibit dispersive power that is two to three orders of magnitude greater than that of conventional materials. The angular dispersion of diffraction gratings (Sec. 2.4B) is stronger than that of ordinary prisms; they are capable of resolving wavelength differences corresponding to a few GHz.

- Wavelength separation may also be implemented by using a bank of filters tuned to the different wavelengths. The incoming light is broadcast to the different filters; each filter transmits a single wavelength channel and blocks all others. Alternatively, the beam may be directed through a sequence of filters with narrow spectral passbands, such as dielectric **thin-film interference filters** (TFFs), each of which transmits one wavelength and reflects all others to the next filter, as illustrated in Fig. 24.2-2(*c*). A GRIN rod may be used to guide the rays between the filters.
- In a similar implementation, the wavelength dependence of the reflectance of a fiber Bragg grating (FBG) (Sec. 7.1C) is exploited to separate wavelength components; the component at the Bragg wavelength $\lambda_B = 2\Lambda$, where Λ is the grating period, is reflected and all other components are transmitted. Multiple Bragg gratings are used to separate multiple wavelengths [Fig. 24.2-2(*d*)].
- In yet another similar implementation, a sequence of **microring-resonator filters**, each tuned to one wavelength, is used [Fig. 24.2-2(*e*)].
- Other implementations make use interferometers such as the **Mach–Zehnder interferometer** and the **arrayed waveguides router**, as will be described subsequently.

Optical Add–Drop Multiplexer (OADM)

An optical add–drop multiplexer (OADM) drops data from, and simultaneously adds data to, selected wavelength channels of a multi-channel optical beam. The individual wavelength channels may be accessed by means of a demultiplexer followed by a multiplexer, as portrayed in Fig. 24.2-1(*c*). A selected wavelength channel, along with its associated data, is separated from the other channels by means of a wavelength-sensitive optical element and extracted (dropped) by detecting (annihilating) it. New data are added via a modulated optical source. We consider two examples of OADMs

based on this arrangement. The first, pictured in Figs. 24.2-3, makes use of a fiber Bragg grating (FBG) as the wavelength-selective element while the second, illustrated in Fig. 24.2-4, uses multiple microring resonators for this purpose.

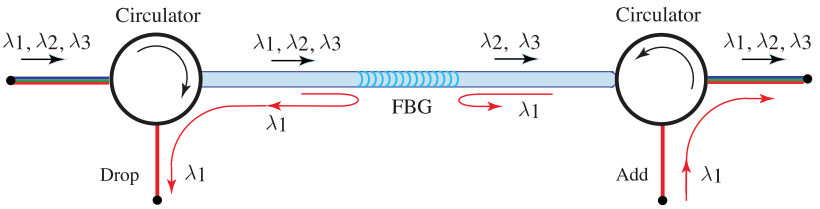


Figure 24.2-3 An optical add–drop multiplexer (OADM). This version makes use of a fiber Bragg grating (FBG) to reflect the wavelength component λ_1, which is then directed by a circulator to a photodetector for annihilation. The remaining components, λ_2 and λ_3, pass through the FBG to the output. Another circulator serves to add light, modulated by new data, at λ_1. The FBG retroreflects light at λ_1.

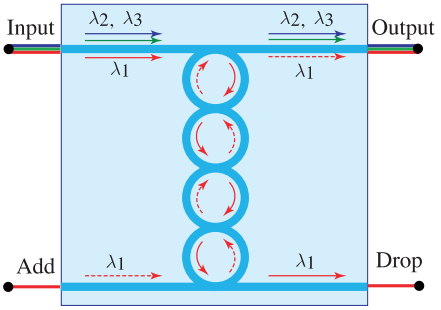


Figure 24.2-4 Another configuration for an OADM. This version makes use of multiple microring resonators to extract the wavelength component λ_1 from the multichannel input beam and to direct it to a photodetector for annihilation. The remaining components, λ_2 and λ_3, pass through to the output. Light at λ_1, modulated by new data, is selected by the resonators and transferred to the output beam. Multiple microring resonators offer greater wavelength selectivity (narrower spectral width and greater rejection ratio) than single microring resonators.

The Mach–Zehnder Interferometer as a Demultiplexer

Since interferometers are sensitive to wavelength, they are suitable for wavelength-division routing. The integrated-photonic Mach–Zehnder interferometer (**MZI**) portrayed in Fig. 24.2-5, for example, may be used as a two-wavelength demultiplexer. To direct the components of wavelengths λ_1 and λ_2 in the input beam to different output ports, the pathlength difference Δd is selected such that the phase difference $\varphi = 2\pi\Delta d/\lambda$ is an even multiple of π at λ_1 and an odd multiple of π at λ_2; i.e., $\Delta d = q_1\lambda_1/2$ and $\Delta d = q_2\lambda_2/2$, where q_1 is an even integer and q_2 is an odd integer.

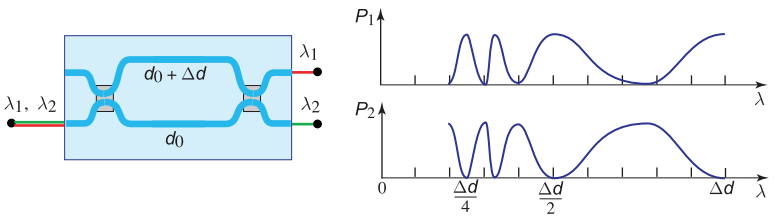


Figure 24.2-5 Wavelength-division routing (demultiplexing) of two wavelengths by use of an integrated-photonic Mach–Zehnder interferometer.

The resolution of the routing device, i.e., the closest wavelengths that can be separated, is determined by writing $|1/\lambda_1 - 1/\lambda_2| = |q_1 - q_2|/2\Delta d$, and taking $|q_1 - q_2| = 1$, so that $|1/\lambda_1 - 1/\lambda_2| = 1/2\Delta d$. The corresponding frequency difference $\Delta\nu = |\nu_1 - \nu_2|$ is therefore

$$\Delta\nu = \frac{c}{2\Delta d}. \tag{24.2-1}$$

For example, if $\Delta d = 1$ mm and $n = 1.5$, then $\Delta\nu = 100$ GHz. Smaller separations $\Delta\nu$ require proportionally longer pathlength differences Δd.

The spectral sensitivity of the MZI router may be determined by writing its interconnection matrix as

$$\mathbf{T} = \begin{bmatrix} 1 & j \\ j & 1 \end{bmatrix} \begin{bmatrix} \exp[-j2\pi(d_0 + \Delta d)/\lambda] & 0 \\ 0 & \exp(-j2\pi d_0/\lambda) \end{bmatrix} \begin{bmatrix} 1 & j \\ j & 1 \end{bmatrix}, \tag{24.2-2}$$

where $d_0 + \Delta d$ and d_0 represent the pathlengths of the interferometer branches. The first and third entries in this matrix product are the interconnection matrices for a 3-dB coupler set forth in (24.1-2). For an input field of unit power at input port 2, the powers received at output ports 1 and 2 are $P_1 = |T_{21}|^2$ and $P_2 = |T_{22}|^2$, respectively, so that

$$P_1 = \cos^2(\pi\Delta d/\lambda) \quad \text{and} \quad P_2 = \sin^2(\pi\Delta d/\lambda). \tag{24.2-3}$$

These powers are plotted in Fig. 24.2-5 as functions of λ. It is clear from this dependence that the smaller the ratio $\lambda/\Delta d$, the more rapidly these functions alternate between 0 and 1 and thus the greater the possibility for demultiplexing closely spaced wavelengths.

Multiple MZIs may be cascaded to separate more than two wavelengths. For example, four wavelengths may be separated in a two-step process, as illustrated in Fig. 24.2-6. The first MZI separates the odd-numbered from the even-numbered wavelengths, and subsequent MZIs implement finer wavelength separations.

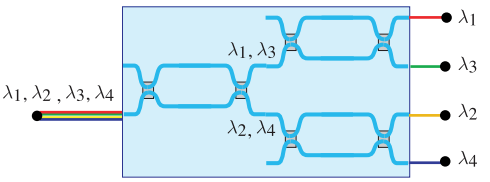


Figure 24.2-6 Wavelength-division routing (demultiplexing) of four wavelengths by use of cascaded integrated-photonic Mach–Zehnder interferometers.

Arrayed Waveguides (AWG) Routers

Other interferometric configurations may be used to provide greater wavelength selectivity. Multipath interferometers, for example, are highly selective to wavelength since they exhibit sharp resonances. Such interferometers may be custom designed using planar waveguides and may be configured to provide routing for a large number of wavelengths in devices that contain many input and output ports. The **arrayed waveguides (AWG) router** operates on the principle that *each* connection between an input port and output port is configured as an independent multipath interferometer that transmits only specific wavelengths. Since this device is similar to a diffraction-grating spectrometer, the AWG router is also known as a **waveguide grating router** (**WGR**).

Multipath interferometer. Before embarking on a discussion of the operation of the AWG, we first review the properties of the multipath interferometer. An L-path interferometer is a connection comprising L optical paths whose lengths increase progressively and linearly so that adjacent paths have exactly the same pathlength difference Δd, as portrayed in Fig. 24.2-7. The wave received at the output port is then the sum of L waves of equal amplitudes and equal phase difference, as considered in Sec. 2.5B. Since the phase difference between adjacent paths is $\varphi = 2\pi\Delta d/\lambda$ at wavelength λ, the power transmittance provided in (2.5-12) becomes

$$T = \frac{\sin^2(L\varphi/2)}{\sin^2(\varphi/2)} = \frac{\sin^2(L\pi\Delta d/\lambda)}{\sin^2(\pi\Delta d/\lambda)}, \tag{24.2-4}$$

which is a periodic function of φ with sharp peaks that occur when φ equals integer multiples of 2π, as shown in Fig. 2.5-7. The dependence of T on λ is not periodic, but rather comprises sharp peaks at $\lambda = \Delta d$ and integer fractions thereof, as illustrated in Fig. 24.2-7. The larger the number of paths L, the sharper the peaks.

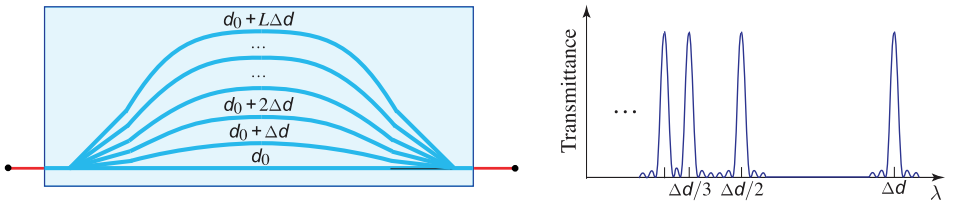


Figure 24.2-7 A multipath interferometer and the wavelength dependence of its transmittance.

The AWG as a wavelength-division demultiplexer. An arrayed waveguides (AWG) router may be used as a $1 \times N$ wavelength-based router that directs each of N wavelength components, $\lambda_1, \lambda_2, \ldots, \lambda_N$, at the input port to one of the N output ports, as shown in Fig. 24.2-8. There are N multipath interferometers, one for each of the output ports. Each interferometer has a unique pathlength difference Δd selected such that only a specific wavelength is transmitted. This is achieved if the connections leading to the mth output port are designed to have a pathlength difference Δd_m that is an integer multiple of λ_m, but not an integer multiple of the other wavelengths.

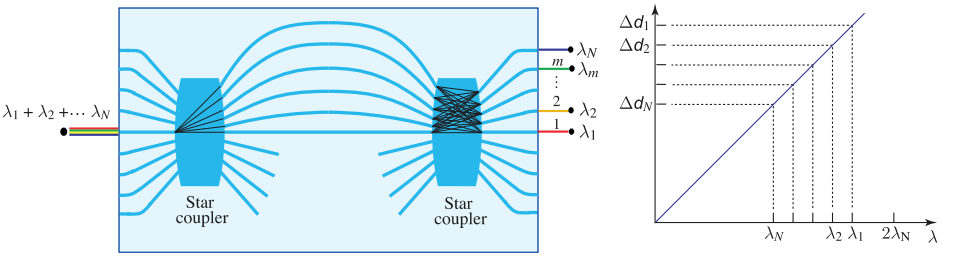


Figure 24.2-8 Wavelength-division demultiplexing using an arrayed waveguides (AWG) router.

The design is simpler if the wavelengths $\lambda_1, \lambda_2, \ldots, \lambda_N$ are distributed uniformly as a decreasing sequence, $\lambda_m = \lambda_0 - m\Delta\lambda$, where $\Delta\lambda$ is the wavelength channel

separation and $\lambda_0 = \lambda_1 - \Delta\lambda$. A necessary condition for operation of the demultiplexer is

$$\Delta d_m = \lambda_m = \lambda_0 - m\Delta\lambda, \quad m = 1, 2, \ldots, N, \tag{24.2-5}$$

indicating that the pathlength difference for the connections to the mth output port decreases linearly with m. The other condition is that Δd_m is not equal to an integer multiple of λ_ℓ for all $\ell \neq m$. This condition is automatically satisfied if the shortest wavelength λ_N is greater than one half of the longest wavelength λ_1, as depicted in Fig. 24.2-8.

In the implementation displayed in Fig. 24.2-8, each pathlength between the input port and an output port is the sum of the waveguide length and the distances traveled in the star couplers. The waveguide lengths may be selected to increase progressively by a fixed length Δd_w. For a star coupler with circular boundaries, the pathlength difference may be approximated by a linearly decreasing function of m, so that

$$\Delta d_m = \Delta d_w + (\Delta d_a - m\,\Delta d_b), \tag{24.2-6}$$

where Δd_a and Δd_b are constants that depend on the geometry of the couplers. The condition provided in (24.2-5) can therefore be satisfied if $\Delta d_w + \Delta d_a = \lambda_0$ and $\Delta d_b = \Delta\lambda$. The resolution of the wavelength demultiplexer, i.e., the minimum wavelength separation $\Delta\lambda$, is therefore limited by the minimum value of the geometrical factor Δd_b.

The AWG as an *N* x *N* wavelength router. The AWG may also be used as a more general $N \times N$ wavelength router. The connections between the ℓth input port and the mth output port form a multipath interferometer with pathlength difference $\Delta d_{\ell m} = \lambda_{00} - (\ell + m)\Delta\lambda$, which decreases linearly with both ℓ and m (λ_{00} and $\Delta\lambda$ are constants that depend on the geometry of the AWG). Light is transmitted between these ports if the wavelength $\lambda_{\ell m}$ equals $\Delta d_{\ell m}$, i.e., if

$$\boxed{\lambda_{\ell m} = \lambda_{00} - (\ell + m)\Delta\lambda, \quad \ell, m = 1, 2, \ldots, N.} \tag{24.2-7}$$

AWG Equation

Equation (24.2-7) is a generalization of (24.2-5). Although the AWG does not implement an arbitrary wavelength routing, it can offer solutions to certain routing problems such as simultaneous wavelength multiplexing operations.

B. Polarization- , Phase- , and Intensity-Based Routers

Polarization-Based Routing

A simple example of passive optical routing is based on polarization. In polarization-division demultiplexing the parallel and orthogonal polarization components of an optical beam are separated by making use of a polarizing beamsplitter (PBS), as illustrated in Fig. 24.2-9. Polarization-based multiplexing is achieved by using the PBS as a beam combiner, with light traveling from right-to-left instead of from left-to-right.

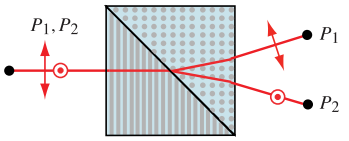


Figure 24.2-9 Polarization-division routing using a polarizing beamsplitter (PBS). For beams traveling from left-to-right, the prism is a demultiplexer. For beams traveling from right-to-left, it is a multiplexer.

Phase-Based Routing

Another simple example of passive optical routing is based on phase. Here a sequence of optical pulses with phases 0 or π are to be sorted based on phase and routed to two output ports. This may be accomplished by making use of a simple Mach–Zehnder interferometer, as shown in Fig. 24.2-10.

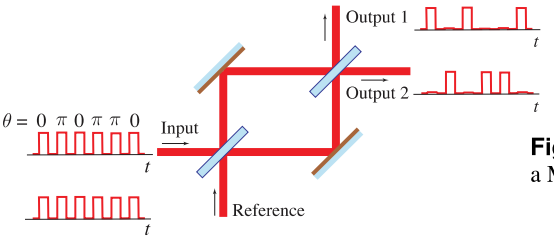


Figure 24.2-10 Phase-based routing using a Mach–Zehnder interferometer.

Intensity-Based Routing

A light beam whose intensity takes on different values at different times may be routed into separate beams based on the value of the intensity. For example, a light beam carrying a sequence of pulses with two intensities, as depicted in Fig. 24.2-11, may be separated into two beams, one containing the high-intensity pulses and the other containing the low-intensity pulses. This demultiplexing operation requires the use of a nonlinear optical element. It is often implemented by converting the intensity variation into phase change via the optical Kerr effect (Sec. 22.3A), as described next.

Nonlinear Mach–Zehnder interferometer (MZI). The nonlinear MZI is a conventional MZI in which a nonlinear optical element, such as a Kerr cell, is placed in one of the interferometer branches. The cell introduces a phase shift proportional to the light intensity. The system is adjusted such that the phase difference between the interferometer branches is an odd multiple of π for one intensity, and an even multiple of π for the other. This diverts the stream of pulses into two output ports, one containing the high-intensity pulses and the other containing the low-intensity pulses, as illustrated in Fig. 24.2-11(a). The interferometer may also be implemented using optical fibers, as depicted in Fig. 24.2-11(b).

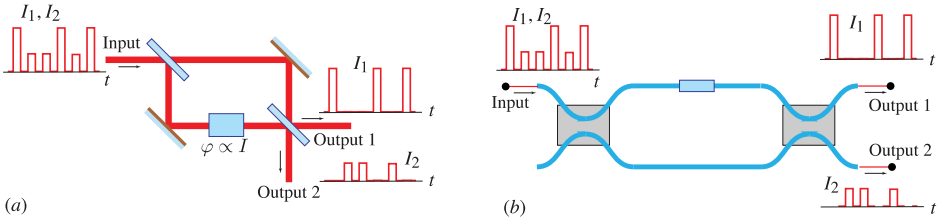


Figure 24.2-11 An intensity-based 1×2 router using a Mach–Zehnder interferometer in which a nonlinear Kerr medium has been placed in one of the interferometer branches. (a) Bulk-optics version. (b) Fiber-optics version.

Nonlinear asymmetric Sagnac interferometer. An intensity-based 1×2 router using a nonlinear fiber Sagnac interferometer is illustrated in Fig. 24.2-12. In this configuration, light enters from fiber 1 and is split into a clockwise wave and a counterclockwise wave. If the optical pathlengths of these waves are identical, constructive interference occurs and the light propagates back into fiber 1 and is directed to output port 1, so that the device acts as a mirror. This occurs if the fiber is linear, or if the fiber is nonlinear and the intensities of the two waves are equal. However, if the coupler feeding the interferometer loop is not symmetric, then the intensities in the two paths are unequal, in which case the phase shifts introduced via the optical Kerr effect are

generally different. When the phase difference is π, destructive interference ensues and light is diverted into fiber 2 and output port 2. Since the phase difference is proportional to the intensity of the incident wave, the system acts as a 1×2 self-controlled intensity-division router (a demultiplexer).

Asymmetry between the clockwise and counterclockwise waves in the Sagnac interferometer may also be introduced by placing an erbium-doped fiber amplifier (EDFA) at an asymmetric location within the loop. This serves to amplify one of the interfering waves during the first half of its trip around the loop, so that it travels more than half a round trip at a high intensity. The other wave is amplified during the second half of its trip around the loop so it travels a shorter distance at high intensity and thus encounters a smaller nonlinear phase shift. This system is known as a **nonlinear optical loop mirror** (**NOLM**).

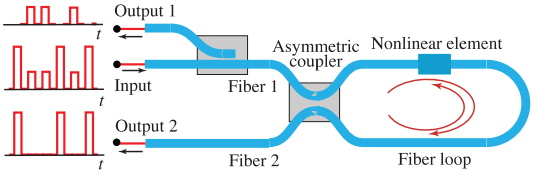


Figure 24.2-12 Intensity-based 1×2 router using a nonlinear Sagnac interferometer that serves as a nonlinear optical loop mirror (NOLM).

Nonlinear directional coupler (NLDC). A waveguide or fiber-optic directional coupler made of a Kerr material can also serve as an intensity-based router, as illustrated in Fig. 24.2-13. If the intensity of the input pulse is low, the medium is approximately linear and the light is periodically coupled from one guide to the other as it travels (Fig. 9.4-6). If the length of the coupler is equal to the transfer distance L_0 and there is no phase mismatch, the light is completely transferred from the input waveguide to the other waveguide. However, for pulses of large intensity the propagation constants are altered by the Kerr effect, creating an intensity-dependent phase mismatch that varies with distance. Propagation then obeys the nonlinear coupled differential equations

$$\frac{d\mathrm{a}_1}{dz} = -j\mathcal{C}\exp(j\Delta\beta z)\mathrm{a}_2(z) - j\gamma|\mathrm{a}_1|^2\mathrm{a}_1 \tag{24.2-8}$$

$$\frac{d\mathrm{a}_2}{dz} = -j\mathcal{C}\exp(-j\Delta\beta z)\mathrm{a}_1(z) - j\gamma|\mathrm{a}_2|^2\mathrm{a}_2, \tag{24.2-9}$$

which are generalizations of the linear coupled equations (9.4-4) for the conventional directional coupler. Here, $\mathcal{C} = \pi/2L_0$ is the coupling coefficient and γ is proportional to the optical Kerr coefficient n_2 [Sec. 22.3A and (23.5-16)]. The system is designed so that the high-intensity pulses exit the coupler from the same waveguide and are separated from the low-intensity pulses, as displayed in Fig. 24.2-13.

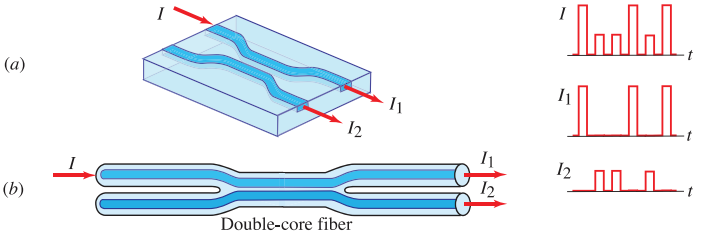


Figure 24.2-13 Intensity-based 1×2 router using a directional coupler fabricated from a nonlinear optical material. Implementation in: (*a*) integrated-photonic technology; (*b*) fiber-optic technology.

Soliton directional coupler. Since the intensity of an optical pulse varies during its time course, so too does the nonlinear refractive index n_2 and the corresponding propagation constant in the nonlinear medium. Different fractions of the pulse power are therefore transferred between the two channels, which can lead to pulse reshaping and possibly to pulse breakup. However, this will not occur in a fiber-optic NLDC [Fig. 24.2-13(b)] if the pulse is an optical soliton (Sec. 23.5B). Because the nonlinear phase shift of an optical soliton is constant over the pulse's envelope, the soliton pulse remains intact as it is routed between the coupled fibers. A further advantage of operating the NLDC in the soliton mode is that the transition between the output ports is a far sharper function of the input pulse power.

24.3 PHOTONIC SWITCHES

A. Space-Switch Architectures

A switch is a device that both establishes and releases connections among transmission paths in a communication or signal-processing system. A control unit processes the commands for connections and sends a control signal to operate the switch in the desired manner. Whereas interconnects always operate on the incoming signals in the same manner, switches are controllable, active, or reconfigurable interconnects that can be modified by an external command. **Space switches** establish transmission paths that route optical beams between specific physical locations, namely the input and output ports of the switch. Examples of such switches are displayed in Fig. 24.3-1.

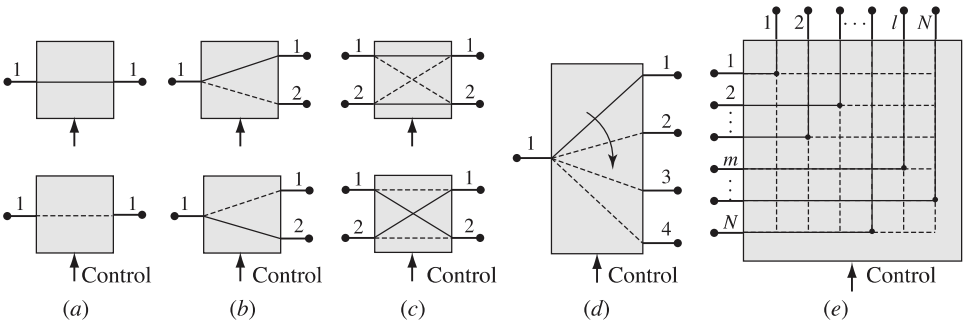


Figure 24.3-1 Examples of space switches. (a) A 1×1 switch connects or disconnects two lines. It is an ON–OFF switch. (b) A 1×2 switch connects one line to either of two lines. (c) A 2×2 crossbar switch connects two lines to two lines. It has two configurations: the bar state and the cross state, and may be regarded as a controllable directional coupler. (d) A $1 \times N$ switch connects one line to one of N lines. (e) An $N \times N$ crossbar switch connects N lines to N lines. Any input line can always be connected to a free (unconnected) output line without blocking (i.e., without conflict).

A 1×1 switch can be used as an elementary unit from which switches of larger sizes can be built. An $N \times N$ crosspoint-matrix (crossbar) switch, for example, may be constructed by using an array of N^2 1×1 switches, organized at the points of an $N \times N$ matrix, to connect or disconnect each of the N input lines to a free output line [Fig. 24.3-1(e)]. In this portrayal, the mth input reaches all elementary switches in the mth row, while the lth output is connected to the outputs of all elementary switches in the lth column. A connection is effected between the mth input and the lth output by activating the (m, l) 1×1 switch. Examples that make use of 1×1 switches are shown in Fig. 24.3-2.

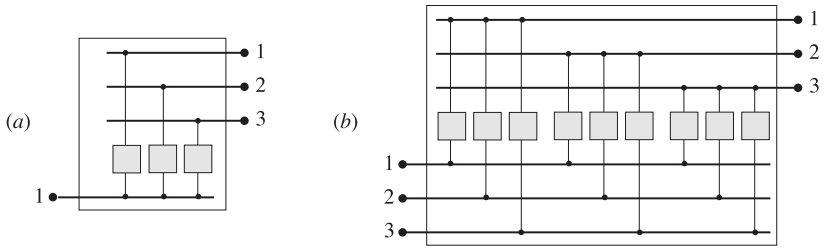


Figure 24.3-2 (a) A 1×3 switch made from three 1×1 switches. (b) A 3×3 switch made from nine 1×1 switches in a broadcast-and-select configuration.

An $N \times N$ switch can also be constructed by using 2×2 switches as building blocks. Examples that make use of 2×2 switches are displayed in Fig. 24.3-3.

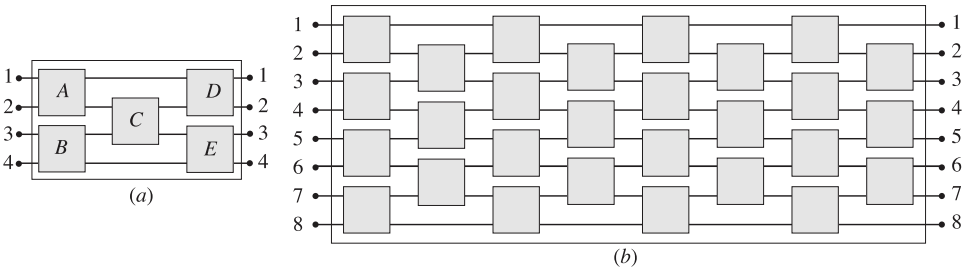


Figure 24.3-3 (a) A 4×4 switch made from five 2×2 switches. Input line 1 is connected to output line 3, for example, if switches A and C are in the cross state and switch E is in the bar state. (b) An 8×8 switch comprising 28 interconnected 2×2 switches.

Photonic Switch Characteristics

A photonic switch is characterized by the following parameters:

- *Size*: number of input and output lines.
- *Direction(s)*: whether data can be transferred in one or two directions.
- *Switching time*: time required for the switch to be reconfigured.
- *Propagation delay time*: time required by the signal to cross the switch.
- *Throughput*: maximum data rate that can flow through the switch.
- *Switching energy*: energy required to activate and deactivate the switch.
- *Power dissipation*: energy dissipated per second in the process of switching.
- *Insertion loss*: decrease in signal power introduced by forging the connection.
- *Extinction ratio*: contrast between the ON and OFF states.
- *Crosstalk*: undesired power leakage to other lines.
- *Blocking probability*: probability that a connection cannot be established because of a conflict with another connection.
- *Physical dimensions*: the physical size is an important consideration when large arrays of switches are built.

B. Implementations of Photonic Space Switches

Optoelectronic Switches

Electronic switches have evolved a great deal since the early years of telephony, generally tracking the steady advances in microelectronics. Nanoscale CMOS electronic

gates can operate at switching times as small as 0.1 ns and with switching energies smaller than 1 fJ. Advanced MOSFET gates can be switched at subpicosecond time scales. Electronic chips for crossbar switching with large numbers of ports (e.g., 128×128) are readily available. It is therefore natural to consider these devices for photonic switching. Unfortunately, optoelectronic switches are cumbersome for photonic switching inasmuch as they require optical-to-electrical conversion at the input of the switch and electrical-to-optical conversion at its output, as illustrated schematically in Fig. 24.3-4. Moreover, the optical/electrical/optical conversions required for the operation of optoelectronic switches introduce substantial time delays and power loss. It is therefore desirable to make use of "transparent" photonic switches that operate directly on the optical signals.

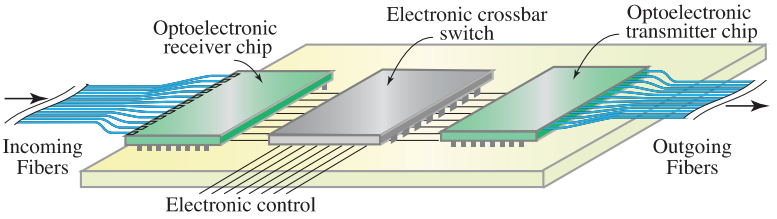


Figure 24.3-4 An optoelectronic crossbar switch. Incoming optical signals carried by optical fibers are detected by an array of photodetectors on an optoelectronic chip, switched using an electronic crossbar switch, and regenerated using an array of light sources (e.g., VCSELs) that feed outgoing optical fibers. The process is cumbersome and inefficient.

Elementary Photonic-Switch Configurations

The most elementary of photonic switches are the optical scanner and the modulator. A scanner that deflects an optical beam into one of N possible directions is a $1 \times N$ switch [Fig. 24.3-5(a)]. An optical modulator operated in ON–OFF mode also serves as a 1×1 switch. Modulation may be *direct*, relying on some physical effect that transmits or blocks the light, or *interferometric*, using for example an optical phase modulator placed in one arm of a Mach–Zehnder interferometer, which converts phase modulation into intensity modulation [Fig. 24.3-5(b)]. Another elementary photonic switch is a directional coupler operated as a 2×2 switch. This may be implemented by making use of a Mach–Zehnder interferometer in which a phase modulator is placed in one or both branches [Fig. 24.3-5(c)]. The two interferometer branches may also represent two orthogonal polarization components, in which case the phase modulator is a wave retarder that introduces a relative phase shift between the two polarizations.

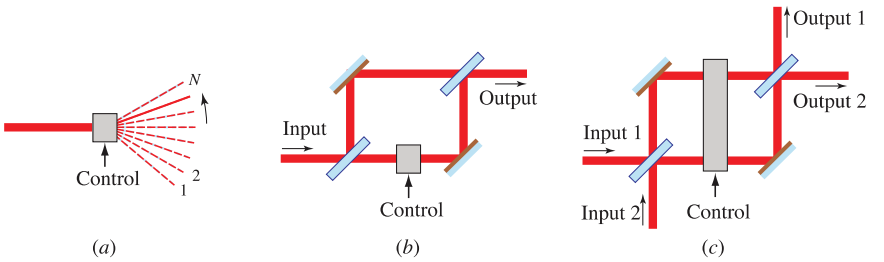


Figure 24.3-5 (a) An optical scanner as a $1 \times N$ switch. (b) An interferometer with a phase modulator as a 1×1 switch. (c) An interferometer with a phase modulator as a 2×2 switch.

Elementary photonic switches may be combined or cascaded in free space or in planar-waveguide technology to create switches of higher dimension. As illustrated in

Fig. 24.3-6, for example, a planar array of 16 optical modulators, each serving as a 1×1 switch, may be configured in an optical system that operates as a 4×4 crossbar switch in the broadcast-and-select configuration.

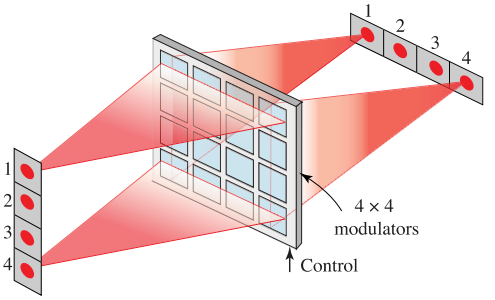
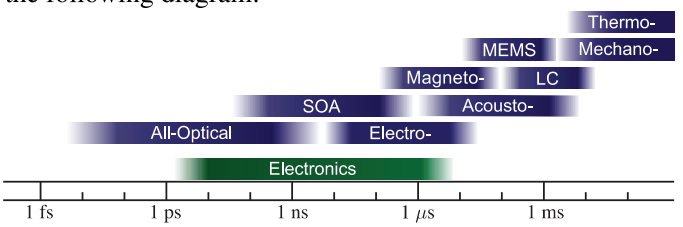


Figure 24.3-6 A 4×4 crossbar switch. Each of the 16 elements is a 1×1 switch that transmits or blocks light depending on a control signal. Light from the mth point of the input, $m = 1, 2, 3, 4$, is broadcast to all switches in the mth row. Light from all switches in the lth column is directed to the lth output point, $l = 1, 2, 3, 4$. The system is an implementation of the 4×4 switch depicted in Fig. 24.3-1(e).

The modulation and deflection of light can be achieved by the use of mechanical; electromechanical; electrical; acoustical; magnetic; and thermal control. Switches that operate under these rubrics are, respectively, mechano-optic (or optomechanical); microelectromechanical systems (MEMS); electro-optic, semiconductor, and liquid-crystal; acousto-optic, magneto-optic, and thermo-optic. The remainder of this section is devoted to providing brief outlines of these technologies. All-optical, or opto-optic, switches are described in Sec. 24.3C. The switching times of these various devices are compared in the following diagram:



Mechano-Optic Switches

A mechano-optic (or optomechanical) $1 \times N$ switch (a scanner) may be implemented by using a moving (rotating or alternating) mirror, prism, or holographic grating that deflects a light beam to a set of directions (Fig. 24.3-7). As illustrated in Fig. 24.3-7(c), an optical fiber can be connected to any of a number of other optical fibers by mechanically moving the input fiber to align with the selected output fiber. Piezoelectric elements may be used for faster mechanical action.

Microelectromechanical systems (**MEMS**) are miniaturized mechanical arrangements powered by electrostatic actuators and fabricated in large arrays using processes similar to those used in microelectronics. Switching times range from 10 μs to 10 ms. A crossbar switch, for example, may be implemented by using a set of MEMS popup mirrors, as shown in Fig. 24.3-8(a), or by using a set of rotating mirrors, as shown in Fig. 24.3-8(b). Another example of a MEMS switch is the digital micromirror device (DMD) developed at Texas Instruments (Fig. 24.3-9). As shown in the inset, it is an array of micromirrors, each assuming one of two possible orientations ($+12^\circ$ or -12°). In one orientation, the mirror deflects incoming light through an imaging lens that displays a bright spot in the image plane, whereas in the other orientation the mirror deflects light to a beam dump so that dark spot is displayed in the image plane. The DMD is a binary (bright/dark) spatial light modulator (SLM) that finds use in commercial projection systems. A gray value at any spot may be obtained by modulating the relative time durations of dark and bright.

Mechano-optic switches generally offer low insertion loss and low crosstalk but have relatively slow response times.

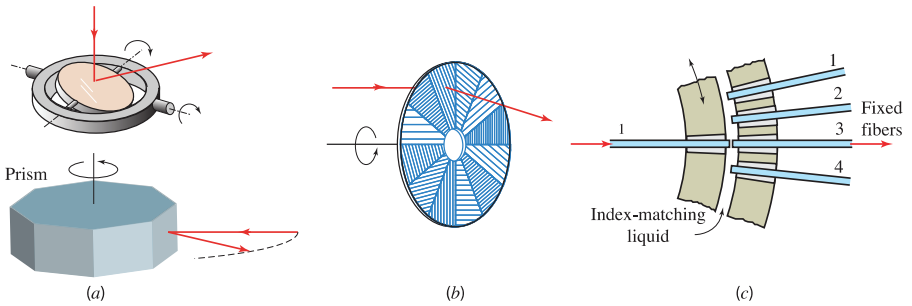


Figure 24.3-7 Examples of the deflection of light into different directions using mechano-optic switches. (a) A rotating mirror or prism. (b) A rotating holographic disk; each sector of the disk contains a grating whose orientation and period determine the scanning plane and scanning angle of the deflected light. (c) An optical fiber attached to a rotating wheel aligned with one of a number of optical fibers attached to a fixed wheel; the fibers are placed in V-grooves and an index-matching liquid is used to ensure good optical coupling.

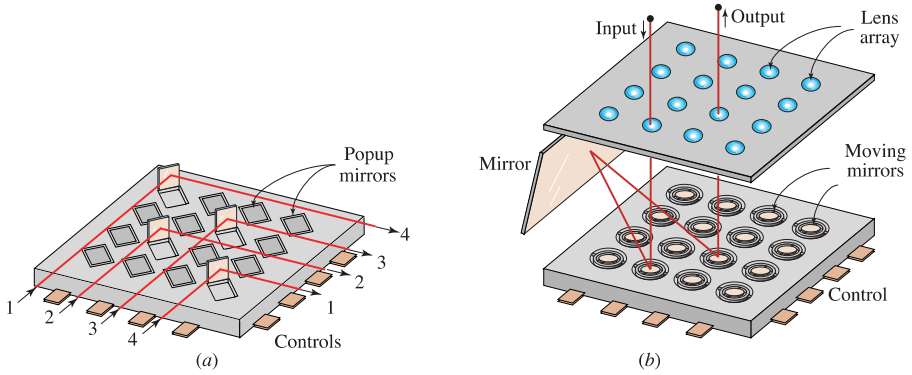


Figure 24.3-8 (a) MEMS popup-mirror switch. (b) MEMS rotating-mirror switch.

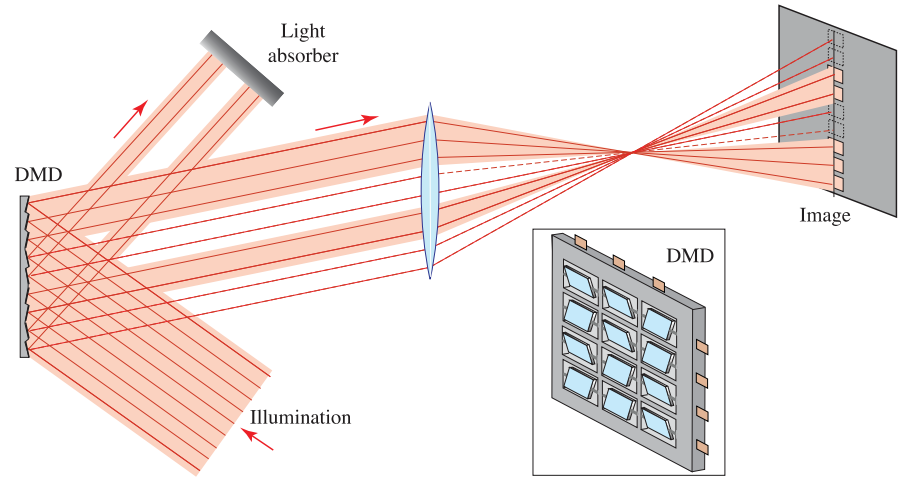


Figure 24.3-9 The digital micromirror device (DMD) is an array of micromirrors switched between two orientations to create a binary image.

Electro-Optic Switches

As discussed in Sec. 21.1, the refractive indices of electro-optic materials are altered in the presence of an electric field. These materials may therefore be used as electrically controlled phase modulators or wave retarders. When placed in one arm of an interferometer, or between crossed polarizers, an electro-optic cell can serve as an electrically controlled light modulator or a 1×1 (ON–OFF) switch (Sec. 21.1B).

Since it is difficult to make large arrays of switches using bulk crystals, the most viable approach to electro-optic switching is via integrated photonics. As described in Sec. 9.3, integrated-photonic waveguides may be fabricated using electro-optic dielectric substrates such as $LiNbO_3$; the diffusion of titanium into the substrate creates strips of slightly elevated refractive index that serve as waveguides. An example of a 1×1 switch using an integrated-photonic Mach–Zehnder interferometer (MZI) is displayed in Fig. 24.3-10(a) (which reproduces Fig. 21.1-5). A 2×2 integrated-photonic switch can also be fashioned from a MZI, as portrayed in Fig. 24.3-10(b).

The directional coupler discussed in Sec. 21.1D also serves as a 2×2 switch. As indicated in Fig. 24.3-10(c) (which reproduces Fig. 21.1-10), two waveguides can be optically coupled by placing them in close proximity. The refractive index may then be altered by applying an electric field, which can be adjusted so that the optical power either remains in the same waveguide or is transferred to the other waveguide. Switches such as these operate at a few volts and at bandwidths of tens of GHz.

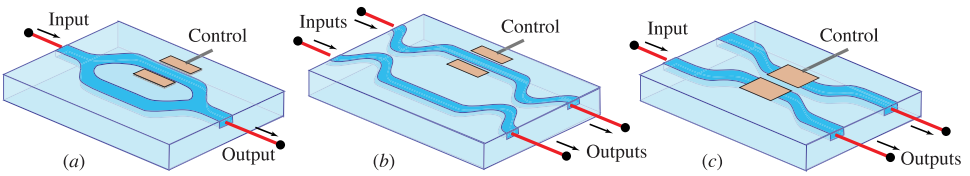


Figure 24.3-10 (a) A 1×1 switch using an integrated-photonic Mach–Zehnder interferometer (MZI). (b) A 2×2 switch using an integrated-photonic MZI. (c) A 2×2 switch using an integrated-photonic directional coupler.

An $N \times N$ integrated-photonic switch can be built by making use of combinations of 2×2 switches. A 4×4 switch, for example, may be implemented by using an arrangement of five 2×2 switches, as suggested in Fig. 24.3-3(a). This configuration can be fabricated on a single substrate in the geometry displayed in Fig. 24.3-11. Lithium niobate electro-optic switches of size 32×32 have been fabricated. The limit on the number of switches per unit area is governed by the relatively large physical dimensions of each directional coupler and the planar nature of the interconnections within the chip. However, intersecting (rather than parallel) waveguides may be used to reduce the dimensions and increase the switch packing density.

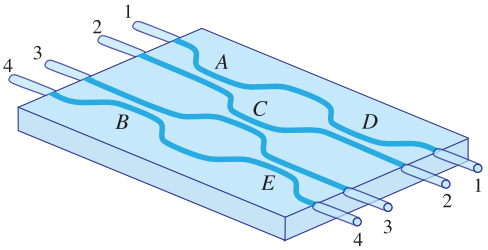


Figure 24.3-11 An integrated-photonic 4×4 switch using five directional couplers (A, B, C, D, and E) implemented on a single substrate.

It is worthy of mention that the rectangular geometry of integrated-photonics technology makes it difficult to obtain efficient coupling to cylindrical waveguides such as optical fibers. Relatively large insertion losses are encountered, especially with single-mode fibers. Also, the coupling coefficient is polarization-dependent, which requires proper selection of the polarization of the guided light and the use of polarization-maintaining input and output connecting fibers (Sec. 10.2B). Elaborate schemes are required to make polarization-independent switches.

Semiconductor Photonic Switches

Electrically controlled semiconductor devices exhibit optical properties that can be exploited for fast photonic switching. As described in Sec. 21.5, electroabsorption, based on the Franz–Keldysh effect in bulk semiconductors and on the quantum-confined Stark effect (QCSE) in multiquantum-well (MQW) structures, is useful for controlling the absorption of light at wavelengths near the bandgap wavelength by means of an electric field. Used as 1×1 switches, such electrically controlled optical modulators can exhibit switching times shorter than 20 ps. They can be fabricated in large arrays and bonded to silicon substrates for operation in a surface-normal configuration, as illustrated schematically in Fig. 24.3-12.

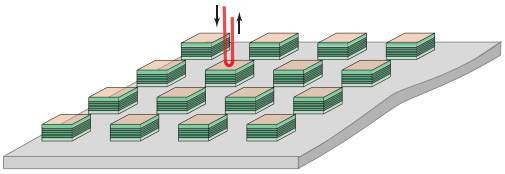


Figure 24.3-12 An array of MQW switches in a surface-normal configuration. Operation is based on the QCSE.

Another device available for photonic switching is the semiconductor optical amplifier (SOA). Since the SOA may be rapidly turned on and off by applying and removing the injected electric current (Sec. 18.2), it can be used as a 1×1 switch with switching times in the nanosecond regime. In the absence of gain (when the device is in the OFF state) it acts as a strong absorber whereas in the presence of gain (when the device is in the ON state) it becomes an amplifier; extinction ratios in excess of 40 dB can be obtained. SOA switches using InGaAsP/InP MQW structures operate at wavelengths in the vicinity of 1.55 and 1.3 μm (Sec. 18.2D).

Arrays of SOA switches may be fabricated and interconnected via optical fibers, as illustrated in Fig. 24.3-13. SOAs operated as amplifiers provide gain so they may be inserted in the circuit to compensate for the large splitting losses. Since SOAs can also function as wavelength converters, they may be used in wavelength switching where optical data carried on a carrier of one wavelength are "copied" onto a carrier of a different wavelength. Because of their nonlinear optical properties, SOAs can also be operated as ultrafast all-optical switches, as discussed in Sec. 24.3C.

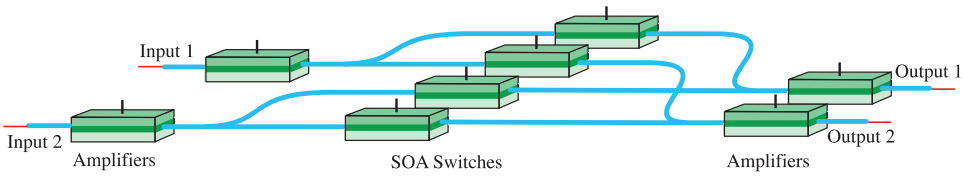


Figure 24.3-13 A 2×2 switch using four 1×1 SOA switches in the broadcast-and-select configuration illustrated in Fig. 24.3-2.

Liquid-Crystal Switches

Liquid crystals (LCs) offer yet another technology that can be used to make electrically controllable photonic switches. As described in Sec. 21.3, a liquid-crystal cell may be configured to act as an electrically controlled wave retarder or polarization rotator; these effects may be converted into intensity modulation by use of crossed polarizers.

A compact configuration for implementing a 2×2 crossbar LC switch is illustrated in Fig. 24.3-14. This is a polarization version of the Mach–Zehnder interferometer displayed in Fig. 24.3-5(c); the LC cell rotates the polarization of the beams in the interferometer arms by 90° if the control signal is present, thereby switching the connections from the bar to the cross state. The switch is polarization-independent so that the beams are directed to the desired ports regardless of their polarization state.

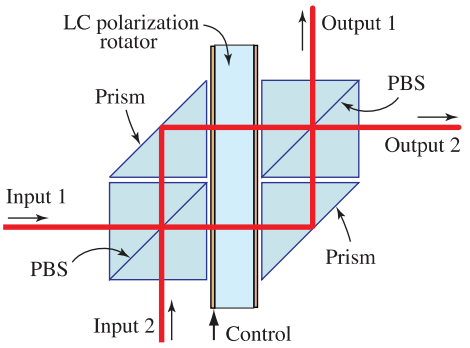


Figure 24.3-14 A 2×2 crossbar liquid-crystal switch. The two polarization components of an input beam are separated by the left-hand polarizing beamsplitter (PBS) and recombined by the right-hand PBS after passage through the liquid-crystal cell (LC), which serves as a $\pi/2$ polarization rotator if the control signal is on. Without polarization rotation, the beams entering at inputs 1 and 2 are directed to outputs 1 and 2, respectively, i.e., the switch is in the bar state. With polarization rotation, the beams are directed to the opposite output ports, corresponding to the cross state.

In an alternate switching configuration, the change of the LC refractive index caused by an applied electric field may be directly used for switching. The incoming light enters the LC at an angle, via another medium whose refractive index is selected such that total internal reflection occurs only when the electric field is applied to the LC. A large array of electrodes placed on a single liquid-crystal panel serves as a set of 1×1 switches (a digital spatial light modulator) that may be used in the broadcast-and-select mode displayed in Fig. 24.3-6 to implement an $N \times N$ crossbar switch.

Because of their relatively long switching times, LC switches are used in applications where speed is not paramount, such as in fault-protection switching and in the reconfigurable optical add–drop multiplexers (ROADMs) used in optical fiber networks (Sec. 25.5B).

Acousto-Optic Switches

Acousto-optic switches rely on the Bragg diffraction of light by sound (Sec. 20.1A). The reflectance of the diffracted light is controlled by the intensity of the sound wave whereas the angle of deflection is controlled by its frequency. An acousto-optic modulator, such as that displayed in Fig. 24.3-15(a) (which reproduces Fig. 20.1-2) is a 1×2 switch. An acousto-optic scanner, such as that illustrated in Fig. 20.2-9, is a $1 \times N$ switch. A 2×2 switch is portrayed in Fig. 24.3-15(b). If different portions of the acousto-optic cell carry sound waves of different frequencies, as shown in Fig. 24.3-15(c) (which reproduces Fig. 20.2-13), the result is an $L \times M$ switch. A bound on the maximum value of the product LM achievable with an acousto-optic device, a quantity known as the interconnection capacity, is set forth in (20.2-9). Acousto-optic switches are generally slow since the response time depends on the transit time of sound across the device. Arrays of acousto-optic cells are also available.

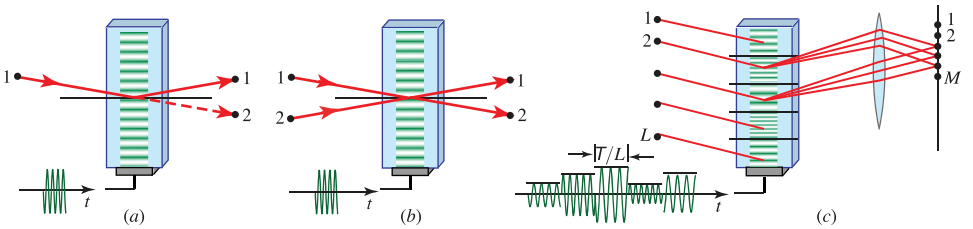


Figure 24.3-15 Acousto-optic switches. (*a*) A 1×2 switch. (*b*) A 2×2 switch. (*c*) An $L \times M$ switch.

Magneto-Optic Switches

The optical properties of magneto-optic materials are altered when a magnetic field is present. Materials exhibiting the Faraday effect, for example, act as polarization rotators in the presence of a static magnetic flux density B (Sec. 6.4B); the rotatory power ρ (angle per unit length) is proportional to the component of B that lies along the direction of propagation. When the material is placed between crossed polarizers, the optical power transmittance $\mathcal{T} = \sin^2 \theta$ is governed by the polarization rotation angle $\theta = \rho d$, where d is the thickness of the cell. This device can thus serve as a 1×1 switch controlled by the magnetic field.

Magneto-optic materials usually take the form of films (e.g., bismuth-substituted iron garnet) deposited on nonmagnetic substrates. The magnetic field is applied by making use of a pair of intersecting conductors carrying electric current. A device operates in binary mode by switching the direction of magnetization. Arrays of magneto-optic switches can be fabricated by etching isolated cells (each of size $\approx 10\ \mu\text{m} \times 10\ \mu\text{m}$) on a single film. Conductors for the electric-current drive lines are subsequently deposited using usual photolithographic techniques. Arrays of such switches (e.g., 1024×1024) are available and operate at switching times of ≈ 100 ns. Devices have also been developed that make use of plasmonic effects. Magneto-optic materials are also used for optical-disc recording, but in that case the process relies on a thermo-magnetic effect whereby the magnetization is altered by heating with a strong focused laser; weak linearly polarized laser light is used for readout.

Thermo-Optic Switches

Thermo-optic switches usually operate on the basis of the **thermo-optic effect**, which is a modification of the refractive index of a material caused by a change in its temperature. The temperature change results in a density change that in turn gives rise to an index change. The resulting index modification is almost always small, however, so that thermo-optic switches are usually operated in an interferometric configuration. For example, the thermo-optic coefficient of silica glass is $dn/dT \approx 10^{-5}$ per °C, although polymers often exhibit larger values.

Thermo-optic integrated-photonic switches are fabricated in fiber-matched silica-on-Si (SOS) waveguide configurations. An example is the Mach–Zehnder interferometer (MZI) switch illustrated in Fig. 24.3-16. A thin-film metal heater deposited directly on the waveguide in one of the interferometer branches is used to control the temperature of the material. A temperature change ΔT results in a phase shift $(2\pi L/\lambda_o)\Delta n = (2\pi L/\lambda_o)(dn/dT)\Delta T$, where L is the length of the heated region. In a silica-based switch with $L/\lambda_o = 2 \times 10^3$, for example, the temperature change required to introduce a phase shift of π is $\Delta T = 25°$ C. Other interferometric switching configurations based on arrayed waveguides (AWG) have also been fabricated using both silica and polymeric waveguides. The principal limitation of these switches is long switching time ($\approx$ ms) but they are suitable for applications such as reconfiguring light paths in optical networks.

A different kind of thermo-optic switching technology is based on the induction of changes in the refractive index of a fluid by a bubble jet initiated by a microheater. As illustrated in Fig. 24.3-17, the resulting refractive-index change in the fluid vacates its index-matching functionality and renders it a total internal reflector instead.

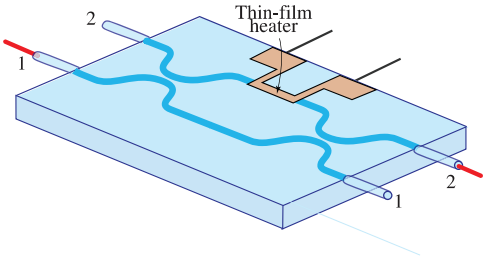


Figure 24.3-16 Thermo-optic Mach–Zehnder interferometer (MZI) switch.

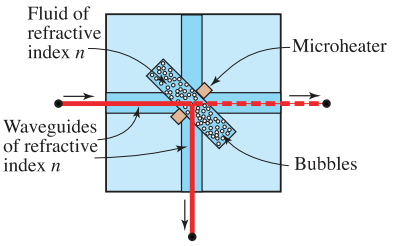


Figure 24.3-17 Bubble-jet switch.

C. All-Optical Space Switches

All-optical space switches, also called **opto-optic space switches**, operate by making use of a nonlinear optical material or device that allows light to control light. The controlling light alters a particular optical property of the nonlinear material, which in turn modifies an attribute of the controlled (signal) light such as its phase, frequency, or polarization. The modified attribute is then used to block the transmission of, or redirect the path of, the signal light. The control and signal light must be distinguishable by at least one feature, e.g., wavelength, polarization, or direction. The nonlinear process may be nonparametric or parametric (Chapter 22) and the configuration may be non-interferometric or interferometric.

As an adjunct to the parameters listed in Sec. 24.3A under the rubric *photonic switch characteristics*, the principal parameters that characterize an all-optical switch are: 1) area illuminated by the control light, 2) intensity of the control light, 3) switching energy, 4) switching time, and 5) extinction ratio. The ideal all-optical switch is an ultracompact device that operates with ultralow control-light intensity and ultralow switching energy at an ultrashort switching time and with high extinction ratio.

Devices that are currently used as all-optical switches include: 1) semiconductor optical amplifiers (SOAs) and 2) nonlinear optical waveguides and fibers. Micro- and nanostructures that make use of solitons, photonic crystals, plasmonic structures, metamaterials, and ring nanocavities promise stronger nonlinearities, lower operating power, and shorter switching times.

SOA Switches

Electrically controlled semiconductor photonic switches were discussed in Sec. 24.3B. In this section we turn to all-optical semiconductor optical amplifier (SOA) switches that operate via nonparametric nonlinear processes: **cross-gain modulation** (**XGM**) and **cross-phase modulation** (**XPM**).

In a XGM-SOA switch [Fig. 24.3-18(a)], the control light has higher power than the signal light, and the two are distinguished by different frequencies or different polarizations. When the control light is absent the signal light is amplified; when it is present the amplifier is saturated and its gain is substantially reduced because of carrier-density depletion, resulting in the absence of amplification. High and low values of the output signal represent the ON and OFF states of the switch, respectively. The contrast ratio between these states typically lies in the range of 10–20 dB.

In a XPM-SOA switch [Fig. 24.3-18(*b*)], the amplifier is unsaturated and the control light, if present, alters the refractive index of the SOA, which is dependent on the carrier density. The signal light then undergoes a phase shift that is sensed by an interferometer. If the presence of the control light results in a phase difference between the interferometer branches that changes from 0 to π, the signal light is directed from one output port to the other and the device functions as a 1×2 switch. Although SOAs are compact devices and can be integrated in arrays, their switching time is limited to the 10–100-ps range.

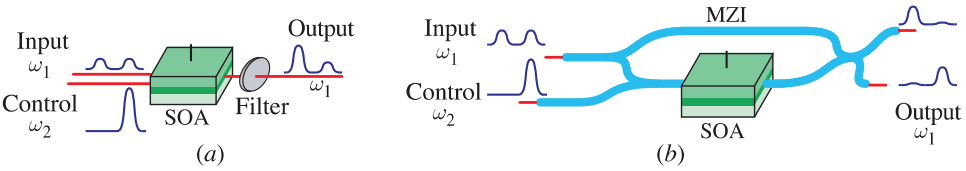


Figure 24.3-18 SOA all-optical switches. (*a*) A 1×1 switch based on cross-gain modulation (XGM). (*b*) A 1×2 switch makes use of cross-phase modulation (XPM) and a Mach–Zehnder interferometer (MZI).

Parametric Switches

As described in Chapter 22, second- and third-order nonlinear materials support numerous nonlinear parametric wave-mixing effects, including sum-frequency generation (SFG), optical Kerr effect, self-phase modulation (SPM), cross-phase modulation (XPM), cross-gain modulation (XGM), four-wave mixing (FWM), optical frequency conversion (OFC), as well as optical solitons. A number of all-optical switches and gates are based on these effects.

SFG switch. Illustrated in Fig. 24.3-19 is an example of a switch based on sum-frequency generation (SFG) in a second-order nonlinear waveguide such as a periodically poled lithium niobate (PPLN) crystal (Sec. 22.2E). If the quasi-phase matching condition is satisfied, the signal and control waves, of frequencies ω_1 and ω_2, respectively, generate a new wave at the sum frequency $\omega_3 = \omega_1 + \omega_2$. The two original waves are depleted in the process, but this only happens when both the signal *and* control waves are present. Hence, in the presence of the control wave, the signal wave is extinguished whereas in the absence of the control wave, the signal wave is transmitted through the waveguide without depletion. The SFG process therefore serves as a 1×1 switch governed by the control wave.

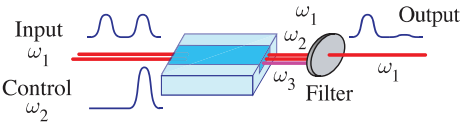


Figure 24.3-19 An all-optical 1×1 switch based on depletion resulting from parametric sum-frequency generation (SFG) in a second-order nonlinear waveguide (WG).

XPM-MZI switch. As illustrated in Fig. 24.3-20, the refractive index of a third-order nonlinear medium that exhibits cross-phase modulation (XPM), such as a **highly nonlinear fiber (HNLF)**, is altered by the control wave. The phase of the signal wave is thus modified and this is sensed by a Mach–Zehnder interferometer (MZI) that directs the signal wave to either of the two output ports, depending on the presence or absence of the control wave. The control wave is distinguished from the signal wave by its different frequency and a filter is used to permit only the signal wave to pass through the interferometer. The result is a 1×2 switch governed by the presence or absence of the control wave.

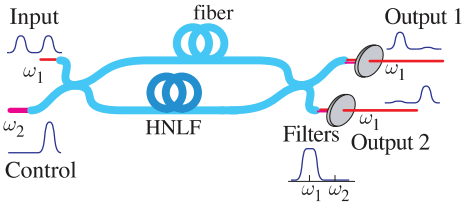


Figure 24.3-20 An all-optical 1×2 switch based on cross-phase modulation (XPM) in a highly nonlinear fiber (HNLF) placed in a Mach–Zehnder interferometer (MZI).

XPM microring switch. An integrated-optic dielectric microring resonator, such as that displayed in Fig. 11.0-1(b), whose resonance frequency is optically controllable can be used as an ultracompact, all-optical switch. As illustrated in Fig. 24.3-21, the input (signal) and control waves are guided through straight waveguide segments adjacent to the microring. In the absence of the control wave, the input wave is coupled into the resonator, which is designed to have a matching resonance frequency, and is dropped. The presence of the control wave introduces a refractive index change $\Delta n = n_2 I$ via the optical Kerr effect (Sec. 22.3A), which alters the resonance frequency of the resonator proportionally. Being off-resonance, the signal wave is then no longer coupled to the resonator and is fully transmitted to the output. The result is a 1×1 switch governed by the presence or absence of the control wave. Microring switches have been successfully implemented in III–V semiconductors and in silicon-on-insulator (SOI) technology. Though the Kerr effect in crystalline silicon is small, other nonlinear mechanisms give rise to a sizable intensity-dependent refractive index in Si: free-carrier effects and two-photon absorption (Example 16.3-4) and the thermo-optic effect (Sec. 24.3B). Switching times of 25 ps at 1.55 μm have been demonstrated.

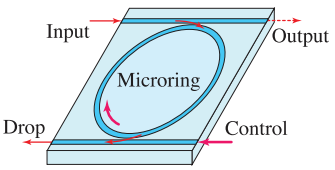


Figure 24.3-21 An all-optical 1×1 switch based on altering the resonance frequency of a microring resonator via a change in refractive index effected by the control wave. The signal wave is blocked when its frequency is on-resonance, and transmitted when it is off-resonance.

XPM retardation switch. This switch is based on the optical Kerr effect in an anisotropic nonlinear medium. Propagation of the control wave through such a medium creates different changes in the principal refractive indices so that the medium serves as a wave retarder for the signal (input) wave, thereby changing its state of polarization. If a birefringent crystal or HNLF is placed between crossed polarizers it can then function as an all-optical ON–OFF switch. When the retardation is 0, the input wave is blocked and the switch is in the OFF state and when the control wave provides a retardation of π, the signal wave is transmitted and the switch is in the ON state. Alternatively, a polarizing beam splitter (PBS) can be used to direct the output from one port to another as the control wave is turned on or off. This 1×2 switch is illustrated in Fig. 24.3-22.

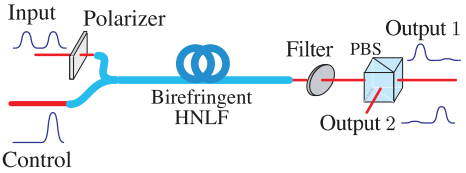


Figure 24.3-22 An all-optical 1×2 switch based on cross-phase modulation (XPM), which alters the retardation in a birefringent HNLF.

XPM frequency-shifting switch. The non-interferometric configuration illustrated in Fig. 24.3-23 may be used for switching ultrashort light pulses. As described in Sec. 23.1A, a phase shift that varies linearly with time is equivalent to a frequency shift proportional to the slope of the control pulse power profile. Accordingly, the time-varying intensity at one edge of the control optical pulse introduces, via XPM, a time-varying phase and an associated frequency shift in the signal pulse. An appropriate bandpass filter (BPF) is used to block the frequency-shifted wave. When the control pulse is absent, the frequency shift does not occur and the signal pulse is transmitted.

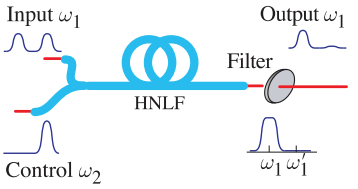


Figure 24.3-23 An all-optical 1×1 switch based on the frequency shift introduced by cross-phase modulation (XPM) near the edge of the control pulse in a highly nonlinear fiber (HNLF).

FWM switch. A switch based on four-wave mixing (FWM) (Sec. 22.3D) in a nonlinear optical fiber is illustrated in Fig. 24.3-24. Signal (input) and control waves with frequencies ω_1 and ω_2, respectively, are launched into the fiber. As they co-propagate, they generate two new waves with frequencies ω_3 and ω_4 satisfying the FWM condition $\omega_1 + \omega_2 = \omega_3 + \omega_4$, and both the signal and control waves are depleted in the process. The system is much like a depleted-pump optical parametric amplifier (OPA) in which the two incoming waves act as pumps. This effect can also be regarded as an example of XGM. The amount of depletion can be signal by adjusting the power of the incoming waves. When the control wave is absent, the signal wave emerges with no significant loss since the parametric interaction process is thwarted.

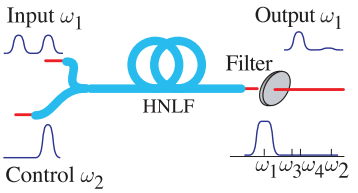


Figure 24.3-24 An all-optical 1×1 switch based on depletion of the signal wave caused by four-wave mixing (FWM) in the presence of the control wave in a highly nonlinear fiber (HNLF).

Soliton Switches

Optical solitons are ultrashort pulses that propagate in nonlinear dispersive optical fibers without spreading (Sec. 23.5B). An all-optical switch may be realized by using one optical soliton to control the routing of another. The interaction between the two solitons may take the form of a collision or a recombination into a single vector soliton. In either case, some optical property of the input soliton is changed by the interaction, and the altered property is used to effect the routing.

Soliton-collision switch. If two solitons with slightly different frequencies, and hence slightly different group velocities, collide (pass through one another), the arrival time and the phase are altered for each soliton. One of the pulses serves as the control pulse, and the other as the signal pulse. Either the time delay or the phase shift that accompanies the collision with the control pulse is used to route the signal pulse. Time-based routing is implemented by making use of an optical gate that opens during a prescribed time window. Phase-based routing is effected by making use of an interferometer.

Vector-soliton switch. A **vector soliton** comprises two orthogonally polarized optical pulses copropagating through a nonlinear birefringent fiber. Since both pulses must be present for the vector soliton to form, the system may be used as a photonic switch with one pulse serving to control the other.

Two pulses with orthogonal polarizations travel in a birefringent fiber at slightly different group velocities and therefore separate in time, a phenomenon known as walk-off (Sec. 23.5A). If the fiber is also nonlinear, cross-phase modulation (XPM) (Sec. 22.3C) results in a frequency upshift in one pulse and a frequency downshift in the other. Because of group velocity dispersion (GVD), these shifts are accompanied by a change in the group velocities. When the group velocity difference due to the birefringence is exactly compensated by that due to GVD (via XPM), the two pulses travel jointly as a single vector soliton, a condition also known as *soliton trapping*.

As illustrated in Fig. 24.3-25, a 1×1 vector-soliton switch may be implemented by using one of the two orthogonally polarized pulses as the control pulse, and the other as the signal to be transmitted or blocked. If the two pulses have the same wavelength λ, when traveling through the nonlinear birefringent fiber they form a vector soliton whose components have wavelengths shifted to $\lambda \pm \delta\lambda$. One of these components is selected by a filter and constitutes the output of the switch. In the absence of the control pulse, the vector soliton is not formed and the wavelength is not shifted, in which case the light is blocked by the filter.

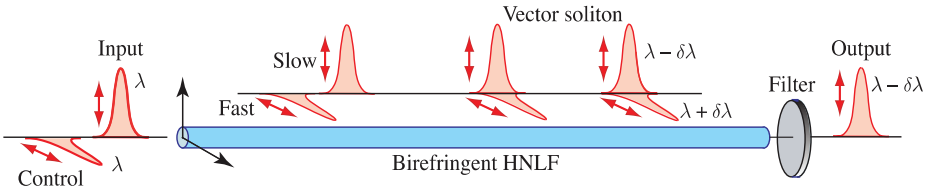


Figure 24.3-25 An all-optical switch using vector solitons in a highly nonlinear fiber (HNLF).

Photonic-Crystal and Plasmonic Switches

The optical Kerr effect may also be used in various all-optical switches implemented in photonic crystal and plasmonic nanostructures and microstructures. In these devices the control light alters the refractive index of the dielectric material, thereby altering a characteristic of the structure through which the signal light propagates. The result is that the light is either transmitted (ON state) or blocked (OFF state).

Photonic-crystal switches. As described in Sec. 7.2, photonic crystals are periodic structures characterized by bandgaps, i.e., frequency bands within which light cannot propagate. The bandgaps can be frequency shifted by altering the refractive index of the material. In the absence of the control light, the frequency of the signal light lies inside the bandgap, but near its edge, so the light cannot propagate and the switch is in the OFF state. When the control light is applied, the bandgap is shifted so that the frequency of the signal light falls outside the bandgap and the light can propagate, so the switch is converted to the ON state. An all-optical switch may also be implemented by using a photonic-crystal nanocavity formed by a lattice defect. Since the resonance frequency of the nanocavity mode depends on the refractive index of the material, it can be altered by the control light. The optical signal then will or will not be coupled to the nanocavity depending on whether its frequency matches or mismatches the nanocavity resonance frequency, thereby enabling switching action. Experimental demonstrations of photonic-crystal, all-optical switching with response times in the subpicosecond range and pump powers of kW/cm^2 have been reported.

Plasmonic switches. The resonance frequency of a surface plasmon polariton (SPP) mode is highly sensitive to the permittivity of the adjacent dielectric material (Sec. 8.2B), which can be altered by the control light via the optical Kerr effect. The signal light is coupled to the SPP mode, and the switch is in the OFF state when its frequency matches that of the of the SPP mode. Otherwise, the signal light propagates uninterrupted (ON state). Plasmonic all-optical switches are suitable for integrated-photonic implementations and offer subpicosecond response times that are limited by the relaxation time of the plasmonic resonance. However, typical propagation distances of propagating SPP waveguides are rather short because of metallic losses.

Switching Time

The switching time of an all-optical switch is limited by the duration of the control optical pulse and the response time of the nonlinear process responsible for the switching action. Since ultrashort optical pulses of a few femtoseconds duration (a few optical cycles) are readily available, the much slower nonlinear interaction process sets the ultimate limit on the switching time, which is highly dependent on the switching material.

Semiconductors such as GaAs, InSb, InAs, and CdS exhibit strong optical nonlinearities as a result of excitonic effects at wavelengths near the band edges. Switch-on times are typically on the order of a few picoseconds while switch-off times, dominated by relatively slow carrier recombination, typically extend to hundred picoseconds. Semiconductor optical amplifier (SOA) switches based on XGM and XPM are similarly limited by the intrinsically slow recovery time of the amplifier gain, exhibiting switch-off times in the 10–100-ps range. In contrast, switches based on XPM and FWM in highly nonlinear optical fibers have far shorter response times, typically < 100 fs. Switches that make use of fiber soliton technology also operate at sub-picosecond switching times.

The non-symmetric nature of the temporal response of the nonlinear effect in semiconductors (short rise time at the onset of the control optical pulse and a far longer decay time following removal of the pulse) can be exploited in designing switches whose switching times are limited by the short rise time rather than by the long decay time, as described in Example 24.3-1.

EXAMPLE 24.3-1. ***Ultrafast Nonlinear Asymmetric Sagnac Interferometer Switch.*** The response time of XPM in a third-order nonlinear-optical material is characterized by a short rise time and a long decay time. The switching time can be reduced by making use of an interferometric configuration in which both branches of the interferometer include the same nonlinear element but the signal light pulse crosses it at different times. This is readily implemented by a fiber Sagnac interferometer with a nonlinear optical element placed at an asymmetric location within the fiber loop, as illustrated in Fig. 24.3-26. When the input signal pulse enters the loop from fiber 1, it is split by a symmetric coupler into a clockwise pulse and a counterclockwise pulse of equal amplitudes. If the two pulses encounter the same phase shift as they make their round-trip paths around the loop, they recombine and return back into fiber 1 and exit from output port 1. If they undergo phase shifts differing by π, on the other hand, they recombine and emerge into fiber 2 and exit from output port 2. These are the two states of a 1×2 switch.

The nonlinear element is controlled by a short control optical pulse that changes its refractive index by Δn. This change builds up with a short rise time t_i and decays with a much longer relaxation time t_r. Since the nonlinear element is placed at an offset location within the fiber loop, the two signal pulses cross it at different times, τ_1 and τ_2. If both pulses cross the nonlinear element when it is active, i.e., in the presence of the full change Δn, they undergo the same phase shift and the recombined pulse is received in fiber 1. This also occurs if both pulses cross the nonlinear element when it is inactive. However, if one pulse crosses when the nonlinear element is active and the other when it is inactive, they undergo different phase shifts, and if the phase difference is π, the pulse enters fiber 2, and emerges from output port 2. The switching action is therefore governed by the time difference

$\tau_1 - \tau_2$, which is proportional to the distance of the nonlinear element from the mid-point of the fiber loop. If $\tau_1 - \tau_2$ is slightly greater than the rise time t_i, the switching action can be controlled with precision limited by the rise time, instead of by the full response time t_r. Femtosecond switching times have been achieved, so that the switch can be operated at terahertz bandwidths. This switch has been used for time-division demultiplexing and is known as the **Terahertz Optical Asymmetric Demultiplexer** (**TOAD**).[†]

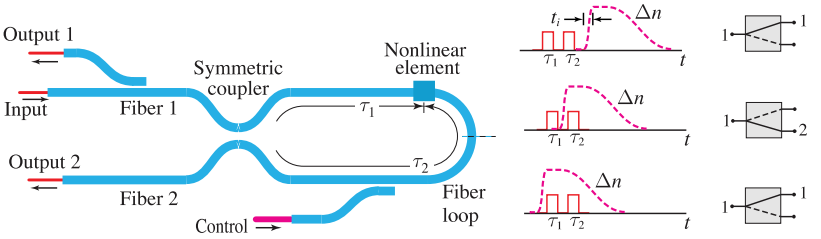


Figure 24.3-26 An all-optical-fiber nonlinear asymmetric Sagnac interferometer used as a 1×2 switch. The switch is controlled by an optical pulse that initiates a refractive index change Δn in a nonlinear element placed at an offset location within the interferometer loop. The input pulse coming from fiber 1 is split into clockwise and counterclockwise pulses that traverse the nonlinear element at different times. The switch changes the connection from output port 1 to output port 2 if one of these pulses arrives just before, and the other just after, the onset of Δn. This results in a phase difference of π and a diversion of the output pulse to output port 2.

Switching Energy

The switching energy $E = TAI_c$ is the product of the switching time T, the switch area illuminated by the control light A, and the intensity of the control light I_c. As indicated previously, the minimum switching time is limited by the response time of the nonlinear process underlying the switching action. Limits on the switch area are governed by diffraction effects, which make it difficult to couple optical power into and out of devices whose dimensions are smaller than a wavelength of light. If T and A are maintained at their lowest possible values, further reduction in the switching energy E can be achieved by making use of a material with stronger nonlinearity, i.e., a material that requires a lower intensity I_c for operation at a high switching efficiency. Certain device configurations, such as those making use of resonant-cavity-enhanced nonlinear interactions, can enhance the effective nonlinearity, but they have the concomitant property of prolonging the response time. In the extreme limit of $T = 10$ fs and $A = 1\ \mu\text{m}^2$, operation with a small switching energy $E = 1$ fJ requires a control-beam intensity $I_c = 1\ \text{kW/cm}^2$, which is modest for conventional nonlinear materials.

The switching energy of GaAs devices typically lies in the range 1–10 pJ, but it is in principle possible to reduce it to the fJ regime. InGaAsP optical amplifiers have been operated with switching energies less than 1 fJ, but doing so comes at the expense of switching times of hundreds of ps. Switches that make use of solitons in optical fibers have also been implemented with switching energies in the range of tens of pJ.

Quantum limit. The minimum switching energy is ultimately limited by the photon nature of light and its inherent uncertainty, which is particularly evident at low light energies. A switching energy $E = 20$ aJ, for example, corresponds to an average photon number $\bar{n} = E/h\nu = 100$ at $\lambda_o = 1\ \mu\text{m}$, but the actual photon number n

[†] See J. P. Sokoloff, P. R. Prucnal, I. Glesk, and M. Kane, A Terahertz Optical Asymmetric Demultiplexer (TOAD), *IEEE Photonics Technology Letters*, vol. 5, pp. 787–790, 1993.

is random. If the light arises from a laser or an LED, the photon number n is Poisson distributed with mean $\overline{n} = 100$ and width $\sqrt{\overline{n}} = 10$, as is understood from Sec. 13.2C. Thus, if the switch is designed to be activated when a fixed threshold number of photons is received, there is always a finite probability that the actual number of photons falls below that threshold and the switch fails. This type of switching error can be minimized by making use of greater switching energies. A switching energy $E = 1$ fJ, for example, corresponds to a larger mean photon number, $\overline{n} = 5000$, and a relatively narrower uncertainty ≈ 70 photons. In that case, the switching error is minimal and switch activation almost always occurs when desired.

Heat dissipation. An important practical limit on all-optical switching arises from the difficulty of thermally transferring the heat dissipated by the switching process. This limitation is particularly severe when the switching is carried out at the maximum allowed switching rate. With a switching energy E, a switching time T, and a maximum switching rate of $1/2T$ operations per second, the power dissipated is $E/2T$, a heat load that can be substantial for large values of E and small values of T. The need to remove this dissipated power can make the combination of very high switching energies and very short switching times difficult. Thermal effects are less restrictive when the switch is operated below its maximum permitted repetition rate, of course, since the energy associated with each switching operation then has more time to be dissipated.

A key issue pertaining to the usefulness of photonic switches is the ability to fabricate them in large arrays on a single chip. Again, heat dissipation can be a limiting factor; for an array of N switches per unit area, the total power density required to be removed is $NE/2T$. Consider, for example, an array of 100×100 switching elements on a 1-cm^2 GaAs chip, so that $N = 10^4$ cm^{-2}. If the switching energy $E = 1$ pJ and the switching time $T = 50$ ps, then the power required to be removed is $NE/2T = 100$ W/cm^2, which is manageable with good thermal engineering. Such a chip can carry out $N/2T = 10^{14}$ switching operations per second, which is large in comparison with electronic supercomputers.

D. Wavelength-Selective Switches

The switches described to this point are space-domain switches, meaning that they establish transmission paths that route optical beams among specific physical positions (the input and output ports of the switch). Their wavelength-domain logical counterparts, known as **wavelength-selective switches** (**WSSs**), are illustrated by the following examples:

EXAMPLE 24.3-2. ***Reconfigurable Wavelength Selector.*** The wavelength selector displayed in Fig. 24.3-27 is an example of an optical device that makes use of a combination of passive wavelength routers and space switches. This switch selects one or more wavelengths from an incoming beam with N wavelengths. It uses a demultiplexer (DMUX) to separate the N wavelength components, followed by a set of N 1×1 switches to select the desired wavelengths, and then a multiplexer (MUX) to reconstitute the output beam, as shown in the figure.

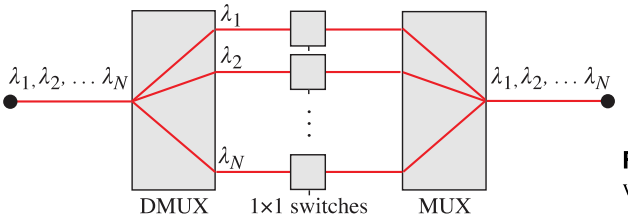


Figure 24.3-27 A reconfigurable wavelength selector.

EXAMPLE 24.3-3. ***Reconfigurable Optical Add–Drop Multiplexer (ROADM).*** The ROADM is a reconfigurable OADM with the option to add, drop, or pass-through specific wavelength channels, as illustrated in Fig. 24.3-28. It uses a demultiplexer (DMUX) and a multiplexer (MUX), along with a 1×2 switch and a 2×1 switch for each add–drop channel. The ROADM is the core element of wavelength-division multiplexing (WDM) networks (Sec. 25.3C). It acts on each wavelength in an incoming optical fiber by either allowing it to pass or routing it to a drop fiber, which directs it to another client or to an outgoing fiber connected to another node in the network. Data in the dropped wavelength may be replaced by new data added from another fiber carrying data from another client or another node. ROADMs deployed in early fiber networks used free-beam diffraction gratings as demultiplexers and multiplexers along with arrays of transmissive liquid crystal shutters as switches. These configurations are bidirectional, handling both left-to-right and right-to-left traffic. ROADMs that are deployed in more modern and larger networks use so-called multi-degree ROADMs (M-ROADMs), which provide wavelength-based interconnection among three or more intersecting fiber routes. The switching core requires an $N \times N$ WSS, which is often implemented by a bank of $1 \times N$ WSSs, each using a route-and-select design based on beam-steering elements such as metal micromirrors or diffractive arrays of liquid-crystal cells.

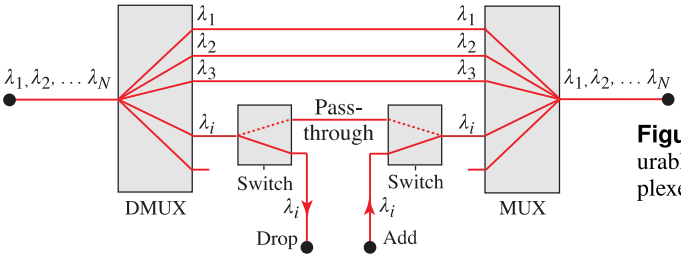


Figure 24.3-28 A reconfigurable optical add–drop multiplexer (ROADM).

EXAMPLE 24.3-4. ***Wavelength-Channel Interchange (WCI).*** The WCI switch, also called the **λ switch**, routes data between wavelength channels in the same optical beam. An $N \times N$ WCI switch may be implemented by mapping the wavelength channels to the space domain using a demultiplexer, converting the wavelengths using a bank of N **wavelength converters** (**WCs**), and recombining the channels into a single beam by use of an $N \times 1$ coupler, as shown in Fig. 24.3-29. A wavelength converter changes the wavelength of a beam without altering the data, i.e., it "copies" the data from one wavelength channel to another.

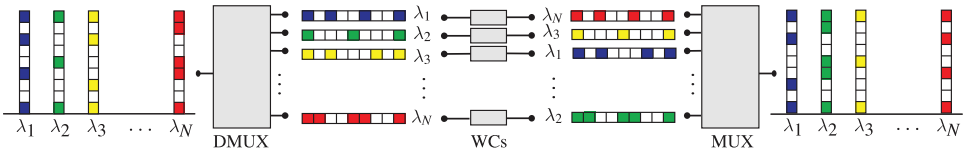


Figure 24.3-29 Implementation of a wavelength-channel interchange (WCI). Data bits are depicted as colored and white squares. In this example, data in wavelength channel 2 (green) of the input beam are routed to wavelength channel 3 (yellow) of the output beam. This switch is implemented by use of a wavelength demultiplexer to separate and direct the wavelength channels to a bank of wavelength converters. A fan-in $N \times 1$ coupler recombines the switched channels into a single beam.

Multidimensional Space–Wavelength Switches

The previous examples of wavelength-domain switches involve a single optical beam with multiple wavelength channels. Switching may also be applied to multichannel multiple beams. Consider, for example, the switching of N beams, each with one of N wavelength channels. The switch redistributes the wavelength channels among the beams. Two implementations are displayed in Fig. 24.3-30.

The first implementation uses a broadcast-and-select router to redirect the wavelength channels to different ports. This is accomplished by means of a star coupler that broadcasts the contents of all N beams to every one of a set of wavelength filters, each of which is tuned to a single wavelength channel [Fig. 24.3-30(a)]. Finally, for further processing, the wavelengths of the switched channels are converted to the original wavelengths (without changing their data content), by use of a bank of wavelength converters (WCs).

The second implementation makes use of two sets of WCs with an arrayed waveguides (AWG) router placed between them, as portrayed in Fig. 24.3-30(b). The first WC converts the wavelengths to values that satisfy the AWG equation (24.2-7) for the appropriate destinations. The AWG switch is more efficient than the broadcast-and-select switch since the latter wastes considerable power at the filters. However, the broadcast-and-select switch has the advantage of being reconfigurable.

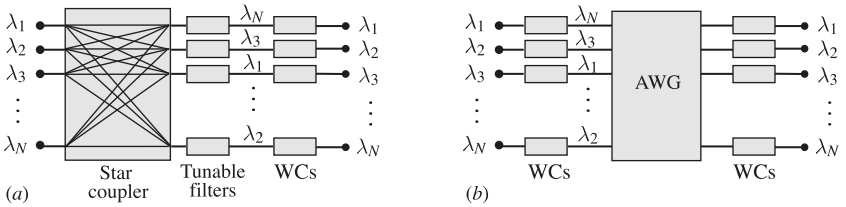


Figure 24.3-30 (a) Broadcast-and-select space–wavelength switch. (b) Arrayed waveguides (AWG) space–wavelength switch.

Implementations of Wavelength Converters

A wavelength converter (WC) transfers data carried by an optical beam at one wavelength to a different wavelength. The wavelengths often represent the channels of a WDM optical fiber communication system (Sec. 25.3C); their wavelength separation is then not large since they lie in the same band. Wavelength converters are implemented by making use of nonlinear optical devices, nonparametric or parametric, similar to those described in Sec. 24.3C for all-optical switching.

SOA WCs In nonparametric WCs, the intensity of the first beam, which is modulated by the data, is used to alter an optical property of a medium, such as the gain coefficient, absorption coefficient, or refractive index of a semiconductor, in proportion to the intensity, so that the data is "written" into the medium. A second beam of different wavelength transmitted through the medium is then modulated by the altered property, so that the data are "read" by, and transferred to, the second beam.

An example is the process of cross gain modulation (XGM) in a saturated semiconductor optical amplifier (SOA), for which the gain is a decreasing function of the intensity, as depicted in Fig. 24.3-31(a). When the original intensity-modulated beam is transmitted through this device, the gain is modulated as an inverted function, and so is the intensity of the read beam. Another example is the process of cross-phase modulation (XPM) in an unsaturated SOA, for which the refractive index is modulated by the write beam since it is dependent on the carrier density. The read beam is therefore phase modulated. An interferometer must be employed to convert the phase modulation into intensity modulation, as depicted in Fig. 24.3-31(b).

Parametric WCs. In a WC based on a parametric interaction, beams of different wavelengths are coupled via the nonlinear effect. In a second-order nonlinear medium

(Sec. 22.2), for example, a wave of frequency ω_1 may be downconverted to a frequency $\omega_2 = \omega_3 - \omega_1$ with the help of an auxiliary wave of frequency ω_3. The amplitude of the downconverted wave is related to that of the original wave, so that the data embedded in the magnitude or in the phase of the original wave are transferred to the downconverted wave. The principal difficulty of using this three-wave mixing process is that if the frequencies ω_1 and ω_2 are close to each other, the frequency ω_3 of the auxiliary wave must be approximately twice as large. If it is desired to use only waves of approximately the same frequencies, a cascade of two nonlinear parametric processes may be implemented. The first process could be a second-harmonic generation (SHG) process in which ω_1 is converted to $2\omega_1$, while the second process is a three-wave-mixing downconversion process in which a wave of frequency $\omega_2 = 2\omega_3 - \omega_1$ is generated. All three waves now have approximately the same frequency.

Alternatively, a four-wave mixing (FWM) process implemented via a third-order nonlinearity, such as occurs in optical fibers, may be implemented, as displayed in Fig. 24.3-31(*c*). As described in Sec. 22.3D, this process involves the mixing of four-waves of frequencies satisfying the relation $\omega_1 + \omega_2 = \omega_3 + \omega_4$. In the partially-degenerate case we have $\omega_3 = \omega_4 = \omega_0$, so that $\omega_2 = 2\omega_0 - \omega_1$.

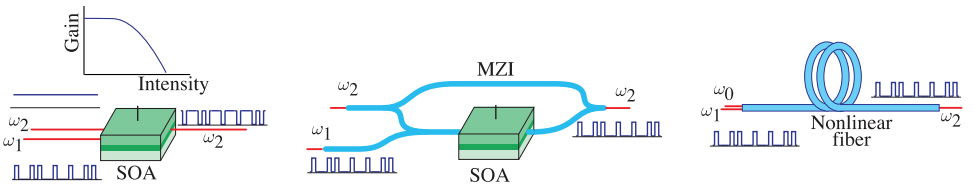


Figure 24.3-31 Wavelength conversion. Data is transferred from a beam of frequency ω_1 to a beam of frequency ω_2. (*a*) Cross-gain modulation (XGM) in a saturated semiconductor optical amplifier (SOA). (*b*) Cross-phase modulation (XPM) in an unsaturated SOA. The phase modulation of the converted beam is transformed into intensity modulation by use of a Mach–Zehnder interferometer (MZI). (*c*) Partially-degenerate four-wave mixing (FWM) in a third-order nonlinear medium, such as an optical fiber, using an auxiliary wave of frequency $\omega_0 = \frac{1}{2}(\omega_1 + \omega_2)$.

E. Time-Domain Switches

The **time-domain switch** routes signals between time slots, as illustrated in Fig. 24.3-32. In digital communication systems, a signal is divided into a sequence of time frames of equal duration, each of which is divided into N time slots where the data reside. An example of a time-domain switch is the **time-slot interchange** (**TSI**) switch, which transfers the data resident in the ℓth time slot of each frame to the mth time slot of the same frame. This is analogous to the wavelength-channel interchange (WCI) switch described in Example 24.3-4.

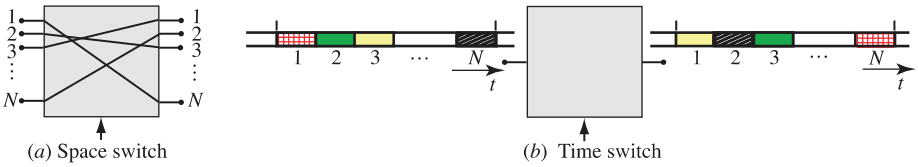


Figure 24.3-32 Correspondence between time- and space-domain switches. (*a*) Space-domain switch. In the example shown, data in line 2 are routed to line 3. (*b*) Time-domain switch implementing a time-slot interchange (TSI). In the example shown, data in time slot 2 are routed to time-slot 3 in each frame.

Two-dimensional space–time switches employ a combination of time-domain and space-domain switches. The switch connects a set of input lines, each carrying a digital signal composed of a sequence of time frames, to a similar set of output lines. Data in each time slot of each input line are transferred to one, or several, time slots in one or several output lines, in accordance with a prescribed rule. An example is the **time–space–time (TST) switch**, which consists of a cascade of a time-slot interchange (TSI), a space switch, and another TSI, as shown in Fig. 24.3-33.

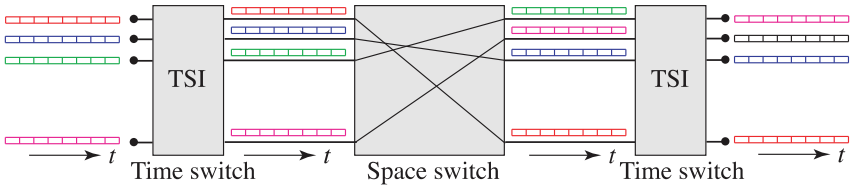


Figure 24.3-33 Time–space–time (TST) switch.

Time-Division Multiplexing and Demultiplexing

A simple example of a space–time switch is the time-division demultiplexer. It has one input line and N output lines, where N is the number of time slots in each frame. The switch routes data in the ℓth time slot of the input line to the ℓth time slot of the ℓth output line; $\ell = 1, 2, \ldots, N$. The process is repeated periodically in all frames. This switch is therefore equivalent to a time-to-space mapping.

In the **time-division demultiplexer** shown in Fig. 24.3-34, for example, there are $N = 4$ time slots per frame. The slots contain data that takes the form of pulses of various heights. The switch directs the first pulse to the first output port, and the second pulse to the second output port, and so on. Such a switch could be constructed by use of a $1 \times N$ space switch connecting the input port sequentially to one of its four output ports.

The inverse of the time-division demultiplexer, called a **time-division multiplexer** (**TDM**), interleaves pulses from N separate ports to form a single sequence of pulses at one output port. This inverse operation is readily visualized in Fig. 24.3-34 by exchanging the roles of the input and output ports so that the pulses travel from right-to-left, instead of from left-to-right. The $1 \times N$ time-division demultiplexer may be implemented by making use of N 1×1 ON–OFF switches, as illustrated in Fig. 24.3-2(a), that are turned on and off sequentially with control pulses from a clock.

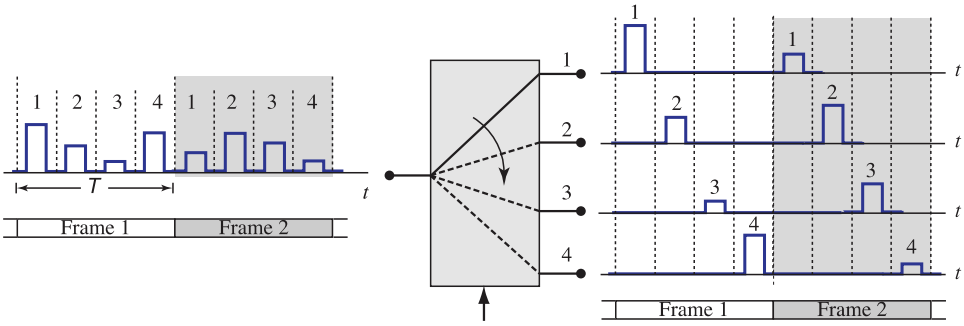


Figure 24.3-34 Time-division demultiplexing with $N = 4$.

Optical Time-Division Multiplexing (OTDM)

An optical implementation of the TDM is illustrated in Fig. 24.3-35(a). Copies of the input beam are transmitted through a set of N 1×1 photonic switches controlled by a set of optical pulses from a clock delayed by multiples of the time delay T/N, where T is the frame period. In an alternate implementation, portrayed in Fig. 24.3-35(b), copies of the input beam are successively delayed by multiples of T/N so that the N input pulses are separated in space but synchronized in time; the 1×1 switches are controlled by the same clock signal. The system is similar to that used to detect the temporal profile of a single optical pulse (Fig. 23.6-5). Optical delays may be implemented by using lengths of optical fiber (approximately 5 ns/m for silica-glass fibers). The 1×1 switches may be implemented optically by using an all-optical nonlinear interferometric switch such as the Terahertz Optical Asymmetric Demultiplexer (TOAD) discussed in Example 24.3-1.

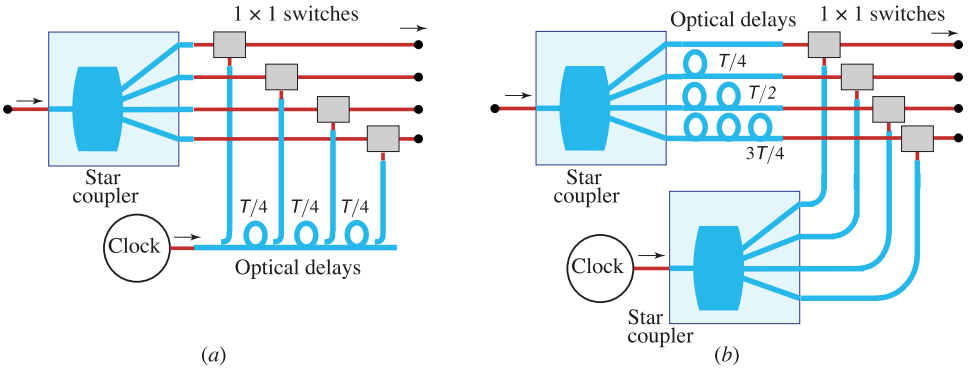


Figure 24.3-35 Two implementations of time-division demultiplexing using star couplers, optical time delays, and 1×1 photonic switches. In this illustration, $N = 4$.

Optical Time-Slot Interchange (TSI)

The TSI switch that was shown in Fig. 24.3-32 is a time-domain switch that interchanges data within the time slots of each frame. Optical implementations may be effected by combining space and space–time switches. The configuration schematized in Fig. 24.3-36, for example, relies on the following sequence of switches: 1) a time-division demultiplexer (DMUX) that routes the time slots to separate lines in space (time-to-space mapping); 2) time delays to synchronize the pulses in a single time slot of duration T/N; 3) an $N \times N$ cross-connect space switch to implement the desired interchanges; 4) a second set of time delays to restore the pulses to their original time slots; and 5) a time-division multiplexer (MUX) to bring these time slots to a single time line (space-to-time mapping).

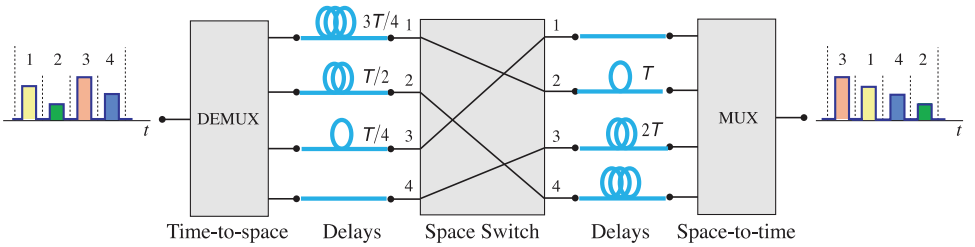


Figure 24.3-36 An implementation of optical time-slot interchange (TSI).

Optical Programmable Time Delays and Buffers

Controllable time delays are essential components for time-domain switching. **Buffers** are memory elements used to temporarily store data or to compensate for differences in data flow rates. As is understood from Figs. 24.3-35 and 24.3-36, such delays may be introduced by making use of optical fibers of appropriate length (silica-glass fibers introduce delays of approximately 5 ns/m). Programmable delays may be implemented by allowing the optical pulses to circulate in a fiber loop for a programmable number of cycles. As illustrated in Fig. 24.3-37, this may be accomplished by using a crossbar switch that permits the pulse to enter the loop at the desired time and releases it after a specified number of cycles.

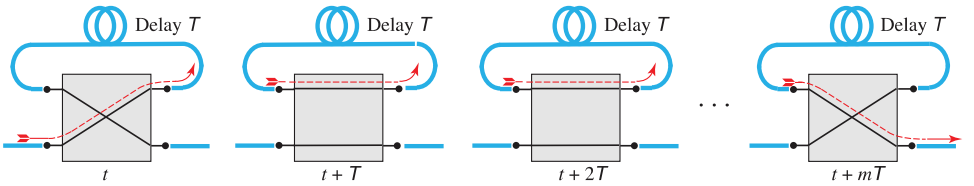


Figure 24.3-37 Programmable delay line using a fiber loop and a crossbar switch. At time $t = 0$, the switch is in the cross state so that the optical pulse is admitted into the loop. At time $t = T$, the pulse returns back to the input port of the switch, which is then placed in the bar state so that the pulse undergoes another round trip that incurs an additional delay T. At time $t = mT$, the pulse is released by reconfiguring the switch to the cross state.

F. Packet Switches

The switches considered so far in this section are relational switches that establish mappings between input and output ports that depend on the state of the switch, which is controlled by external control signals that are not dependent on the data entering the input ports. This type of switching is called **circuit switching**.

Another type of switch, called a **packet switch**, sets the switch configuration in accordance with destination information contained in the input data itself. As illustrated in Fig. 24.3-38, the data are organized in packets, each with a *header* containing the address of the packet's destination and a *payload*. The packet switch contains a header recognition unit that reads the address and sends a control signal that configures the switch appropriately.

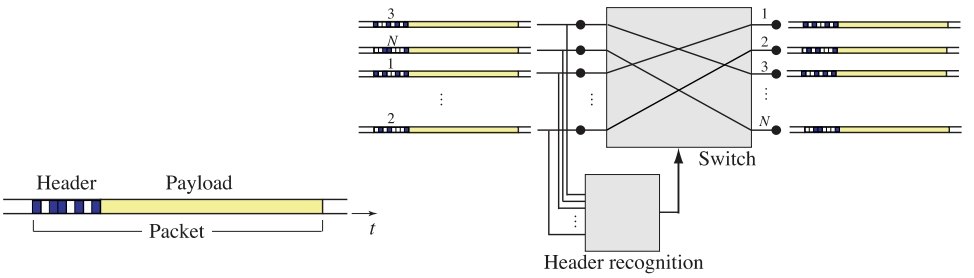


Figure 24.3-38 Packets and packet switches.

A header-address recognition system may use a bank of correlators that correlate the bit sequence representing the address of the incoming packet with the bit sequences representing each of the possible addresses in a lookup table, and identifies the address with the highest correlation. For example, if the address of the incoming

packet is the bit sequence $(a_1, a_2, \ldots, a_N)$ and that of one of the addresses in the table is $(b_1, b_2, \ldots, b_N)$, the correlation is the sum $a_1b_1 + a_2b_2 + \cdots + a_Nb_N$. Since the bits of the incoming header arrive sequentially in time, implementation of the correlation operation requires the use of delays, multipliers, and an adder. One optical implementation uses an optical fiber with N fiber Bragg grating (FBG) reflectors placed at equal distances, as shown in Fig. 24.3-39. The reflectors have reflectances $(b_1, b_2, \ldots, b_N)$ and serve as the multipliers. The round-trip delays introduced by the fiber segments bring the bits of the incoming header into synchrony so that they add up to yield the correlation sum.

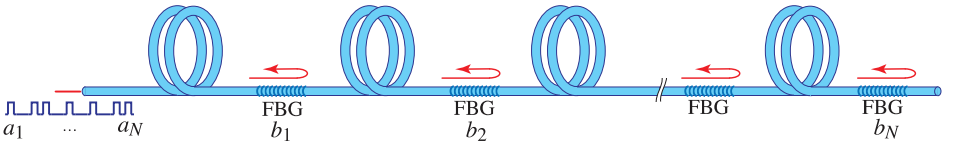


Figure 24.3-39 Optical correlator for recognizing the header address.

A packet switch may also be implemented sequentially by using a set of elementary 2×2 switches, each of which routes the incoming packet to its upper or lower output port depending on one bit in the header address. For example, if the bit is 1 or 0, the switch routes the data to the upper or lower output port, respectively. In other systems, the 2×2 switch sorts its two incoming packets and directs the packet with the greater address number to the lower output port and the other packet to the upper output port.

An example is provided by the 8×8 three-stage switching configuration illustrated in Fig. 24.3-40. This device, called a **Banyan switch**, employs twelve 2×2 self-routing switches. The address of each packet is expressed as a binary number (x_1, x_2, x_3). Routing in the first stage is based on the most significant bit x_1, while routing in stages 2 and 3 is based on bits x_2 and x_3, respectively. In each case, if the bit is 1, the packet is routed to the lower output port; otherwise, it goes to the upper output port. The switch is configured in such a way that after three stages, the packet arrives at its desired destination. However, it is not difficult to show that a conflict may arise when two packets are to be routed to the same output port of a 2×2 switch. More complex configurations, such as systems that make use of combinations of sorting and routing units, have been devised to avoid such internal blocking.

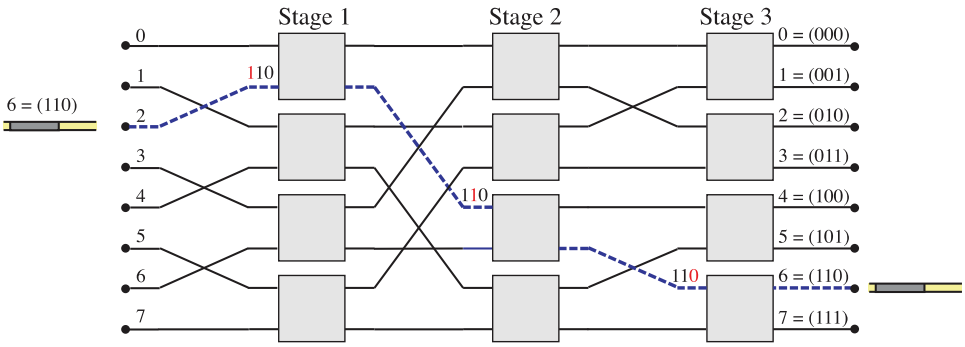


Figure 24.3-40 Configuration of an 8×8 three-stage Banyan switch. An incoming packet at input port number 2 with header address number 6 is directed to its destination, output port number 6, after passage through three 2×2 self-routing switches. Since the address is represented by the binary number 6 = (110), the packet is directed to the (lower, lower, upper) output ports of these switches, respectively, following the path indicated by the dashed lines, and ultimately reaches output port 6 = (110).

Contention occurs when packets from different input ports are simultaneously destined to the same output port. Methods for *contention resolution* include routing the conflicting packet via a different path or delaying it to a later time by using a buffer. In the optical domain, the packet may also be converted to a different wavelength and transmitted along a different wavelength channel. Wavelength converters and optical buffers were described in Secs. 24.3D and 24.3E, respectively.

24.4 PHOTONIC LOGIC GATES

Highly sophisticated digital electronic systems, such as digital computers, contain large numbers of interconnected basic units: switches, logic gates, and memory elements. As illustrated in Fig. 24.4-1, *all-optical logic gates* may be implemented by making use of the all-optical switches described in Sec. 24.3C. For example, the AND logic operation may be implemented by making use of a 1×1 switch in which the input and control signals (A and B, respectively) represent the two input bits of the gate, while the output signal (C) represents the output bit [Fig. 24.4-1(a)]. The AND gate may also be implemented by using two 1×1 switches connected in series, with the control signals for the two switches serving as the input bits of the gate [Fig. 24.4-1(b)]. Similarly, two 1×1 switches connected in parallel implement the OR gate [Fig. 24.4-1(c)], while a cascade of a 1×2 switch and a 2×1 switch implements the XOR (exclusive OR) gate [Fig. 24.4-1(d)].

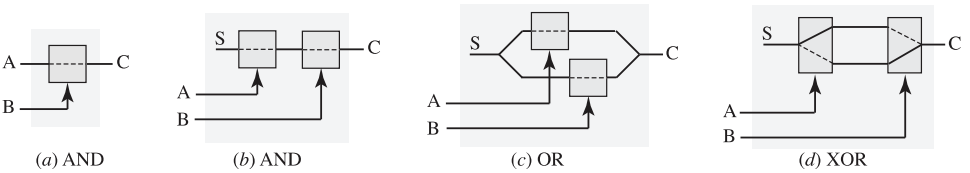


Figure 24.4-1 Implementation of various logic gates using switches. The input bits are denoted A and B and the output bit is denoted C. The source bit, denoted S, is in the "1" state.

While digital systems operate using *binary* (ON–OFF) *signals* (bits), the nonlinear optical interactions underlying the operation of all-optical switches do not necessarily adhere to this format. Binary signaling enables fan-out, input–output isolation, and the cascadability of logic operations. What is required, in essence, is an "optical transistor." In this section, we introduce the basic operating principles of bistable (flip-flop) optical systems and devices that find use in certain digital optical systems.

A. Bistable Systems

A **bistable system** is a two-state system whose output can assume only one of two distinct stable values, whatever input is applied. Switching between these two values may be achieved by changing the input level in particular ways. In the system illustrated in Fig. 24.4-2, for example, the output is seen to assume its low value for small input levels and its high value for large input levels. An input signal that starts at a low level and increases causes the output to suddenly jump from its low to its high value when the input exceeds a certain critical level, the *threshold* ϑ_2. If this large input signal is subsequently decreased, the output will jump back to its low value when the input crosses a different threshold $\vartheta_1 (< \vartheta_2)$, so that the input–output relation forms a hysteresis loop.

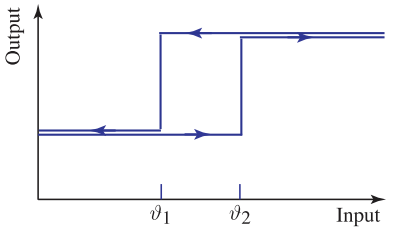


Figure 24.4-2 Input–output hysteresis relation for a bistable system.

There is an intermediate range of input levels, between ϑ_1 and ϑ_2, for which either low or high output values are possible, depending on the history of the input. Within this range, the system acts like a seesaw (Fig. 24.4-3). If the output is low, a large positive input spike will flip it to high. When in the high state, a large negative input spike will flip it to low. The system exhibits a flip-flop behavior; its state depends upon whether the last spike was positive or negative, i.e., on its history. Bistable devices can serve as switches, logic gates, and memory elements.

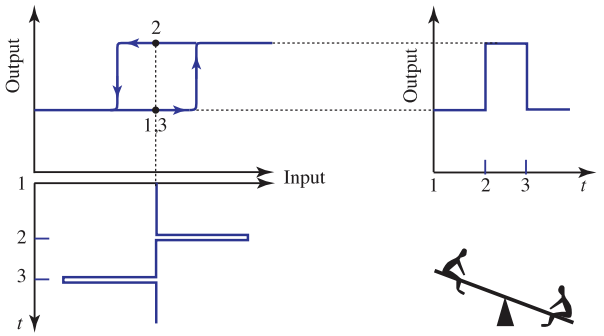


Figure 24.4-3 The flip-flop behavior of a bistable system. At time 1 the output is low. A positive input pulse at time 2 flips the output from low to high. The output remains at its high value until a negative pulse at time 3 flips it back to its low value. The system acts as a latching switch or a memory element.

The bistable device parameters may be adjusted so that the two critical values (the thresholds ϑ_1 and ϑ_2) coalesce into a single value ϑ. The result is then a single-threshold steep $\int$-shaped nonlinear input–output relation, as shown in Fig. 24.4-4.

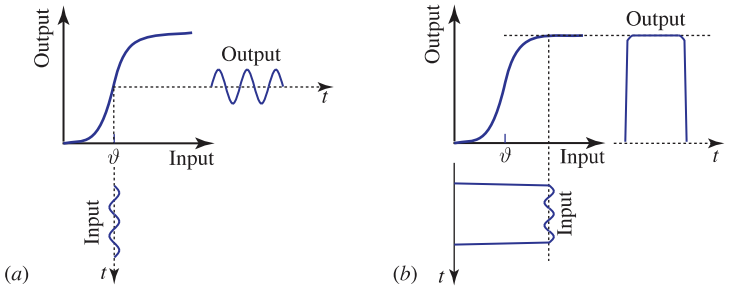


Figure 24.4-4 The bistable device as: (a) An amplifier; (b) A thresholding device, pulse shaper, or limiter.

When biased appropriately, the device can exhibit large differential gain and can be used as an amplifier [Fig. 24.4-4(a)], much like a transistor. It can also be used as a

thresholding element for which the output switches between two values as the input exceeds a threshold; or it can be used as a pulse shaper or limiter [Fig. 24.4-4(*b*)]. Stable threshold and stable bias are required to carry out these operations.

Bistable devices may also be used as logic elements. The binary data are represented by pulses that are added and their sum is used as the input. With an appropriate choice of pulse heights in relation to the threshold (Fig. 24.4-5), the device can be made to switch to its high value only when both pulses are present, for example, in which case it behaves as an AND gate.

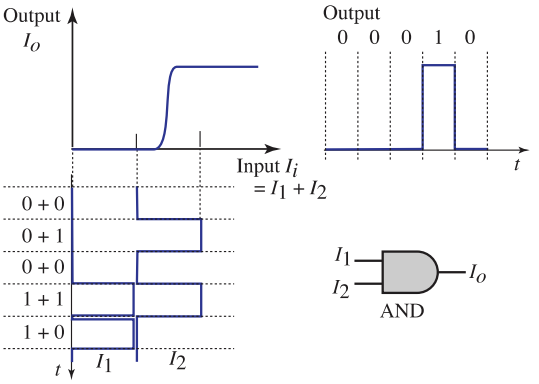


Figure 24.4-5 The bistable device as an AND logic gate. The input I_i is $I_1 + I_2$, where I_1 and I_2 are pulses representing the binary data. The output I_o is high if and only if both inputs are present.

B. Principles of Optical Bistability

Two features are required for the operation of a bistable device: *nonlinearity* and *feedback*. An *electronic* bistable (flip-flop) circuit is made by connecting the output of each of two transistors to the input of the other, as may be understood by consulting a textbook on digital electronics. An *optical* bistable system is realized by making use of a nonlinear optical element whose output beam is used in a feedback configuration to control the transmission of light through the element itself.

Consider the generic optical system illustrated in Fig. 24.4-6. The output intensity (or power) I_o is made to control the transmittance $\mathcal{T}$ of the system by means of feedback, so that $\mathcal{T}$ is some nonlinear function $\mathcal{T} = \mathcal{T}(I_o)$. Since $I_o = \mathcal{T} I_i$, we have

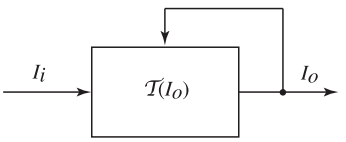


Figure 24.4-6 An optical system whose transmittance $\mathcal{T}$ is a function of its output I_o.

$$I_i = \frac{I_o}{\mathcal{T}(I_o)}. \tag{24.4-1}$$

Input–Output Relation for a Bistable System

If $\mathcal{T}(I_o)$ is a nonmonotonic function of I_o, such as the bell-shaped function used for purposes of illustration in Fig. 24.4-7(*a*), then $I_i = I_o/\mathcal{T}(I_o)$ will also be a nonmonotonic function of I_o, as shown in Fig. 24.4-7(*b*). Consequently, I_o must be a multivalued function of I_i, indicating that some values of I_i will have more than one corresponding value of I_o, as portrayed in Fig. 24.4-7(*c*). The system therefore exhibits bistable behavior, as schematized earlier in Fig. 24.4-2.

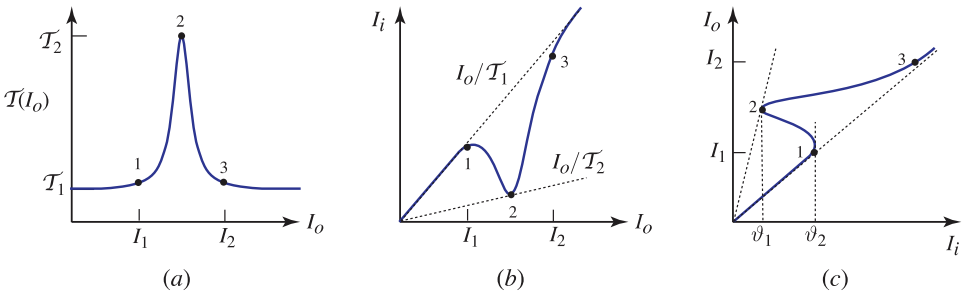


Figure 24.4-7 (a) Transmittance $\mathcal{T}(I_o)$ versus output I_o. This nonmonotonic function is chosen for illustrative purposes. (b) Input $I_i = I_o/\mathcal{T}(I_o)$ versus output I_o. For $I_o < I_1$ and $I_o > I_2$, $\mathcal{T}(I_o) = \mathcal{T}_1$ and $I_i = I_o/\mathcal{T}_1$, which is a linear relation of slope $1/\mathcal{T}_1$. At the particular intermediate value of I_o where $\mathcal{T}$ has its maximum value $\mathcal{T}_2$ (point 2), I_i dips below the line $I_i = I_o/\mathcal{T}_1$ and touches the lower line $I_i = I_o/\mathcal{T}_2$. (c) Output I_o versus input I_i. This curve is obtained by replotting the curve in (b) with the axes exchanged. This is achieved by rotating the diagram in (b) by 90° in a counterclockwise direction and forming a mirror image about the vertical axis.

The plot of I_o vs. I_i displayed in Fig. 24.4-7(c) is shown in more detail in Fig. 24.4-8. For small inputs $(I_i < \vartheta_1)$ or large inputs $(I_i > \vartheta_2)$, each input level has only a single corresponding output value I_o. Increasing the input level from a small value results in the output jumping to its high value when the threshold reaches ϑ_2. When the input is subsequently decreased, the output follows its upper branch until it reaches ϑ_1, at which point it jumps to its low value.

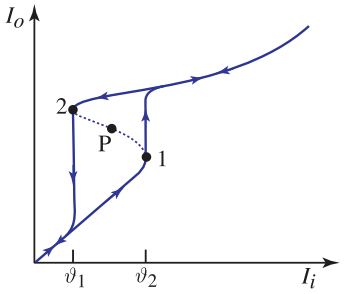


Figure 24.4-8 Expanded view of Fig. 24.4-7(c). The dotted curve that contains point P represents an unstable state, as explained in the text.

In the intermediate range $\vartheta_1 < I_i < \vartheta_2$, however, each input level corresponds to three possible output values. The upper and lower values are stable, but the intermediate output value (shown as the dotted curve) is unstable since any slight perturbation added to the input will force the output to move to either the upper or the lower branch. This may be understood by considering point P on the dotted curve joining points 1 and 2. A small increase in the output I_o will cause a sharp increase in the transmittance $\mathcal{T}(I_o)$ since the slope of $\mathcal{T}(I_o)$ is positive and large, as is evident from Fig. 24.4-7(a). This increase in transmittance results in a further increase in I_o, which serves to increase $\mathcal{T}(I_o)$ yet further. The net result is a transition to the upper stable state at point 2. Similarly, a small decrease in I_o at point P will cause a transition to the lower stable state at point 1.

The nonlinear bell-shaped function $\mathcal{T}(I_o)$ displayed in Fig. 24.4-7(a) was constructed for purposes of illustration. Many other nonlinear transmittance functions, coupled with systems patterned on Fig. 24.4-6, lead to bistability, and sometimes to *multistability* (where more than two stable values of the output exist for a single value of the input).

EXERCISE 24.4-1

Nonlinear Transmittance Functions That Exhibit Bistability. Plot the relation between I_o and $I_i = I_o/\mathcal{T}(I_o)$ for each of the following functions:

(a) $\mathcal{T}(x) = 1/[(x-1)^2 + a^2]$.
(b) $\mathcal{T}(x) = 1/[1 + a^2 \sin^2(x+\theta)]$.
(c) $\mathcal{T}(x) = \frac{1}{2} + \frac{1}{2}\cos(x+\theta)$.
(d) $\mathcal{T}(x) = \text{sinc}^2\sqrt{a^2+x^2}$.
(e) $\mathcal{T}(x) = (x+1)^2/(x+a)^2$.

Select appropriate values for the constants a and θ to generate a bistable relation. Some of these functions represent bistable systems that will be discussed subsequently.

Embedded bistable systems. The foregoing analysis dealt with a system whose transfer function $\mathcal{T}(I_o)$ is a function of its own output. In practice, however, a nonlinear element embedded within an optical system is illuminated not only by a portion of the output but also by a portion of the input, as depicted in Fig. 24.4-9. In this configuration, the open-loop system is described by a transfer function $\mathcal{T}(I)$ that depends on the light intensity I illuminating the nonlinear element, which is the sum of a component proportional to I_t and another component proportional to I_i, namely $I = \mathcal{T}_i I_i + \mathcal{R}_o I_t$. The transfer function $\mathcal{T}(I)$ relates the transmitted intensity to the input intensity via $I_t = \mathcal{T}(I) I_i$. The output intensity of the closed-loop system is $I_o = \mathcal{T}_o I_t$, where $\mathcal{T}_o$ is a transmittance factor. Combining these relations leads to the following two equations:

$$I_i = \frac{I}{\mathcal{T}_i + \mathcal{R}_o \mathcal{T}(I)} \quad \text{and} \quad I_o = \frac{\mathcal{T}_o}{\mathcal{R}_o}\left(I - \mathcal{T}_i I_i\right). \tag{24.4-2}$$

The first equation is a nonlinear relation analogous to (24.4-1). It can be similarly inverted to obtain I as a function of I_i and exhibits bistability if $\mathcal{T}(I)$ is a nonmonotonic function. The second equation is a linear relation that gives the final output I_o in terms of I and I_i.

The details of the function $\mathcal{T}(I)$, along with the constants $\mathcal{T}_i$, $\mathcal{T}_o$, and $\mathcal{R}_o$, are determined by the embedded bistable system under study, as will be understood from the devices considered in Sec. 24.4C.

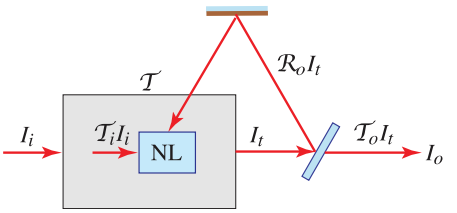


Figure 24.4-9 An optical system with an embedded nonlinear element NL and a feedback loop that directs a portion of the output power to NL. A portion of the input power also illuminates NL.

Intrinsic bistable systems. The feedback required for bistability can also be provided entirely internally. The system shown in Fig. 24.4-10, for example, consists of a resonator containing an optical nonlinear medium whose transmittance $\mathcal{T}(I)$ is controlled solely by the internal light intensity I within the resonator, rather than by the output light intensity I_o. Since $I_o = \mathcal{T}_o I$, where $\mathcal{T}_o$ is the transmittance of the resonator output mirror, the action of the internal intensity I is the same as that of the external intensity I_o, except for a constant factor. Examples of intrinsic bistable optical systems are provided in Sec. 24.4C.

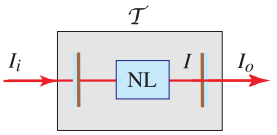


Figure 24.4-10 Intrinsic bistable device. The internal light intensity I controls the nonlinear medium and therefore the overall transmittance of the system $\mathcal{T}(I)$.

C. Bistable Optical Devices

Many schemes are available for implementing the foregoing bistability principles in optical configurations. The principal types of nonlinear optical elements that can serve in this capacity include:

- **Dispersive nonlinear elements**, for which the refractive index n is a function of the optical intensity. A medium exhibiting the optical Kerr effect, for example, has a refractive index $n(I) = n_0 + n_2 I$, where I is the intensity and n_0 and n_2 are constants, as discussed in Sec. 22.3A.
- **Dissipative nonlinear elements**, for which the absorption coefficient α is a function of the optical intensity. The saturable absorber discussed in Sec. 15.4A is an example in which the absorption coefficient $\alpha(I) = \alpha_0/(1 + I/I_s)$ is a nonlinear function of I, where α_0 is the small-signal absorption coefficient and I_s is the saturation intensity.
- **Amplifying nonlinear elements**, in which the gain coefficient γ is a function of the optical intensity. An example is a medium with saturable gain $\gamma(I) = \gamma_0/(1 + I/I_s)$, as considered in Sec. 15.4A.
- **Nonlinear elements with combined dispersion and dissipation/gain**. Intensity-dependent attenuation or amplification is often combined with an intensity-dependent refractive index. In a semiconductor optical amplifier (SOA), for example, increasing the optical intensity depletes the carrier density, which reduces the gain coefficient and also alters the refractive index.

Examples of optical configurations that include a nonlinear element and feedback include the following:

Optical Kerr Medium in a Mach–Zehnder interferometer. As discussed in Sec. 2.5A, the transmittance of a Mach–Zehnder interferometer (MZI) is a function of the phase difference φ between its branches: $\mathcal{T} = \frac{1}{2} + \frac{1}{2}\cos\varphi$. If an optical Kerr medium of refractive index $n(I) = n_0 + n_2 I$ is placed in one branch, as illustrated in Fig. 24.4-11, then $\varphi = k_o d n + \varphi_0 = k_o d n_2 I + \varphi_1$, where d is the length of the nonlinear medium; $k_o = 2\pi/\lambda_o$ where λ_o is the free-space wavelength; and φ_0 and φ_1 are constants. The transmittance of the system then depends on intensity as

$$\mathcal{T}(I) = \tfrac{1}{2} + \tfrac{1}{2}\cos\left(k_o d n_2 I + \varphi_1\right). \tag{24.4-3}$$

As Fig. 24.4-11 illustrates, this function is a periodic repetition of a bell-shaped function of intensity, such as that used earlier to demonstrate bistability [Fig. 24.4-7(*a*)].

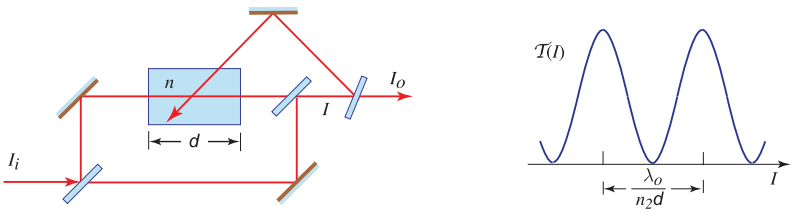


Figure 24.4-11 A Mach–Zehnder interferometer in which one branch contains a nonlinear medium of refractive index $n(I)$ controlled by the reflected intensity I via the optical Kerr effect.

Optical Kerr medium in a Fabry–Perot resonator. The transmittance of a Fabry–Perot resonator is expressible as $\mathcal{T} = \mathcal{T}_{\max}[1 + (2\mathcal{F}/\pi)^2 \sin^2(\varphi/2)]^{-1}$, where φ is the round-trip phase shift and $\mathcal{T}_{\max}$ and $\mathcal{F}$ are constants [see (2.5-18)]. For an optical Kerr medium of length d placed between the mirrors [Fig. 24.4-12(a)], we again have $\varphi = k_o d n + \varphi_0 = k_o d n_2 I + \varphi_1$, where φ_0 and φ_1 are constants. The transmittance of the system then depends on the internal intensity I as

$$\mathcal{T}(I) = \frac{\mathcal{T}_{\max}}{1 + (2\mathcal{F}/\pi)^2 \sin^2(k_o d n_2 I + \varphi_1)} . \tag{24.4-4}$$

As illustrated in Fig. 24.4-12(b), this function consists of a periodic sequence of sharply peaked, bell-shaped functions, and the system is bistable.

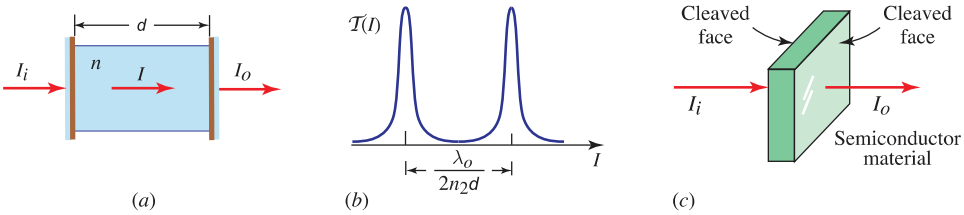


Figure 24.4-12 (a) A Fabry–Perot resonator containing a medium of nonlinear refractive index $n(I)$ that is controlled by the internal light intensity I. (b) Intensity-dependent transmittance $\mathcal{T}(I)$. (c) Bistable Fabry–Perot etalon made of a thin layer of semiconductor material with parallel reflecting surfaces.

This configuration has been used to demonstrate optical bistability in a number of materials (e.g., sodium vapor, carbon disulfide, and nitrobenzene). Since the nonlinear refractive index n_2 is small for these materials, however, the effect is hard to observe. Semiconductors such as GaAs are more suitable since they exhibit far stronger optical nonlinearities; in fact, a bistable device may be fabricated from a thin layer of semiconductor material whose cleaved faces serve as mirrors, as portrayed in Fig. 24.4-12(c). Multiquantum-well semiconductor structures (Secs. 17.1G and 18.2D) also exhibit bistability, as do organic materials, graphene, nonlinear photonic crystals, and surface plasmon polaritons.

Saturable absorber in a Fabry–Perot resonator. A saturable absorber placed inside a Fabry–Perot resonator of length d that is tuned for peak transmission (Fig. 24.4-13) has transmittance

$$\mathcal{T} = \frac{\mathcal{T}_1}{(1 - \mathcal{R}e^{-\alpha d})^2} , \tag{24.4-5}$$

where $\mathcal{R} = \sqrt{\mathcal{R}_1 \mathcal{R}_2}$; $\mathcal{R}_1$ and $\mathcal{R}_2$ are the mirror reflectances; and $\mathcal{T}_1$ is a constant (see Secs. 2.5B and 11.1A).

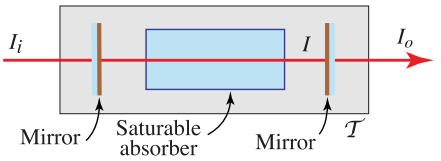


Figure 24.4-13 A bistable device consisting of a saturable absorber in a Fabry–Perot resonator.

If the medium is optically thin so that $\alpha d \ll 1$, a Taylor-series expansion provides $e^{-\alpha d} \approx 1 - \alpha d$, whereupon

$$\mathcal{T} \approx \frac{\mathcal{T}_1}{[1-(1-\alpha d)\mathcal{R}]^2}. \tag{24.4-6}$$

Because α is a nonlinear function of I, $\mathcal{T}$ is also a nonlinear function of I. Using the relations $I = I_o/\mathcal{T}_o$ and $\alpha(I) = \alpha_0/(1 + I/I_s)$, together with (24.4-6), we arrive at

$$\mathcal{T}(I_o) = \mathcal{T}_2 \left[\frac{I_o + I_{s1}}{I_o + (1+a)I_{s1}}\right]^2, \tag{24.4-7}$$

where $\mathcal{T}_2 = \mathcal{T}_1/(1-\mathcal{R})^2$, $a = \alpha_0 d\mathcal{R}/(1-\mathcal{R})$, and $I_{s1} = I_s\mathcal{T}_0$. For certain values of a, the system is bistable [see example (e) in Exercise 24.4-1].

Nonlinear amplifier in a Fabry–Perot resonator. Suppose now that the saturable absorber considered above is replaced by an amplifying medium with saturable gain. The system is then nothing but an optical amplifier with feedback, i.e., a laser. If $\mathcal{R}\exp(\gamma_0 d) < 1$, the laser is below threshold. But if $\mathcal{R}\exp(\gamma_0 d) > 1$, the system becomes unstable and laser oscillation ensues. Though the topic of laser bistability was not discussed in Chapters 16 and 18, lasers do indeed exhibit bistable behavior under certain circumstances. In some sense, the dispersive bistable optical system is the nonlinear-refractive-index analog of the nonlinear-gain laser.

SOA-MZI bistable device. A bistable system that makes use of a pair of identical semiconductor optical amplifiers (SOA 1 and SOA 2) in a Mach–Zehnder interferometer (MZI) configuration is illustrated in Fig. 24.4-14. The SOAs are placed in the two branches of the MZI and the output of the MZI is fed back to SOA 1 via a feedback loop implemented by optical fibers and couplers. The saturated gain of the SOA is a decreasing function of the optical power and the phase is also power-dependent, so gain and dispersive effects both come into play.

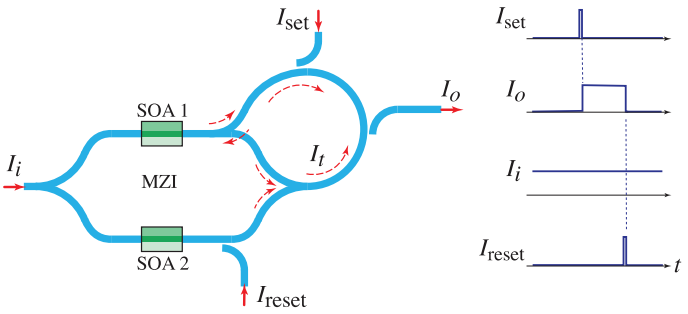


Figure 24.4-14 A bistable device that makes use of two semiconductor optical amplifiers, SOA 1 and SOA 2, in the two branches of a Mach–Zehnder interferometer (MZI). The output of the MZI is fed back into SOA 1. The "set" and "reset" short pulses initiate transitions to states of high and low output power I_o, respectively. The input I_i has constant power and serves as a bias.

With light of constant power I_i presented to the input, the MZI is balanced and the optical power I_t at its output is zero, as is the output I_o of the overall system. This stable condition is the lower state of the bistable system. Another stable state is that for which the optical power in SOA 1 is high so that its gain is depleted and its phase differs from that of SOA 2. The MZI is then unbalanced and the output I_o is high.

Flipping between the two states of this system is accomplished by making use of external triggers. When the system is in the low state, it may be switched to the

high state by injecting a short pulse I_{set} into SOA 1 through the feedback loop. This reduces the gain of SOA 1 and alters the phase shift introduced by this laser. As a consequence, the MZI becomes unbalanced and its transmittance increases. A portion of the input optical power I_i now reaches the feedback loop and is coupled back into SOA 1, keeping it in the depleted state even after the termination of the set pulse. An unbalanced MZI and a stable state of high output power results. The system may be flipped back to its lower state by injecting a pulse I_{reset} directly into SOA 2, thereby reducing its gain. Power also reaches SOA 1 via the feedback loop. A new balanced MZI condition is attained for which the output power $I_o = 0$, and the system remains in the low state.

Coupled microring-laser bistable device. A microring laser (Sec. 18.5B) has two independent (uncoupled) lasing modes, one that propagates clockwise (CW) and the other that propagates counterclockwise (CCW). Two such lasers with close resonant frequencies may be connected via a waveguide such that light from the CW mode of laser A is coupled to the CW mode of laser B, or light from the CCW mode of laser B is coupled to the CCW of laser A (Fig. 24.4-15). This type of mutual feedback gives rise to a bistable system with two stable "master–slave" states in which laser A is the master and laser B is the slave, or vice-versa. The first state emerges if more light from the CW mode of laser A is coupled into the laser B ring. As it undergoes resonant amplification, it injection-locks the CW mode of laser B whereupon self-oscillation in laser B is extinguished and so is its CCW mode. The CW mode then acquires greater power since its pump energy is supporting only one mode [Fig. 24.4-15(a)]. The second state is established when the two lasers reverse roles; the CCW mode of laser A is then injection-locked and acquires greater power, while its CW mode is suppressed [Fig. 24.4-15(b)].

The system is set into the first stable state (higher power in the laser B CW mode) by injecting a "set" optical pulse that favors the CW mode of laser B [Fig. 24.4-15(a)]. The system may then be flipped into the second stable state (higher power in the laser A CCW mode) by injecting a "reset" optical pulse that strengthens the CCW mode of laser A [Fig. 24.4-15(b)]. A coupled microring-laser flip-flop implemented in the form of an InP/InGaAsP photonic integrated circuit (PIC) exhibits a switching time of 20 ps and a switching energy of 5.5 fJ.[†]

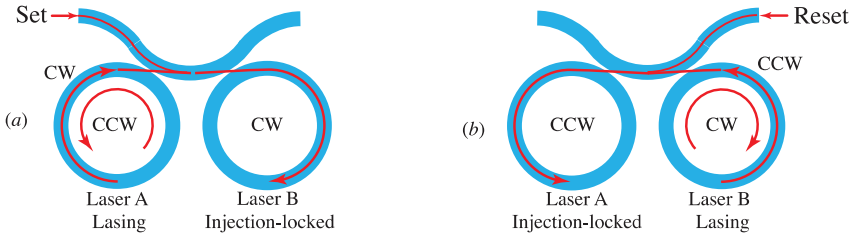


Figure 24.4-15 A bistable device using two microring lasers connected via a waveguide. The clockwise (CW) modes of the two lasers are coupled, and so too are the counter-clockwise (CCW) modes. (a) The "set" optical pulse initiates a state for which laser A acts as a master that injection-locks laser B and suppresses its CCW mode. (b) The "reset" optical pulse initiates a state for which laser B is a master that injection-locks laser A and suppresses its CW mode.

[†] See M. T. Hill, H. J. S. Dorren, T. de Vries, X. J. M. Leijtens, J. H. den Besten, B. Smalbrugge, Y.-S. Oei, H. Binsma, G.-D. Khoe, and M. K. Smit, A Fast Low-Power Optical Memory Based on Coupled Micro-Ring Lasers, *Nature*, vol. 432, pp. 206–209, 2004.

Hybrid Bistable Optical Devices

All of the bistable optical systems discussed thus far are all-optical systems. Hybrid electrical/optical bistable systems that involve electric fields have also been devised. In one example, a Pockels cell is placed inside a Fabry–Perot resonator; the output light is detected using a photodetector, and a voltage proportional to the detected optical intensity is applied to the cell so that its refractive index varies in proportion to the output intensity. The optical transmittance of the resonator is consequently a nonlinear function of the output optical intensity, and since feedback is provided by the resonator, the prerequisites for bistable behavior are present. A related example makes use of a Pockels-cell wave retarder placed between crossed polarizers. Again the output light intensity is detected and a proportional voltage is applied to the cell. The transmittance of the modulator is then a nonmonotonic function of the transmitted intensity, and the system is bistable. These systems are readily implemented in integrated-photonic technology.

Spatial light modulators (SLMs) may be used to construct arrays of bistable elements. In an optically addressed liquid-crystal SLM (Sec. 21.3B), for example, the reflectance of each element is a nonlinear function of the intensity of the light illuminating its write side. By using feedback, the write intensity is proportional to the intensity of the beam reflected from the element itself, so that bistable behavior is exhibited. Different points on the surface of the device can be addressed separately so that the SLM serves as an array of bistable optical elements.

The **self-electro-optic-effect device** (**SEED**) is an electro-optic semiconductor device that exhibits bistability. The SEED consists of a p–i–n photodiode in which the intrinsic region comprises a semiconductor multiquantum-well (MQW) heterostructure. The diode is reverse-biased and a large electric field is created in the MQW. By virtue of the quantum-confined Stark effect (QCSE) (Sec. 21.5), the optical absorption coefficient is a nonlinear function of the voltage across the MQW. Consequently, the optical transmittance is a nonlinear function of that voltage. Bistable behavior is exhibited in the SEED as a result of the feedback mechanism introduced by the photodiode electrical circuit, which renders the voltage dependent on the incident optical power. This occurs since the absorbed light creates a proportional photocurrent that flows into the external circuit, resulting in a voltage drop. SEED devices can be fabricated in the form of arrays that operate at moderately high speeds and at low powers.

READING LIST

Optical Interconnects

F. Chang, ed., *Datacenter Connectivity Technologies: Principles and Practice*, River Publishers, 2018.

T. Tekin, R. Pitwon, A. Håkansson, and N. Pleros, eds., *Optical Interconnects for Data Centers*, Elsevier-Woodhead, 2017.

Special issue on optical interconnects, *Journal of Lightwave Technology*, vol. 34, no. 12, 2016.

A. Siokis, K. Christodoulopoulos, and E. Varvarigos, Multipoint Architectures for On-Board Optical Interconnects, *Journal of Optical Communications and Networking*, vol. 8, pp. 863–877, 2016.

N. Bamiedakis, K. A. Williams, R. V. Penty, and I. H. White, Integrated and Hybrid Photonics for High-Performance Interconnects, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

J. Orcutt, R. Ram, and V. Stojanović, CMOS Photonics for High Performance Interconnects, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

C. Kachris, K. Bergman, and I. Tomkos, eds., *Optical Interconnects for Future Data Center Networks*, Springer-Verlag, 2013.

M. A. Taubenblatt, Optical Interconnects for High-Performance Computing, *Journal of Lightwave Technology*, vol. 30, pp. 448–458, 2012.

C. Kachris and I. Tomkos, A Survey on Optical Interconnects for Data Centers, *IEEE Communications Surveys & Tutorials*, vol. 14, pp. 1021–1036, 2012.

M. J. R. Heck, H.-W. Chen, A. W. Fang, B. R. Koch, D. Liang, H. Park, M. N. Sysak, and J. E. Bowers, Hybrid Silicon Photonics for Optical Interconnects, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, pp. 333–346, 2011.

D. A. B. Miller, Device Requirements for Optical Interconnects to Silicon Chips, *Proceedings of the IEEE*, vol. 97, pp. 1166–1185, 2009.

M. Haurylau, G. Chen, H. Chen, J. Zhang, N. A. Nelson, D. H. Albonesi, E. G. Friedman, and P. M. Fauchet, On-Chip Optical Interconnect Roadmap: Challenges and Critical Directions, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 12, pp. 1699–1705, 2006.

L. Pavesi and G. Guillot, eds., *Optical Interconnects: The Silicon Approach*, Springer-Verlag, 2006.

S. Kawai, ed., *Handbook of Optical Interconnects*, CRC Press/Taylor & Francis, 2005.

G. A. Keeler, B. E. Nelson, D. Agarwal, C. Debaes, N. C. Helman, A. Bhatnagar, and D. A. B. Miller, The Benefits of Ultrashort Optical Pulses in Optically Interconnected Systems, *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 9, pp. 477–485, 2003.

D. A. B. Miller, Rationale and Challenges for Optical Interconnects to Electronic Chips, *Proceedings of the IEEE*, vol. 88, pp. 728–749, 2000.

S. H. Lee, ed., *Selected Papers on Optical Interconnects and Packaging*, SPIE Optical Engineering Press (Milestone Series Volume 142), 1998.

J. W. Goodman, F. I. Leonberger, S. Y. Kung, and R. A. Athale, Optical Interconnections for VLSI Systems, *Proceedings of the IEEE*, vol. 72, pp. 850–866, 1984.

Photonic Routers, Switches, and Gates

See also the reading lists in Chapters 9, 10, 20, 21, and 22.

F. Testa and L. Pavesi, eds., *Optical Switching in Next Generation Data Centers*, Springer-Verlag, 2018.

Z. Chai, X. Hu, F. Wang, X. Niu, J. Xie, and Q. Gong, Ultrafast All-Optical Switching, *Advanced Optical Materials*, vol. 5, 1600665, 2017.

H. Venghaus and N. Grote, eds., *Fibre Optic Communication: Key Devices*, Springer-Verlag, 2nd ed. 2017.

P. R. Prucnal and B. J. Shastri, *Neuromorphic Photonics*, CRC Press/Taylor & Francis, 2017.

E. Cohen, S. Dolev, and M. Rosenblit, All-Optical Design for Inherently Energy-Conserving Reversible Gates and Circuits, *Nature Communications* **7**, 11424 doi: 10.1038/ncomms11424, 2016.

P. J. Winzer, C. J. Chang-Hasnain, A. E. Willner, R. C. Alferness, R. W. Tkach, and T. G. Giallorenzi, eds., *A Third of a Century of Lightwave Technology: January 1983–April 2016*, IEEE–OSA, 2016.

S. Wabnitz and B. J. Eggleton, eds., *All-Optical Signal Processing: Data Communication and Storage Applications*, Springer-Verlag, 2015.

M. D. Feuer and S. L. Woodward, ROADMs: Reconfigurable Optics for Agile Networks, *Optics & Photonics News*, vol. 26, no. 3, pp. 36–43, 2015.

K. Bergman, L. P. Carloni, A. Biberman, J. Chan, and G. Hendry, *Photonic Network-on-Chip Design*, Springer-Verlag, 2014.

L. Thylen, P. Holmström, L. Wosinski, B. Jaskorzynska, M. Naruse, T. Kawazoe, M. Ohtsu, M. Yan, M. Fiorentino, and U. Westergren, Nanophotonics for Low-Power Switches, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

S. L. Woodward, M. D. Feuer, and P. Palacharla, ROADM-Node Architectures for Reconfigurable Photonic Networks, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-B: Systems and Networks*, Academic Press/Elsevier, 6th ed. 2013.

S. L. Woodward and M. D. Feuer, Benefits and Requirements of Flexible-Grid ROADMs and Networks, *Journal of Optical Communications and Networking*, vol. 5, pp. A19–A27, 2013.

L. K. Oxenløwe, A. Clausen, M. Galili, H. C. H. Mulvad, H. Ji, H. Hu, and E. Palushani, Ultra-High-Speed Optical Time Division Multiplexing, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

S. Frisken, I. Clarke, and S. Poole, Technology and Applications of Liquid Crystal on Silicon (LCoS) in Telecommunications, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

K. Ishii, S. Mitsui, H. Hasegawa, K. Sato, S. Kamei, M. Okuno, and H. Takahashi, Development of Hierarchical Optical Path Cross-Connect Systems Employing Wavelength/Waveband Selective Switches, *Journal of Optical Communications and Networking*, vol. 3, pp. 559–567, 2011.

B. Li and S. J. Chua, eds., *Optical Switches: Materials and Design*, Elsevier-Woodhead, 2010.

T. S. El-Bawab, *Optical Switching*, Springer-Verlag, 2006.

I. Glesk, B. C. Wang, L. Xu, V. Baby, and P. R. Prucnal, Ultra-Fast All-Optical Switching in Optical Networks, in *Progress in Optics*, vol. 45, pp. 53–117, E. Wolf, ed., Elsevier, 2003.

A. Marrakchi, ed., *Selected Papers on Photonic Switching*, SPIE Optical Engineering Press (Milestone Series Volume 121), 1996.

C. S. Tsai, Integrated Acoustooptic and Magnetooptic Devices for Optical Information Processing, *Proceedings of the IEEE*, vol. 84, pp. 853–869, 1996.

J. E. Midwinter, ed., *Photonics in Switching*, Volume 1, *Background and Components*, Academic Press/Elsevier, 1993.

J. E. Midwinter, ed., *Photonics in Switching*, Volume 2, *Systems*, Academic Press/Elsevier, 1993.

Bistable Optical Devices

A. Joshi and M. Xiao, *Controlling Steady-State and Dynamical Properties of Atomic Optical Bistability*, World Scientific, 2012.

L. N. Binh and D. Van Liet, eds., *Nonlinear Optical Systems: Principles, Phenomena, and Advanced Signal Processing*, CRC Press/Taylor & Francis, 2012.

H. Kawaguchi, *Bistabilities and Nonlinearities in Laser Diodes*, Artech, 1994.

D. A. B. Miller, D. S. Chemla, T. C. Damen, T. H. Wood, C. A. Burrus, Jr., A. C. Gossard, and W. Wiegmann, The Quantum Well Self-Electrooptic Effect Device: Optoelectronic Bistability and Oscillation, and Self-Linearized Modulation, *IEEE Journal of Quantum Electronics*, vol. 21, pp. 1462–1476, 1985.

H. M. Gibbs, *Optical Bistability: Controlling Light with Light*, Academic Press, 1985.

L. A. Lugiato, Theory of Optical Bistability, in *Progress in Optics*, vol. 21, pp. 69–216, E. Wolf, ed., North-Holland, 1984.

PROBLEMS

24.1-3 **Interconnection Hologram for a Conformal Map.** Design a hologram to realize the geometric transformation defined by

$$x' = \psi_x(x, y) = \ln \sqrt{x^2 + y^2}$$
$$y' = \psi_y(x, y) = \tan^{-1} \frac{y}{x}.$$

This is a Cartesian-to-polar transformation followed by a logarithmic transformation of the polar coordinate $r = \sqrt{x^2 + y^2}$. Determine an expression for the required phase function $\varphi(x, y)$ of the hologram.

24.2-1 **Cascaded MZI MUX/DMUX.** Three Mach–Zehnder interferometers (MZIs) are cascaded as shown in Fig. 24.2-6 to multiplex or demultiplex four wavelength channels with wavelength separation $\Delta\lambda = 0.2$ nm and central wavelength 1550 nm. Determine the pathlength differences Δd required for each interferometer if the refractive index is $n = 2.3$.

24.2-2 **AWG DMUX.** An arrayed waveguides (AWG) router (Fig. 24.2-8) is used to demultiplex four wavelength channels with wavelength separation $\Delta\lambda = 0.2$ nm and central wavelength 1550 nm. Determine the pathlength difference parameter Δd_b that must be introduced by the star coupler if its refractive index is $n = 2.3$.

24.2-3 **AWG as a 2×2 Wavelength Router.** An AWG device is configured as a 2×2 wavelength router. Input port 1 has two wavelength channels, λ_1 and λ_2, and input port 2 has two wavelength channels, λ_3 and λ_4. Design a router that transposes the input wavelengths among the two output ports, i.e., directs the λ_1 and λ_3 channels to output port 1, and the λ_2 and λ_4 channels to output port 2. Write the routing conditions in terms of the four optical pathlength differences Δd_{11}, Δd_{12}, Δd_{21}, and Δd_{22} of the multipath interferometers connecting each of the input ports to each of the output ports.

24.3-1 **Power Loss and Crosstalk.** A 4×4 switch may be implemented by use of five 2×2 switches. If each of these switches introduces a power loss of 0.5 dB and a crosstalk of -30 dB, determine the worst-case power loss and crosstalk for the 4×4 switch.

24.3-2 **MZI Crossbar Switch.** An electro-optic Mach–Zehnder interferometer is used as a crossbar switch. The application of a voltage $V = V_\pi$ to the electro-optic material in one arm of the interferometer introduces a phase shift of π. If the switch is set in the bar state when $V = 0$, what must the applied voltage V be to change the switch to the cross state? Determine the crosstalk (in dB) caused by a 1% error in that applied voltage.

24.3-3 **TSI Switch.** As shown in Fig. 24.3-36, the time-slot interchange (TSI) switch may be implemented by a five step process: time-to-space routing, time delays, space switching, time delays, and space-to-time routing. Construct another implementation using the programmable delay lines shown in Fig. 24.3-37.

24.4-2 **Photonic Logic Gate.** Figure 24.4-5 illustrates how a nonlinear thresholding optical device may be used to make an AND gate. Show how a similar system may be used to make NAND, OR, and NOR gates. Is it possible to make an XOR (exclusive OR) gate? Can the same system be used to obtain the OR of N binary inputs?

24.4-3 **Bistable Interferometer.** A crystal exhibiting the optical Kerr effect is placed in one of the arms of a Mach–Zehnder interferometer. The transmitted intensity I_o is fed back and illuminates the crystal. Show that the intensity transmittance of the system is $I_o/I_i = \mathcal{T}(I_o) = \frac{1}{2} + \frac{1}{2}\cos(\pi I_o/I_\pi + \varphi)$, where I_π and φ are constants. Assuming that $\varphi = 0$, sketch I_o versus I_i and derive an expression for the maximum differential gain dI_o/dI_i.

CHAPTER

25

OPTICAL FIBER COMMUNICATIONS

25.1 FIBER-OPTIC COMPONENTS 1226
- A. Optical Fibers
- B. Sources for Optical Transmitters
- C. Optical Amplifiers
- D. Photodetectors for Optical Receivers
- E. Photonic Integrated Circuits

25.2 OPTICAL FIBER COMMUNICATION SYSTEMS 1238
- A. Evolution of Optical Fiber Communication Systems
- B. Performance of Optical Fiber Communication Systems
- C. Attenuation- and Dispersion-Limited Systems
- D. Attenuation and Dispersion Compensation and Management
- E. Soliton Optical Communications

25.3 MODULATION AND MULTIPLEXING 1257
- A. Modulation
- B. Multiplexing
- C. Wavelength-Division Multiplexing
- D. Space-Division Multiplexing

25.4 COHERENT OPTICAL COMMUNICATIONS 1266

25.5 FIBER-OPTIC NETWORKS 1274
- A. Network Topologies and Multiple Access
- B. Wavelength-Division Multiplexing Networks

Sir Charles Kuen Kao (1933–2018) received the Nobel Prize in 2009 for recognizing that high-purity glass, such as fused silica, could make optical fiber communications a reality.

Ivan Paul Kaminow (1930–2013) made pioneering contributions to photonic components, systems, and networks that are widely used in optical fiber communications today.

Philip St John Russell (born 1953) invented the photonic-crystal fiber; various versions of such fibers have found use in many applications, including optical fiber communications.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.

©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Until the mid-1970s, virtually all communication systems relied on the transmission of information over electrical cables or made use of radio-frequency or microwave electromagnetic radiation propagating in free space. Light would seem to have been a more natural choice for communications since, unlike electricity and radio waves, it did not have to be discovered. However, low-loss conduits for carrying light were not available and obstructions such as clouds, fog, and haze hindered the passage of light through free space.

The invention of the laser in the early 1960s stirred interest in using light to communicate but early lasers were bulky, inefficient, and difficult to modulate. On the brighter side, suitable photodetectors were available. The advent of optical fiber communications has its roots in two critical events in the annals of photonics: the invention of compact and efficient semiconductor sources such as light-emitting diodes (LEDs) and laser diodes (LDs), and the development of low-loss optical fibers to carry light with minimal attenuation. The technology of optical fiber communications offers enormous transmission capacity, long link lengths between amplifiers, immunity from electromagnetic interference, information security, and relative ease of installation. Indeed, billions of kilometers of optical fiber have been deployed around the globe (the circumference of the earth is a mere 40 Mm).

The ever-increasing volume of data, voice, video, and telemetry transmitted over both short- and long-haul links is driven by the voracious human appetite for media content, social networking, internet applications, and cloud services. Optical fiber communications is the only technology that has been able to meet the vast and exponentially increasing demands of broadband communications, on scales that stretch from the individual dwelling to the globe. All-fiber local, metropolitan, regional, long-haul, and submarine networks interconnect organizations and cities, states and countries, and continents. It is a remarkable, and continuing, success story.

This Chapter

This chapter provides an introduction to optical fiber communication systems and fiber-optic networks. A point-to-point communication link comprises three basic elements, as illustrated in Fig. 25.0-1: a compact light source modulated by an electrical signal, a low-loss/low-dispersion optical fiber, and a photodetector that converts the optical signal back into an electrical signal. These optical components are discussed in detail in Chapters 18, 10, and 19, respectively. Optical amplifiers have also proved useful in many fiber systems and these devices are discussed in Chapter 15.

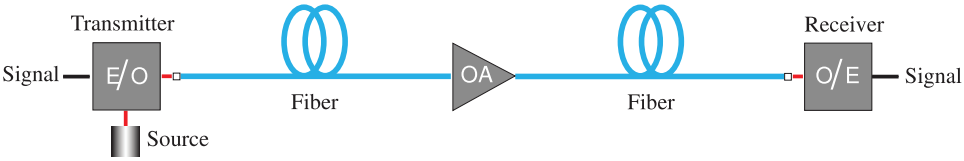


Figure 25.0-1 Schematic of an optical fiber communication system. An electrical signal is converted into an optical signal (E/O) by modulating an optical source. The optical signal is transmitted through the fiber to the receiver, where it is converted back into an electrical signal via a photodetector and demodulator (O/E). For long links, optical amplifiers (OAs) can be used to boost the attenuated optical signal.

To make the chapter self-contained, Sec. 25.1 provides a brief summary of the pertinent properties of fibers, sources, amplifiers, and detectors, and also examines their role in the context of the overall design, operation, and performance of an optical fiber

communication link. Other optical accessories such as splices, connectors, couplers, switches, and multiplexing devices are also essential for the successful operation of fiber links and networks; the principles underlying the operation of many of these devices are described in Chapter 24 and in other parts of this book.

Section 25.2 summarizes the evolution of optical fiber communication systems from a historical perspective and considers the basic design principles applicable to long-distance intensity-modulated digital and analog links. The maximum fiber span available for transmitting data, at a given rate and with a prescribed performance level, is determined. Performance deteriorates if the data rate exceeds the fiber bandwidth, or if the received power falls below the receiver sensitivity in which case the signal cannot be distinguished from noise.

This is followed, in Sec. 25.3, by an introduction to the various forms of modulation and multiplexing used in optical fiber systems, including field, intensity, and digital modulation. Multiplexing, which enables more than one signal to be sent on a single communication link, exists in many forms, including time-division, code-division, wavelength-division, and space-division versions thereof.

Digital coherent fiber communication systems, which are introduced in Sec. 25.4, use light not as a source of controllable power but rather as an electromagnetic wave of controllable amplitude, frequency, and/or phase. Coherent optical fiber systems are the natural extension to higher frequencies of conventional radio and microwave communication systems. Modern high-speed electronics, spectrally efficient coding, and digital signal processing have permitted coherent systems to provide unparalleled performance in receiver sensitivity and information-transmission capacity.

Fiber-optic networks are communication links controlled by a set of routers and switches that interconnect multiple users distributed over some geographic area (e.g., a local-area network or LAN). Section 25.5 provides an introduction to such networks, including wavelength-division multiplexed (WDM) versions.

25.1 FIBER-OPTIC COMPONENTS

A. Optical Fibers

An optical fiber is a cylindrical dielectric waveguide made of low-loss materials, usually fused silica glass (SiO_2) of high chemical purity. In its simplest configuration, called a **step-index fiber**, the core of the waveguide has a constant refractive index that is slightly higher than that of the cladding (the outer medium) so that light is guided by total internal reflection along the direction of the fiber axis.

The transmission of light through the fiber may be most simply understood by examining the trajectories of the guided *rays* within the core (Sec. 10.1). A more complete analysis, based on electromagnetic theory, teaches that light travels in the fiber in the form of guided *waves* (Sec. 10.2); each is a mode with a distinct spatial distribution, polarization, propagation constant, group velocity, and attenuation coefficient. There is, however, a correspondence between each mode and a ray that bounces within the core in a distinct trajectory.

The step-index fiber is characterized by its core radius a; the refractive indices of its core and cladding, n_1 and n_2, respectively; and the fractional refractive index difference $\Delta \approx (n_1 - n_2)/n_1$, which is usually very small ($0.001 \leq \Delta \leq 0.02$). Light rays making angles with the fiber axis that are smaller than the complement of the critical angle, $\bar{\theta}_c = \cos^{-1}(n_2/n_1)$, are guided within the core by multiple total internal reflections at the core–cladding boundary. The angle $\bar{\theta}_c$ in the fiber corresponds to an acceptance angle $\theta_a = \sin^{-1}(\text{NA})$ for rays incident from air into the fiber, where the

numerical aperture NA is given by

$$\text{NA} = \sin\theta_a = \sqrt{n_1^2 - n_2^2} \approx n_1\sqrt{2\Delta}\,. \tag{25.1-1}$$

Numerical Aperture

Multimode Fibers (MMFs)

Step-index fibers. The number of guided modes M supported by a **step-index multimode fiber** is governed by the fiber V parameter, $V = 2\pi(a/\lambda_o)\,\text{NA}$, where a/λ_o is the ratio of the core radius to the free-space wavelength (Sec. 10.2A). A fiber with $V \gg 1$ supports a large number of modes: $M \approx V^2/2$. Since the modes travel with different group velocities, this results in pulse spreading, which increases linearly with the fiber length, an effect known as **modal dispersion**. When an impulse of light travels a distance L in the fiber, it arrives as a sequence of pulses centered at the modal delay times, as illustrated in Fig. 25.1-1(a). The composite pulse has an approximate RMS width

$$\sigma_\tau \approx \frac{\Delta}{2c_1} L\,, \tag{25.1-2}$$

Response Time (Step-Index MMF)

where $c_1 = c_o/n_1$. For example, if $n_1 = 1.46$ and $\Delta = 0.01$, the time-width increase per km is approximately $\Delta/2c_1 \approx 24$ ns/km; for a 100-km-long fiber, an impulse spreads to a width of 2.4 μs. To minimize σ_τ, it is clearly desirable to use fibers with small values of Δ.

Graded-index fibers. Modal dispersion can be reduced by making use of **graded-index (GRIN) fibers** [Fig. 25.1-1(b)]. These fibers are designed such that the refractive index of the core varies gradually from a maximum value n_1 on the fiber axis to a minimum value n_2 at the core–cladding boundary (Sec. 10.2C). Rays then follow curved trajectories, with paths that are shorter than those in the step-index fiber. The axial ray travels the shortest distance but at the smallest phase velocity (largest refractive index), while the oblique rays travel longer distances but at higher phase velocities (smaller refractive indices), so that the delay times are approximately equalized. If the fiber is optimally graded (using an approximately parabolic profile), the pulse spreading rate (ps/km) is proportional to that of the equivalent step-index fiber, with a proportionality constant $\Delta/2$. For $\Delta = 0.01$, for example, the pulse spread of the GRIN fiber is theoretically reduced by a factor of 500 relative to that of the step-index fiber; in practice, however, the improvement is generally more moderate because of the difficulty of achieving the ideal index profile.

Short-haul multimode fiber-optic links. Silica multimode fibers serve as efficient, low-cost, links for short-haul communications. These systems often operate at 850 nm and are used in local area networks (LANs), datacenters, and financial centers. Various MMF grades at this wavelength are specified in terms of their optical mode (OM) designators: OM1 and OM2 operate at bit rates of 1 Gb/s over distances of 300 m and 600 m, respectively; OM3 and OM4 operate at 10 Gb/s at distances of 300 m and 550 m, respectively.

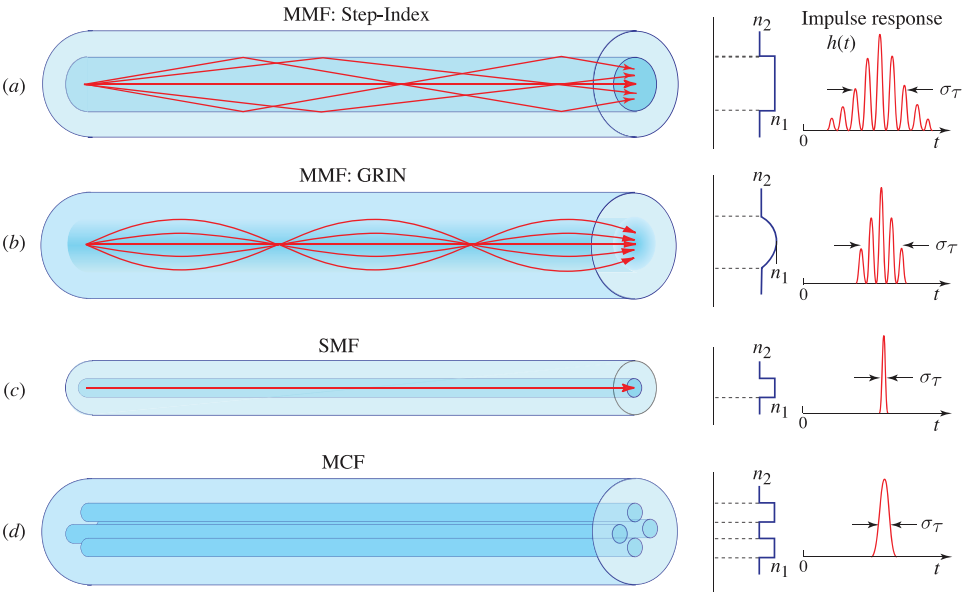


Figure 25.1-1 (a) Step-index multimode fiber (MMF): relatively large core diameter; uniform refractive indices in core and cladding; large pulse spreading arising from modal dispersion. (b) Graded-index multimode fiber (MMF: GRIN): graded refractive index of core; fewer modes; reduced pulse broadening arising from modal dispersion. (c) Single-mode fiber (SMF): small core diameter; no modal dispersion; pulse broadening arises only from material and waveguide dispersion. (d) Multicore fiber (MCF): well-separated cores serve as independent optical single-mode or multimode waveguides; closer cores support coupled core propagation in the form of supermodes.

Single-Mode Fibers (SMFs)

When the core radius a and the numerical aperture NA of a step-index fiber are sufficiently small so that $V < 2.405$, only a single mode is allowed and the fiber is called a **single-mode fiber** [Fig. 25.1-1(c)]. One advantage of using a SMF is the elimination of pulse spreading caused by modal dispersion (Sec. 10.2B). Pulse spreading occurs, nevertheless, since the initial pulse has a finite spectral linewidth and the group velocities (and therefore the delay times) are wavelength dependent. This effect is known as **chromatic dispersion**. There are two origins of chromatic dispersion: **material dispersion**, which results from the dependence of the refractive index on the wavelength, and **waveguide dispersion**, which is a consequence of the dependence of the group velocity of the mode on the ratio between the core radius and the wavelength. Material dispersion is usually larger than waveguide dispersion.

A short optical pulse of spectral width σ_λ spreads to a temporal width

$$\sigma_\tau = |D|\sigma_\lambda L, \tag{25.1-3}$$

Response Time (SMF)

which is proportional to the propagation distance L (km) and to the source linewidth σ_λ (nm). The dispersion coefficient D (ps/km-nm) involves a combination of material and waveguide dispersion. For weakly guiding fibers ($\Delta \ll 1$), D may be separated into a sum $D_\lambda + D_w$ for the material and waveguide contributions, respectively.

Consider, as an example, a SMF with a light source of spectral linewidth $\sigma_\lambda = 1$ nm (such as that emitted from a typical single-mode laser) and a fiber dispersion coefficient $D = 1$ ps/km-nm (for a silica fiber operating near $\lambda_o = 1300$ nm with

minimal waveguide dispersion). The response time calculated from (25.1-3) turns out to be $\sigma_\tau/L = 1$ ps/km; for a 100-km-long fiber, an impulse spreads to a width of 100 ps, far less than that for the multimode step-index fiber considered above.

The geometries, refractive-index profiles, and pulse broadening for step-index multimode, graded-index, and single-mode fibers are schematically compared in Fig. 25.1-1. **Multicore fibers** (Sec. 10.2D), which have multiple cores embedded in the same cladding, are included.

Material Attenuation and Dispersion

The wavelength dependence of the attenuation coefficient for fused silica-glass fibers is illustrated in Fig. 25.1-2. As the wavelength increases beyond the visible band, the attenuation coefficient α decreases to about 0.3 dB/km at $\lambda_o = 1300$ nm and, aside from a slight bump arising from residual OH-ion absorption near 1380 nm (Sec. 10.3A), falls to a minimum of ≈ 0.16 dB/km at $\lambda_o = 1550$ nm. The attenuation rises sharply as the wavelength rises beyond 1700 nm.

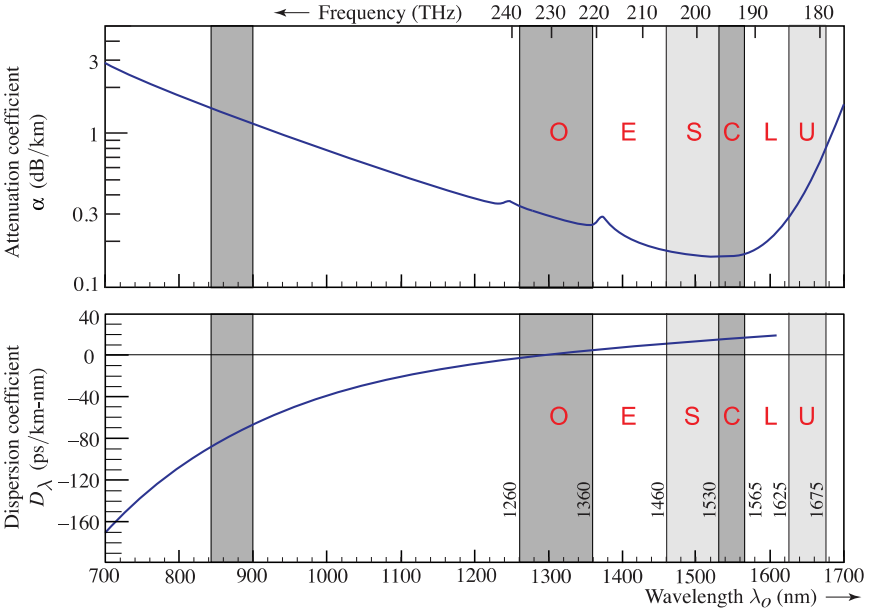


Figure 25.1-2 Wavelength dependence of the attenuation coefficient α (dB/km) and the material dispersion coefficient D_λ (ps/km-nm) for silica-glass fibers with suppressed OH absorption. Three telecommunications bands are highlighted in dark shading: the band centered at 870 nm, which was used in the earliest systems, has $\alpha \approx 1.5$ dB/km and $D_\lambda \approx -80$ ps/km-nm; the O (original) band centered at 1310 nm, for which $\alpha \approx 0.3$ dB/km and the dispersion is minimal; and the C (conventional) band centered at 1550 nm, for which the attenuation is minimal ($\alpha \approx 0.16$ dB/km) and $D_\lambda \approx +17$ ps/km-nm. The additional bands used in wavelength-division multiplexing (WDM) systems include: E = Extended, S = Short, L = Long, and U = Ultra-long.

The wavelength dependence of the dispersion coefficient D_λ for fused silica glass fibers is also displayed in Fig. 25.1-2. It exhibits negative values at short wavelengths that become positive at long wavelengths, and it passes through zero at $\lambda_o \approx 1312$ nm (Sec. 10.3B). In a medium with a negative dispersion coefficient, the shorter-wavelength components of a pulse travel more slowly than the longer-wavelength components, and therefore arrive later. This condition is known as **normal dispersion**. The opposite situation, called **anomalous dispersion**, occurs in a medium that exhibits a positive dispersion coefficient (Sec. 5.7). Though the sign of the

dispersion coefficient does not affect the pulse-broadening rate, it does play an important role in pulse propagation through media comprising cascades of materials with dispersion coefficients of different signs (Secs. 25.2D and 23.3).

Dispersion-Modified Fibers

As described in Sec. 10.3B, advanced designs of single-mode fibers make use of graded-index cores with special refractive-index profiles. These are selected such that the overall chromatic dispersion coefficient D attains desired values at particular wavelengths, or assumes a wavelength dependence that is useful in optical fiber communication systems, as in the following examples:

- In **dispersion-shifted fibers** (**DSFs**), D vanishes at λ_o = 1550 nm, where attenuation is minimum, rather than at 1312 nm [see Fig. 10.3-6(a)]. In **non-zero dispersion-shifted fibers** (**NZ-DSFs**), D is significantly reduced in the 1500–1600 nm window, but it is not zero. Indeed, a small amount of dispersion can be useful in mitigating nonlinear distortions encountered by narrow intense pulses. The wavelength dependence of D in DSF and NZ-DSF fibers is illustrated in Fig. 25.1-3.
- In **dispersion-flattened fibers** (**DFFs**), D vanishes at two wavelengths and is reduced at intermediate wavelengths [see Fig. 10.3-6(b)].
- In **dispersion-compensating fibers** (**DCFs**), D is proportional to that of the conventional step-index fiber over an extended wavelength band, but has the opposite sign. A short length of fiber with a reversed large dispersion coefficient can be used to compensate the pulse spreading introduced in long lengths of conventional fiber [see Fig. 10.3-6(c)].

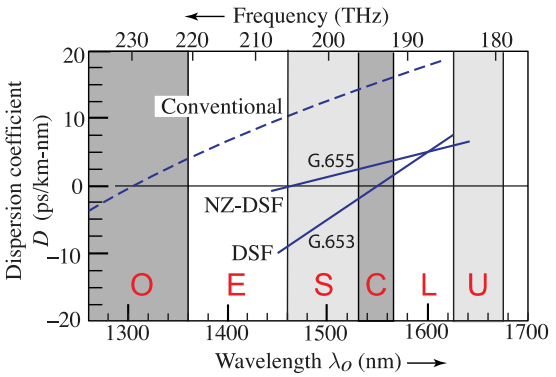


Figure 25.1-3 Wavelength dependence of the chromatic dispersion coefficient D for a conventional fiber and for examples of a dispersion-shifted fiber (DSF) and a non-zero dispersion-shifted fiber (NZ-DSF). The designations G.653 and G.655 are specifications of the ITU (International Telecommunications Union).

Other dispersion-modified fibers include **holey fibers** and **photonic-crystal fibers** (PCFs), as considered in Sec. 10.4. In these fibers, chromatic dispersion is dominated by waveguide dispersion, which is strongly dependent on the geometry of the holes. Dispersion flattening over broad wavelength ranges can be achieved as can dispersion shifting to wavelengths shorter than the zero-material-dispersion wavelength. A holey fiber may be designed to operate as a single-mode waveguide over a broad range of wavelengths (endlessly single-mode fibers). In fibers with a hollow core and a cladding with holes arranged in a periodic structure, light is guided in the core by reflection from the surrounding photonic-crystal cladding. Since the light travels in the hollow core, it suffers lower losses and reduced nonlinear effects.

Polarization Mode Dispersion

Another form of pulse spreading, known as **polarization mode dispersion** (**PMD**), is caused by random anisotropic changes in the fiber introduced by environmental and structural factors along its length. Random variations in the magnitude and orientation

of the birefringence introduce differential delays between the two polarization modes and, as described in Sec. 10.3B, the RMS value of the pulse broadening associated with PMD is proportional to the square root of the fiber length,

$$\sigma_{\text{PMD}} = D_{\text{PMD}}\sqrt{L}, \tag{25.1-4}$$

Polarization Mode Dispersion

where D_{PMD} is a dispersion parameter that typically ranges from 0.1 to 1 $\text{ps}/\sqrt{\text{km}}$. Polarization mode dispersion becomes important at high data rates when other forms of dispersion are compensated.

Nonlinear Optical Effects

Silica-glass fibers exhibit two kinds of optical nonlinear effects that are germane to their use in optical fiber communication systems — third-order nonlinearity, which underlies the optical Kerr effect; and nonlinear inelastic scattering, which includes stimulated Raman and Brillouin scattering. When high-power optical pulses are transmitted through single-mode fibers with small cross-sectional areas, the optical intensity may be sufficient for these nonlinear interactions to cause deleterious effects that damage signal integrity and limit transmission distance and speed:

- *Self-phase modulation* (SPM) is a form of nonlinear dispersion caused by the optical Kerr effect (a dependence of the refractive index, and hence the phase velocity, on the optical intensity, as described in Sec. 22.3A). Pulse spreading ensues since different segments of the optical pulse travel at different velocities (Sec. 23.3B). The optical Kerr effect may also result in crosstalk between counterpropagating waves in two-way communication systems.
- *Cross-phase modulation* (XPM) results from nonlinear wave mixing wherein the phase velocity of a wave at one wavelength depends on the intensities of waves at other wavelengths traveling simultaneously in the same fiber (Sec. 22.3C). In wavelength-division-multiplexed (WDM) systems, XPM can cause substantial crosstalk among different channels.
- *Four-wave mixing* (FWM) is also associated with third-order nonlinear effects (Sec. 22.3D). It causes crosstalk between four waves of different wavelengths traveling simultaneously in the same fiber since the waves may exchange energy. This introduces an intensity-dependent gain/loss into the channels of a WDM system.
- *Stimulated Raman scattering* (SRS) and *stimulated Brillouin scattering* (SBS) are inelastic scattering processes that involve interactions between light and molecular or acoustic vibrations of the medium. In these processes, two optical waves of different wavelengths interact via a molecular vibration mode (SRS) or an acoustic vibration mode (SBS) (Secs. 14.5C, 15.3D, and 16.3C). Such interactions also lead to undesirable crosstalk among channels in a WDM system.

The deleterious effects of nonlinear phenomena in optical fiber communications can be mitigated by increasing the fiber core diameter, thereby reducing the energy density. This can sometimes be achieved by making use of large-mode-area photonic-crystal fibers (Sec. 10.4), although this sometimes requires the deployment of new fiber. However, the nonlinear properties of fibers can also be harnessed for useful applications in communication systems. Nonlinear dispersion via SPM may be adjusted to compensate for chromatic dispersion in the fiber, giving rise to optical solitons (Sec. 23.5B). Nonlinear interactions can also be used to provide useful gain via FWM or SRS. Raman and Brillouin fiber amplifiers are discussed in Secs. 15.3D and 25.1C.

B. Sources for Optical Transmitters

The requirements for the light source used in an optical fiber communication system depend on the nature of the intended application, e.g., long-haul communication or short-haul local-area network.

A number of features are important in determining the source of choice:

- *Wavelength*. The wavelength of the source must be compatible with the fiber medium, usually near the wavelength of its minimum loss or minimum dispersion.
- *Power*. The source power must be sufficiently high so that, after transmission through the fiber and any associated amplification, the received signal is detectable with the required accuracy.
- *Speed*. The source power must be able to be modulated at the rate desired for imparting information.
- *Linewidth*. The spectral linewidth must be sufficiently narrow so that phase noise and the effect of fiber chromatic dispersion is minimized.
- *Noise*. Random fluctuations in the source power and frequency are to be avoided, particularly for coherent communication systems.
- *Other features*. Other important features include ruggedness, compactness, reliability, low cost, long lifetime, and insensitivity to environmental variables such as temperature.

Laser-Diode and VCSEL Sources

Quantum-confined laser diodes (LDs) and vertical-cavity surface-emitting lasers (VCSELs), discussed in Secs. 18.4 and 18.5A, respectively, are both widely used as sources in optical fiber communication systems. Long-haul and short-haul systems, considered in turn, principally make use of LDs and VCSELs, respectively.

Long-haul systems. Long-haul, high-bit-rate communication systems generally rely on single-mode silica-glass fiber, billions of kilometers of which span the globe. Operation at wavelengths in the 1.3–1.6 μm telecommunications band, the region of minimal dispersion and attenuation (Fig. 25.1-2), is readily accommodated by making use of laser diodes fabricated from InGaAsP, a direct-bandgap, quaternary III–V semiconductor (Sec. 18.1C). As discussed in Sec. 18.4A, multiquantum-well lasers, and their strained-layer counterparts, are widely used in such systems because of their superior properties. Edge-emitting distributed-feedback (DFB) laser diodes are particularly good candidates. As illustrated in Fig. 25.1-4, these devices make use of a corrugated-layer grating, placed adjacent to the active region, that acts as a distributed reflector and imposes single-frequency operation.

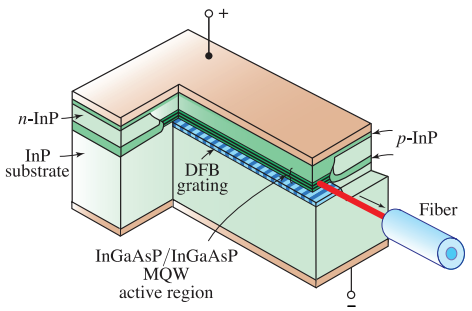


Figure 25.1-4 Buried-heterostructure MQW DFB laser used for long-haul optical fiber communications in the 1.3–1.6-μm telecommunications band. This single-frequency laser operates on a single spatial and a single longitudinal mode. It generates an output power $P_o > 1$ W with a spectral width $\Delta\nu_L$ of a few MHz.

These devices can operate on a single spatial and longitudinal mode and deliver optical powers of watts to tens of watts, power-conversion efficiencies of 70%, modulation rates of tens of Gb/s, spectral widths of a few MHz, and life spans of years.

The operation of these low-noise lasers is robust in the presence of external temperature variations and modulation since they are immune to the deleterious effects of frequency chirping, the change in laser frequency that results from refractive-index variations accompanying fluctuations in the carrier concentration as the drive current is modulated. Typical values of the threshold current and differential responsivity are $i_{\mathrm{t}} < 10$ mA and $\mathrm{R}_d \approx 0.4$ W/A, respectively.

Short-haul systems. Short-haul communication systems also often make use of silica fibers but rely on single-mode VCSELs that operate in the 1.3–1.6-μm telecommunications band. These devices consume little electrical power and can be modulated at bit rates exceeding 10 Gb/s. Their low optical power can be mitigated by the use of amplifiers. Some short-haul silica-fiber-based systems operate at 850 nm using AlGaAs devices. Other, less expensive, plastic-fiber systems operate at 650 nm and rely on AlInGaP devices; 1-mm-diameter, polymer-core fibers can yield data rates of 1 Gb/s. Modules that include VCSELs with WDM functionality are available for high-bandwidth communications (Sec. 25.1E).

Other Sources

A number of sources other than LDs and VCSELs are used, or are potentially useful, as optical transmitters.

LEDs. Whether edge-emitting or surface-emitting (Sec. 18.1), LEDs are generally inferior to LDs and VCSELs for use in optical fiber systems. This stems from their lower power, lower conversion efficiency, and lower modulation rate, along with their larger spectral width and larger light-emission angle, which makes it difficult to couple the light into an optical fiber. The principal advantage of using an LED source is low cost. AlInGaP/InGaP resonant-cavity LEDs operating at 650 nm (Fig. 18.1-20), and InGaAsP LEDs operating at 1.3 μm (Fig. 18.1-19), have indeed been used as sources in short-haul, modest-bit-rate systems. By-and-large, however, LEDs have ceded their ground to VCSELs, which have superior performance.

MQD lasers. Provided that they emit sufficiently high power, multiquantum-dot lasers (Sec. 18.4C) have the distinct merits of small size, low power consumption, low threshold, reduced linewidth, enhanced modulation bandwidth, and resistance to operating-temperature variations.

Fiber DFB lasers. Some fiber systems, such as those that make use of WDM and coherent communications, respectively), require sources whose wavelength can be tuned. Coherent communication systems with advanced modulation formats, in particular, impose stringent requirements on source stability, linewidth, phase noise, and local-oscillator tunability. External-cavity wavelength-tunable laser diodes (Sec. 18.3C) often serve as sources for such systems, but more robust and less expensive alternatives, such as fiber distributed-feedback lasers, are under development.

Multimode lasers. Though they provide greater power, multimode laser diodes are often avoided since they suffer from partition noise. When subjected to chromatic dispersion in the fiber channel, the random distribution of laser power among the modes leads to random intensity fluctuations and reshaping of the transmitted pulses. Multimode VCSELs that operate in the 750–850-nm wavelength region are nevertheless sometimes used for short-haul optical fiber communications.

Mid-IR QCLs. Interest in mid-infrared optical fiber communications stems from several relatively recent developments in infrared photonics: 1) advances in the development of fluoride and other soft-glass fibers (Sec. 10.5), which exhibit substantially reduced Rayleigh scattering and absorption in comparison with silica glass; 2) the operation of quantum cascade lasers (QCLs) in the mid infrared (Sec. 18.4D); and

3) the evolution of mid-infrared photodetectors that make use of materials such as HgCdTe and VOx (Secs. 19.4 and 19.5).

C. Optical Amplifiers

Optical amplifiers are indispensable components in modern long-haul optical fiber communication systems. They find use as power amplifiers (also called postamplifiers), line amplifiers, and preamplifiers. As illustrated in Fig. 25.1-5, power amplifiers augment the optical power before light is launched into an optical fiber, line amplifiers serve to boost the signal in the course of transmission (Fig. 25.0-1), and preamplifiers provide gain before photodetection.

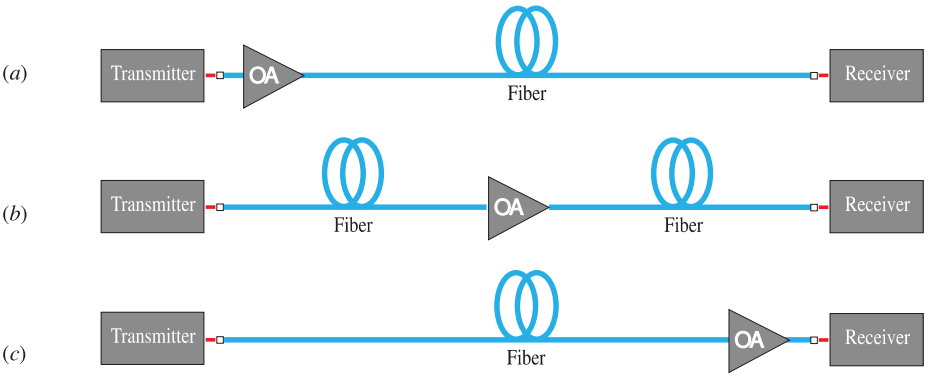


Figure 25.1-5 Optical fiber amplifiers are used in three configurations in optical fiber communication systems: (*a*) power amplifiers; (*b*) line amplifiers; and (*c*) preamplifiers.

In this section, we consider in turn the three types of optical fiber amplifiers (OFAs) that are typically used in optical fiber communication systems:

- Erbium-doped fiber amplifiers (EDFAs) (Sec. 15.3C).
- Rare-earth-doped fiber amplifiers (REFAs) (Sec. 15.3C).
- Raman fiber amplifiers (RFAs) (Sec. 15.3D).

Though semiconductor optical amplifiers (Sec. 18.2) are compact and compatible with photonic integrated circuits, their disadvantages outweigh their merits in the domain of optical fiber communications. In particular, they are inferior to OFAs in terms of fiber geometry, gain, interchannel interference, intersymbol interference, noise, and temperature sensitivity. Similarly, optical parametric amplifiers (Sec. 22.2C) offer substantial gain and broadband tunability but are inferior to OFAs in that they require phase matching and suffer from nonlinearities and sensitivity to polarization.

Erbium-Doped Fiber Amplifiers (EDFAs)

Erbium-doped fiber amplifiers (EDFAs), which were the first OFAs to be developed, are widely used in optical fiber communication systems. As discussed in Sec. 15.3C, they offer high polarization-independent gain, high output power, high efficiency, low insertion loss, low noise, and a broad transition that offers gain in the vicinity of $\lambda_o = 1550$ nm (corresponding to the wavelength of minimum loss for silica optical fibers, as shown in Fig. 25.1-2). Quasi-three-level pumping is achieved by longitudinally coupling light into the optical fiber, in the forward or backward direction, or bidirectionally (Fig. 15.3-5). The pump light is usually generated by strained quantum-well InGaAs laser diodes operating at $\lambda_o = 980$ nm, although in-band pumping at 1480 nm is also used. Ytterbium is usually added to erbium as a co-dopant to increase efficiency.

Gains in excess of 50 dB can be achieved in EDFAs with tens of mW of pump power; signal output powers in excess of 100 W are readily generated. Line-amplifier spacings are often ≈ 50 km. The available bandwidth is $\Delta\lambda \approx 40$ nm, corresponding to $\Delta\nu \approx 5.3$ THz, which accommodates the C-band. The L-band can be accommodated by modifying the optimization parameters of the EDFAs. The large gain and bandwidth offered by these amplifiers make them ideal for use in wavelength-division multiplexing (WDM) systems (Sec. 25.3C). The mixed homogeneous/inhomogeneous broadening leads to a wavelength-dependent gain profile that can require gain equalization. However, EDFAs are often operated in the saturated regime and exhibit minimal crosstalk between different signals that are simultaneously transmitted through them.

Rare-Earth-Doped Fiber Amplifiers (REFAs)

Rare-earth-doped fiber amplifiers (REFAs) other than Er^{3+} that are useful for optical fiber communications include Tm^{3+} and Pr^{3+}. Good performance can be obtained from Tm^{3+}-doped REFAs operating in the S- and U-bands and from Pr^{3+}-doped REFAs operating in the O-band (Fig. 25.1-2). Though neither of these REFAs offers the kind of gain and efficiency achievable with EDFAs, mixing and matching Er^{3+} and Tm^{3+} fiber amplifiers can provide a channel bandwidth of $\Delta\lambda \approx 150$ nm, corresponding to $\Delta\nu \approx 18.8$ THz at 1550 nm.

Raman Fiber Amplifiers (RFAs)

Raman fiber amplifiers (RFAs) operate on the basis of stimulated Raman scattering (Secs. 14.5C, 15.3D, and 16.3C). As discussed in Sec. 15.3D, there are two standard RFA configurations: (1) distributed RFAs where the signal and pump are both sent through a transmission fiber that serves as the gain medium; and (2) lumped RFAs in which a short length of highly nonlinear fiber serves as the amplifier and provides gain. As with EDFAs, pumping can be in the forward or backward direction, or bidirectional.

RFAs typically offer greater bandwidths than EDFAs. The bandwidth over which Raman gain is available in germanium-doped silica fiber is about 100 nm (corresponding to about 12.5 THz at 1550 nm). Moreover, multiple pumps at different frequencies can be combined to provide far greater bandwidths; indeed, Raman amplification can, in principle, be employed over the entire region of fiber transparency. The gain of a RFA, which can reach ≈ 20 dB, is substantially lower than that of an EDFA, as are the efficiency and gain efficiency. However, this can be mitigated in part by making use of dispersion-compensating fiber to simultaneously achieve gain and accommodate the different signal- and pump-pulse frequencies. The relative merits of EDFAs and RFAs have been detailed in Sec. 15.3D. In spite of the apparent shortcomings of RFAs in comparison with EDFAs, their wider bandwidths, arbitrary operating wavelengths, and compatibility with existing systems render them of interest for certain applications. They are particularly attractive for use in those telecommunications bands where EDFAs and other REFAs are unavailable or inefficient.

D. Photodetectors for Optical Receivers

A comprehensive discussion of various photodetectors has been provided in Chapter 19. The two most commonly used types of detectors employed in optical fiber communication systems are p–i–n (PIN) photodiodes and avalanche photodiodes (APDs). By virtue of their larger bandwidths, PIN photodiodes are widely used for systems that rely on coherent communications and that make use of high-efficiency modulation techniques. Schottky-barrier photodiodes, which can have bandwidths ≈ 100 GHz, are also used.

Systems that make use of direct detection enjoy a competitive advantage in simplicity and cost, and are tolerant to chromatic dispersion, intersymbol interference,

and polarization mode dispersion. Direct-detection systems often make use of APDs, which have the advantage that they provide gain before the first electronic receiver amplification stage, thereby reducing the detrimental effects of circuit noise. However, the multiplication mechanism inherent in APDs introduces intrinsic gain noise and engenders an avalanche multiplication time that can reduce receiver bandwidth. Furthermore, APDs require greater voltage and more complex circuitry than PIN detectors, and can often require temperature stabilization. Nevertheless, low-noise APDs with enhanced gain–bandwidth products are widely used in high bit-rate telecommunication systems. The signal-to-noise ratio and sensitivity of receivers using p–i–n photodiodes and APDs has been considered in Sec. 19.6.

Photodetectors in the 1300–1600-nm Wavelength Range

As illustrated in Fig. 25.1-2, the attenuation and dispersion properties of silica optical fibers favor operation in the 1.3–1.6-μm wavelength region comprising the O, E, S, C, L, and U telecommunications bands. Silicon is not photosensitive in this wavelength range as it is transparent ($\lambda_g = 1.11\ \mu\text{m} < \lambda_o$; see Table 17.1-2).

InGaAs PIN and APD Photodetectors. The most widely used photosensitive material in this wavelength range is $In_{0.53}Ga_{0.47}As$, lattice-matched to InP, for which the bandgap energy $E_g = 0.75$ eV ($\lambda_g = 1.65\mu$m). Both PIN and SACM (separate absorption, charge, and multiplication) InGaAs/InGaAsP/InP (A/C/M) APDs are extensively used for long-haul systems that operate in this wavelength range. Waveguide structures offer larger bandwidths.

A typical InGaAs PIN photodiode operating at 1550 nm has quantum efficiency $\eta \approx 0.80$, responsivity $\mathcal{R} \approx 0.95$ A/W (Fig. 19.3-9), and bandwidth $B \approx 10$ GHz.

A typical InGaAs/InGaAsP/InP SACM APD operated at a reverse bias of tens of volts provides a mean gain $\overline{G} \approx 10$ and has a bandwidth $B \approx 10$ GHz (Example 19.4-2). Though normal-incidence InGaAs APDs achieve excellent receiver sensitivities for bit rates as high as 10 Gb/s, three factors limit their performance at higher bit rates: 1) since $\alpha \approx 10^4$ cm^{-1} for InGaAs in this wavelength range, the absorption region must be $\approx 2.5\ \mu$m thick to attain $\eta > 90\%$; 2) the associated transit time then limits the bandwidth to $B \approx 10$ GHz at low gain values; at higher values of the gain, where the avalanche buildup time comes into play, the relatively low gain–bandwidth product ($GB < 100$ GHz) restricts the frequency response; and 3) the ionization ratio $1/\text{k} \approx 0.3$ for InP results in significantly higher excess noise than that attainable with Si. On the other hand, the gain noise of the APD can be reduced, and its speed increased, by making use of $Al_{0.48}In_{0.52}As$ multiplication layers, which offer a reduction of the ionization ratio $1/\text{k} \approx 0.2$, and further by making these layers sufficiently thin so that dead-space (history-dependent ionization) effects play a salutary role (Sec. 19.6B). The gain–bandwidth product can then be increased to $GB \approx 235$ GHz.

Ge-on-Si PIN and APD Photodetectors. Germanium can also serve as a photosensitive material for PIN and APD detectors in the 1.3–1.6-μm wavelength region. Though the use of Ge by itself has a number of drawbacks, group-IV-photonics monolithic devices that make use of Ge-on-Si have the merit that they are CMOS-compatible and hence available for on-chip integration. Waveguide-based devices offer both high quantum efficiency and high speed since they decouple the light absorption and carrier collection. There are two principal limitations associated with the use of Ge-on-Si devices, however: 1) the photosensitivity of Ge falls off rapidly for $\lambda_o > 1.55\ \mu$m, and 2) the dark current is relatively large as a result of the lattice mismatch between Ge and Si.

The particular design for a Ge-on-Si waveguide PIN photodiode operating at 1550 nm described in Example 19.3-2 relies on the butt coupling of light from a Si waveguide to the i region of a lateral p–i–n Ge-on-Si photodiode integrated at its

end. This device offers a responsivity R $\approx$ 1 A/W and a bandwidth $B > 50$ GHz. Its performance is comparable to that obtained via the hybrid integration of InGaAs on Si.

In the normal-incidence Ge-on-Si SACM APD considered in Example 19.4-3, photons impinge on a Ge absorber layer grown on top of a layer of Si, where carrier multiplication takes place. To restrict the dark current, low electric field in the absorption region is maintained by a Si charge layer. With unintentionally doped Ge and Si layers of thicknesses 1 and ½ μm, respectively, and a 0.1-μm p-type Si charge layer, this APD exhibits a mean gain $\overline{G} \approx 50$, a gain–bandwidth product $GB \approx 350$ GHz (which substantially exceeds that of an InGaAs/InP APD), a responsivity R $\approx$ 5.9 A/W at $\lambda_o = 1.3$ μm, an ionization ratio k ≈ 0.09, and operation at bit rates of 25 Gb/s.

AlInAsSb APD Photodetectors. Recent efforts for improving APD performance have focused on the development of new structures and materials that achieve lower noise and higher speeds, while maintaining adequate gain. An example is the AlInAsSb SACM APD, a III–V direct-bandgap device that operates across the 1.3–1.6-μm telecommunications band. A device such as that described in Example 19.4-4 has a quantum efficiency $\eta \approx 0.4$ and a dark current that is somewhat greater than that of the InGaAs/AlInAs APD discussed above, but substantially smaller than that of the Ge-on-Si APD. The gain–bandwidth product is anticipated to be $GB > 300$ GHz. The ionization ratio k $\approx$ 0.01 at a mean gain $\overline{G} = 10$ is comparable to that of Si (Example 19.6-4).

Photodetectors in the 800–900-nm Wavelength Range

As discussed in Sec. 25.2A, first-generation optical fiber communication systems operated at $\lambda_o \approx 870$ nm to match the wavelength of the GaAs LEDs and laser diodes that were developed in the early 1960s (Sec. 18.1C). Short-haul silica-fiber-based systems that are implemented today in this wavelength range make use of $Al_xGa_{1-x}As$ sources (Sec. 25.1B).

Si PIN and APD Photodetectors. Silicon p–i–n photodiodes and APDs are highly effective in this wavelength range. A commercially available Si PIN photodiode operating at 850 nm, for example, has a quantum efficiency $\eta \approx 0.9$, a responsivity R $\approx$ 0.6 A/W (Fig. 19.3-9), and a bandwidth $B \approx 15$ GHz.

In the domain of Si APDs, a separate absorption and multiplication (SAM) reach-through APD has its peak sensitivity at a wavelength of 800 nm (Example 19.6-3). Under specified operating conditions, this device has a quantum efficiency $\eta = 0.8$, a mean gain $\overline{G} = 50$, a gain–bandwidth product $GB = 350$ GHz, and an ionization ratio k $= 0.02$.

Photodetectors in the Mid Infrared

As discussed in Sec. 25.1B, interest in mid-infrared optical fiber communications has been fostered by advances in fluoride and other soft-glass fibers (Sec. 10.5), by the advent of quantum cascade lasers (Sec. 18.4D), and by improvements in mid-infrared detectors (Sec. 19.4).

HgCdTe APD Photodetectors. Substantial progress has been made in the quality of $Hg_xCd_{1-x}Te$ SAM APDs, which have cutoff wavelengths that extend from 2 to 11 μm in the mid IR. A device such as that described in Example 19.4-5 enjoys high quantum efficiency ($\eta \approx 0.9$), high gain ($\overline{G} > 1000$), and large gain–bandwidth product ($GB > 1$ THz), although it requires cryogenic cooling. The ionization ratio k $\approx$ 0 is superior to that of Si (Example 19.6-4) so that the excess noise factor $F = \langle G^2 \rangle / \langle G \rangle^2$, defined in (19.6-24), is limited to $F \leq 2$ and is independent of $\langle G \rangle$. The gain–voltage characteristic is exponential.

E. Photonic Integrated Circuits

Contemporary optical fiber communication systems make use of advanced modulation formats to achieve high capacities. Highly coherent lasers with narrow linewidths are often required. **Silicon photonics** provides a useful platform for fabricating high-quality on-chip light sources that rely on the integration of direct-bandgap III–V photon emitters with CMOS-compatible silicon substrates (Sec. 18.1D). Quantum-dot active regions often replace their quantum-well cousins because of reduced sensitivity to temperature variations and to dislocations arising from the juxtaposition of the different materials (Sec. 18.4C).

As discussed in Chapter 9, **integrated photonics** (integrated optics) is the technology of combining, on a single chip, collections of optical devices and components to achieve a particular purpose or carry out a specific function. The design and fabrication of **photonic integrated circuits (PICs)** draws on integrated photonics and silicon photonics. PICs serve to miniaturize and increase the density of photonic circuitry in much the same way that electronic integrated circuits (ICs) miniaturize and increase the density of electronic circuitry. In **monolithic PICs**, the various elements are simultaneously fabricated on the same chip, and that there are many chips per wafer. Passive PICs are also called **planar lightwave circuits (PLCs)** and PICs that incorporate both photonic and electronic components are also referred to as **optoelectronic integrated circuits (OEICs)** Group-IV PICs can comprise Si waveguides, together with Ge-based lasers, modulators, and photodetectors.

The telecommunications industry long relied on PICs fabricated wholly from III–V materials since these semiconductors readily emit light, accommodate modulation via the electro-optic effect, and can be lattice-matched to a range of related materials. Though thousands of devices can be integrated into an InP-based PIC with high yield, the fabrication process is complex and expensive. In contrast, the compatibility of Si-based PICs with CMOS, and the manifold benefits of group-IV materials, yield far higher integration densities along with more efficient and less expensive integration. Moreover, silicon-photonics-based PICs can offer improved functionality by drawing on the different salutary features of III–V materials and silicon. As an example, high gain can be provided by a III–V direct-bandgap material while photon storage is relegated to a ring resonator in an undoped silicon layer with low loss and high Q.

PICs often incorporate collections of standard on-chip building blocks, such as lasers, modulators, couplers, splitters, amplifiers, and photodiodes. Optical waveguides link the various components. PICs find use in generating, focusing, splitting, combining, isolating, polarizing, coupling, modulating, transporting, multiplexing, switching, and detecting light. Higher-level PICs can incorporate both transmitter and receiver functionalities on the same chip and function as **optical transceivers** (**optochips**) and **coherent optical transceivers**. These devices find application in a broad variety of optical communication systems and networks, including **telecom** (long-haul, short-haul, and undersea), **datacom** (LANs and datacenters), and **computer-com** (chip interconnects and high-performance computing — see Sec. 24.1D). Transceiver PICs that incorporate standard on-chip building blocks along with filters and wavelength multiplexers/demultiplexers are commercially available at bit rates exceeding 100 Gb/s. PICs also find use as chip-based phase-sensitive sensors and in applications such as night-vision monitoring, integrated-optical gyroscopy, and lidar.

25.2 OPTICAL FIBER COMMUNICATION SYSTEMS

The simplest communication system is a point-to-point link. The information is carried by a physical variable (e.g., electrical, electromagnetic, or optical) containing a signal that is transmitted at one point and received at another. Transmitting more than one

signal simultaneously through the same link requires that the signals be marked by some distinct attribute (e.g., time, frequency, or wavelength) or identified by some distinct code. This scheme is called *multiplexing*.

In an optical fiber communication system, the link is an optical fiber that carries a light wave modulated by the signal. The modulated physical variable carrying the information may be the optical intensity, amplitude, frequency, phase, or polarization. The most straightforward example is an intensity-modulated optical communication system, as illustrated in Fig. 25.2-1.

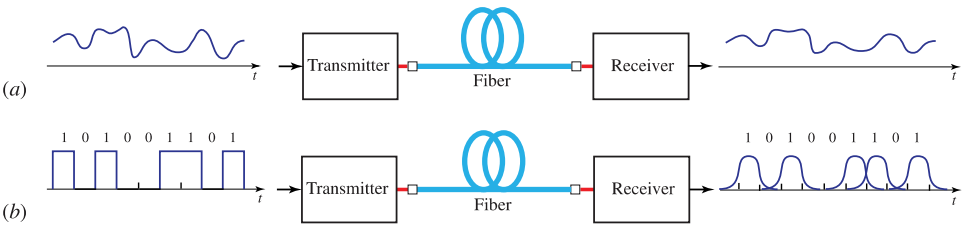


Figure 25.2-1 Optical fiber communication systems using intensity modulation. (*a*) Analog system: the power of the light source is proportional to the signal, which is a continuous function of time that could represent, for example, an audio or video waveform. (*b*) Digital ON–OFF keying system: the states "1" and "0" of a bit are represented, respectively, by the presence and absence of an optical pulse.

The simplest example of optical multiplexing is wavelength-division multiplexing (WDM), in which multiple optical signals are transmitted through the same fiber at distinct optical wavelengths, as illustrated in Fig. 25.2-2.

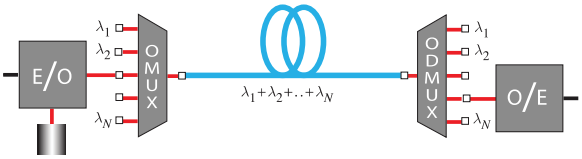


Figure 25.2-2 Wavelength-division multiplexing (WDM).

One measure of the performance of an analog communication system is its **bandwidth** B (Hz). This is the maximum frequency at which modulated optical power may be transmitted through the link such that the received signal is detectable with a prescribed signal-to-noise ratio. The bandwidth is determined by the response time of the overall communication channel as well as by the attenuation and the noise level at the receiver.

An analogous measure of the performance of a digital communication system is the maximum **bit rate** B_0 (bits per second, or b/s) at which bits of the received signal are discernible with an error rate not exceeding a prescribed value. This data rate is determined by the attenuation and pulse spreading introduced by the system, as well as by the noise level at the receiver. The bit rates displayed in Table 25.2-1 represent optical carrier (OC) levels defined by the *Synchronous Optical Network* (SONET) standard for optical telecommunications technology.

Table 25.2-1 Approximate bit rates for the SONET standard.

OC-1	OC-3	OC-12	OC-24	OC-48	OC-192	OC-768	OC-1920
52 Mb/s	156 Mb/s	622 Mb/s	1.25 Gb/s	2.5 Gb/s	10 Gb/s	40 Gb/s	100 Gb/s

This section begins with an overview of the evolution of optical fiber communication systems and is followed by a quantitative analysis of the performance limits of simple digital and analog systems that make use of intensity modulation.

A. Evolution of Optical Fiber Communication Systems

As illustrated in Fig. 25.1-2, the minimum attenuation in silica-glass fibers occurs at $\lambda_o \approx 1550$ nm, whereas the minimum material dispersion occurs at $\lambda_o \approx 1312$ nm. The choice of which of these wavelengths should be used to build a link depends on the availability of an appropriate light source as well as on the relative importance of power loss and pulse spreading, as explained in Sec. 25.2B. First-generation optical fiber communication systems operated at $\lambda_o \approx 870$ nm, the wavelength of the earliest available light-emitting diodes and laser diodes, which were fabricated from GaAs. However, optical-fiber attenuation and material dispersion are both relatively high at this wavelength, so subsequent systems were constructed using longer wavelengths as more advanced semiconductor materials became available. Second- and third-generation systems operated near 1310 and 1550 nm, respectively.

There are many possible combinations of operating wavelengths, devices, materials, and fibers that can be used to build an optical link — some of these are portrayed in Fig. 25.2-3. As new materials and capabilities have come to the fore, the implementation of fiber systems has generally proceeded along the following paths: (1) from shorter to longer wavelengths; (2) from multimode fibers (MMFs) to single-mode fibers (SMFs); (3) from light-emitting diodes (LEDs) to laser diodes (LDs); (4) from p–i–n photodiodes (PINs) to avalanche photodiodes (APDs); (5) from semiconductor optical amplifiers (SOAs) to optical fiber amplifiers (OFAs); and (6) from direct detection to coherent detection. This evolution has been made possible in large part by advances in bandgap engineering and glass engineering. The former allowed remarkable quaternary semiconductor components to be developed for use at longer wavelengths while the latter facilitated effective new designs for silica-glass optical fibers.

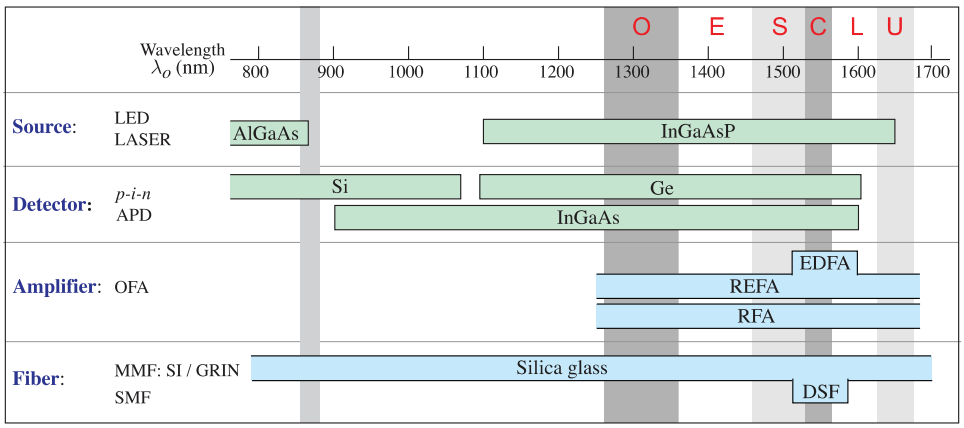


Figure 25.2-3 Materials commonly used for optical sources, detectors, amplifiers, and fibers, along with the ranges of wavelengths over which they operate. Gray and white vertical columns represent various optical fiber telecommunications bands. The first three generations of fiber-optic links operated at wavelengths near 870 nm, 1310 nm, and 1550 nm, respectively.

The evolution of fiber components and systems has been motivated by the desire to increase the transmission bit rate B_0 (b/s) as well as the length L (km) of the

communication link. Both B_0 and the product LB_0 (usually expressed in units of km-Gb/s) serve as measures to gauge the advancement of fiber communication systems.

The nine system generations discussed in the following pages characterize this evolution and Fig. 25.2-4 depicts the increase realized in B_0 and LB_0 as time has progressed. The first three systems, which are often referred to as the first three generations of optical fiber systems, achieved a 1000-fold increase in LB_0 from 1974 to 1990. For simplicity, we use these technologies as examples for evaluating system performance, as discussed in Sec. 25.2B. As is evident in Fig. 25.2-4, subsequent progress has extended these basic systems in a number of directions, and has led to an increase of B_0 and LB_0 by an additional eight orders of magnitude from 1990 to 2015. This tenfold increase every four years is known as "**optical Moore's law**."

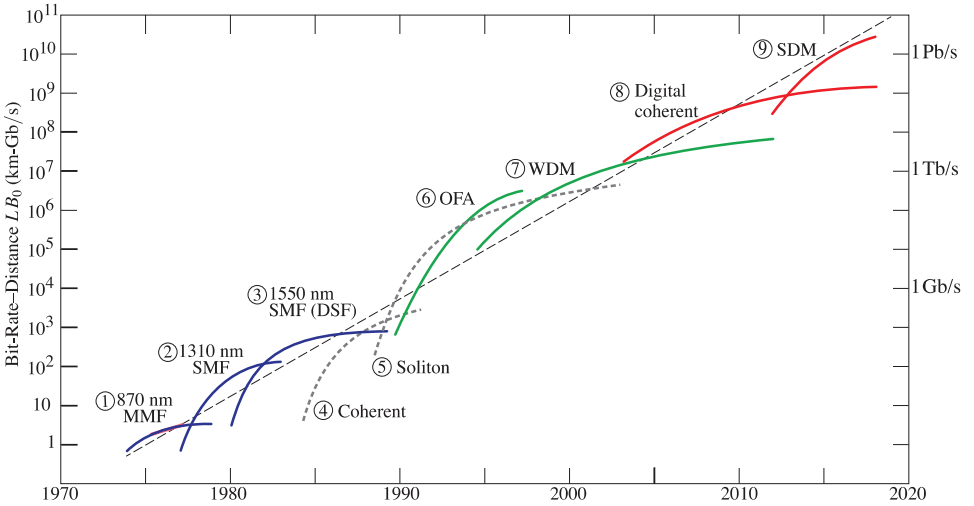


Figure 25.2-4 The development of optical fiber communication systems over the years reveals a continuous growth of bit rate B_0 (right ordinate) and bit-rate–distance product LB_0 (left ordinate, generally with $L = 10\,000$ km). Solid blue curves represent the earliest systems. Dotted black curves indicate systems that were developed subsequently but not widely implemented. Green curves represent systems that have been widely deployed over land and under sea. Red curves indicate systems developed more recently. Systems ⑦–⑨ incorporate both OFAs and WDM. The superb performance of the newer systems has led to the decommissioning of most commercial links based on Systems ①–⑥. The straight dashed line represents an "optical Moore's law" that reveals a tenfold increase in bit rate every four years.

System ①: *Multimode fiber (MMF) at 870 nm.* This was the early technology of the 1970s. Fibers were either step-index or graded-index. The light source was either an LED or a laser diode (initially GaAs and subsequently AlGaAs). Both Si p–i–n and APD photodetectors were used. System performance was limited by high fiber attenuation and modal dispersion. A typical communication link of this era operated at B_0 = 100 Mb/s, and had a length of L = 10 km, yielding $LB_0 \approx 1$ km-Gb/s. Several optical links could be concatenated to form a longer link for intercity communication by inserting optical-electrical-optical (OEO) units, known as *repeaters* or *regenerators*, between consecutive links. Each such repeater implemented three processes: photodetection, electrical amplification, and optical signal regeneration.

System ②: *Single-mode fiber (SMF) at 1310 nm.* The move to single-mode fibers and to a wavelength region with minimal material dispersion in the 1980 time frame led to a substantial improvement in performance, which was limited by fiber attenuation. InGaAsP laser diodes were used with either InGaAs p–i–n or APD photodetectors (sometimes Ge APDs). A typical long-haul link in this class operated at OC-12 (622 Mb/s) with an OEO repeater spacing of $L = 40$ km, yielding $LB_0 \approx 25$ km-Gb/s.

System ③: *Single-mode fiber (SMF) at 1550 nm.* Silica-glass fibers were used in the 1550-nm wavelength range, where their attenuation is lowest. Performance was limited by material dispersion, which was reduced by employing low-chirp single-frequency distributed-feedback (DFB) laser diodes (InGaAsP). APDs were often used. The subsequent use of dispersion-shifted fibers (DSF) reduced the deleterious effects of dispersion and boosted performance. An example of this system is a long-haul terrestrial or undersea link operating at 2.5 Gb/s (OC-48) over a distance of $L = 100$ km, so that $LB_0 \approx 250$ km-Gb/s. Further advances in transmitters and receivers boosted the bit rate of this system to 10 Gb/s (OC-192), bringing LB_0 to ≈ 1 km-Tb/s.

System ④: *Coherent receiver.* Coherent detection is a technique in which light from a local source (called the local oscillator) is mixed with the received signal light at a photodetector (see Sec. 25.4). This is in contrast to direct detection, in which the intensity of the signal light is directly detected by the photodetector. Though the use of coherent detection enhances receiver sensitivity, thereby allowing greater distances between repeaters, this benefit comes at the expense of increased system complexity. As a result, the commercial implementation of coherent systems lagged behind that of direct-detection systems, particularly because of the enhancement provided by the emergence of the optical fiber amplifier (OFA).

System ⑤: *Optical solitons.* Solitons are short (typically 1 to 50 ps) optical pulses that can travel through long optical fibers without changing the shape of their pulse envelope. As discussed in Sec. 23.5B, the effects of fiber dispersion and nonlinear self-phase modulation (arising, for example, from the optical Kerr effect) precisely cancel, so that the pulses act as if they were traveling through a linear, nondispersive medium. Both erbium-doped and Raman fiber amplification can be effectively used in conjunction with soliton transmission to overcome absorption and scattering losses. In 2002, Deutsche Telekom conducted a trial using Lucent's LAMBDAXTREME DWDM system with dispersion-managed transmission and Raman amplification. The system offered 128 WDM channels, each with a capacity of 10 Gb/s, providing $B_0 = 1.28$ Tb/s and yielding $LB_0 \approx 5120$ km-Tb/s over a fiber length of $L = 4000$ km. However, the dispersion characteristics of existing fiber links are often not well-suited to optical-soliton systems and the technology has not been avidly pursued.

System ⑥: *Optical fiber amplifiers (OFAs).* The advent of the erbium-doped fiber amplifier (EDFA) in 1987 (see Sec. 25.1C) had an extraordinary impact on the performance of optical fiber communication systems. Placed periodically along the fiber, these amplifiers compensate for loss by optically boosting the signal, thus dramatically increasing the distance over which information can be transmitted. The first transpacific link to employ OFAs was the TPC-5 cable network, which comprised four single-mode fibers in the form of a ring that operated at bit rate $B_0 = 10$ Gb/s over a distance of $L = 22\,500$ km, yielding $LB_0 \approx 225$ km-Tb/s.

System ⑦: *Wavelength-division multiplexing (WDM).* WDM systems make use of multiple wavelengths (channels) transmitted through the same fiber, which provides a dramatic increase in system capacity. Broadband optical fiber amplifiers provide simultaneous amplification for all channels. Single-wavelength long-haul links can be upgraded to WDM status by simply replacing the equipment at the ends of the link, while retaining the existing OFAs. Capacity is increased with the addition of dispersion-managed transmission and forward error correction. An example of a WDM link is the TPE transpacific cable network between the U.S. and China completed in 2010. The bit rate for each fiber of this eight fiber-pair cable is 10 Gb/s per channel and B_0 = 5.12 Tb/s. With L = 18 000 km, $LB_0 \approx 92$ km-Pb/s.

System ⑧: *Digital coherent receiver with spectrally efficient coding.* The success of WDM in the 1990s and 2000s was followed by an effort to enhance system spectral efficiency [(b/s)/Hz]. This fueled a resurgence of interest in coherent systems, coupled with the use of WDM and optimized digital coding and signal processing. The first transpacific link employing digital coherent technology was the FASTER submarine cable network, which began service in 2016. This system makes use of six pairs of extremely low-loss fibers without dispersion-compensation sections (digital signal processing compensates for the cumulative dispersion at the end of the cable, thereby avoiding optical-amplifier noise). Each of the twelve fibers operates at 100 Gb/s per channel; with 100 WDM channels the overall bit rate is thus $B_0 = 60$ Tb/s in each direction. Transmission over a distance L = 9000 km between Oregon and Japan yields $LB_0 \approx 540$ km-Pb/s. Making use of the C+L-bands, the 2018 Pacific Light Cable Network (PLCN) offers a capacity $B_0 = 120$ Tb/s, double that of the FASTER system. Stretching from Los Angeles to Hong Kong, a distance L = 12 800 km, the PLCN network attains $LB_0 \approx 1.5$ km-Eb/s.

System ⑨: *Space-division multiplexing (SDM).* With spectral bandwidth exhausted in the service of WDM, and spectral efficiency and phase modulation optimized for digital coherent detection, the only degree of freedom remaining available for increasing system capacity is space. This has led to renewed consideration of exploiting multimode fibers in which distinct spatial-mode distributions serve as independent channels. Also being developed for SDM are multicore fibers in which each core carries a few spatial modes. Multiplexers and demultiplexers, along with digital signal processing techniques to compensate for inherent crosstalk, are being developed. As an example, the availability of 50 spatial channels and 100 WDM channels provides a total of 5000 channels; at 200 Gb/s per channel, a single fiber can support a bit rate of 1 Pb/s.

B. Performance of Optical Fiber Communication Systems

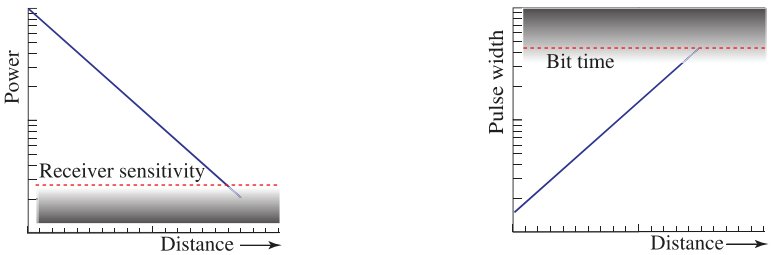
The first step in assessing the performance of an optical fiber communication system is to construct a mathematical model that describes the effect of the various system components, principally the optical fiber, on the modulated signal. This permits the shape of the received distorted signal to be estimated, and hence permits a determination of the signal-to-noise ratio for analog systems and the expected bit error rate for digital systems.

In most applications, the fiber may be treated as a linear system described by an impulse response function $h(t)$ or its Fourier transform, the transfer function $H(f)$, where f is the modulation frequency (see Appendix A). Three important parameters characterize these functions:

- *Power transmission*. This is the fraction of steady (unmodulated) input optical power received at the output. It is given by the transfer function $H(f)$ at $f = 0$; $H(0) = \int h(t)\,dt$ is the area under $h(t)$ since $H(f)$ is the Fourier transform of $h(t)$. For a fiber of length L and attenuation coefficient α (dB/km), $H(0) = \exp(-\alpha L)$, where α is the attenuation coefficient in units of km^{-1}, which is related to α via $\alpha \approx 0.23\alpha$. Localized power losses at couplers may also be included in α in distributed units of dB/km.
- The *response time* σ_τ is the width of $h(t)$. It determines the temporal spreading of optical pulses and therefore sets the maximum data rate that can be used in digital systems. In a single-mode fiber, for example, (25.1-3) provides that $\sigma_\tau = |D|\sigma_\lambda L$, where σ_λ (nm) is the source linewidth and D (ps/km-nm) is the dispersion coefficient. The response time is proportional to the fiber length.
- The *bandwidth* σ_f (Hz) is the width of the transfer function $|H(f)|$. In an analog system, the bandwidth determines the maximum frequency at which the input power may be modulated and successfully detected by the receiver. Since $H(f)$ and $h(t)$ are related by a Fourier transform, the bandwidth σ_f is inversely proportional to the response time σ_τ. The coefficient of proportionality depends on the specific profile of $h(t)$ (see Appendix A, Sec. A.2). Here, we use the relation $\sigma_f = 1/2\pi\sigma_\tau$ for purposes of illustration.

The maximum fiber length that can be used to transmit a signal with a desired performance level is set by the following principal impairments introduced by the system:

- *Attenuation* results in an exponential decrease of the optical power as a function of distance [Fig. 25.2-5(a)]. At a distance for which the received power becomes smaller than the receiver sensitivity (the minimum power required by the receiver), the system's performance becomes unacceptable.
- *Dispersion* results in an increase of the width of the optical pulses that represent data bits in a digital system as a function of distance [Fig. 25.2-5(b)]. When the width exceeds the bit interval, adjacent pulses overlap, resulting in **intersymbol interference** (ISI), which introduces undesirable errors. In an analog system, dispersion washes out high-frequency components of the modulated signal and reduces the system's bandwidth.
- *Noise* added by optical components, such as optical amplifiers, and by random propagation effects, such as polarization mode dispersion, introduces additional errors.
- *Nonlinear distortion* associated with intense optical pulses results in the cross mixing of spectral components, and the introduction of interference between multiplexed signals in wavelength-division multiplexing (WDM) systems.



(a) Attenuation (b) Dispersion

Figure 25.2-5 (a) Dependence of the optical power on distance. (b) Dependence of the pulse width on distance. The maximum length of the optical link is set by either (a) attenuation, when the received power drops below the receiver sensitivity; or (b) dispersion, when the pulse width exceeds the bit time.

The communication system is more sensitive to transmission impairments at high bit rates (or high modulation frequencies) because of the following effects:

- For a fixed average power, a higher bit rate corresponds to fewer photons per bit, and therefore to greater photon noise. Other noise sources in the receiver also become more important at high data rates. The receiver sensitivity is therefore an increasing function of the bit rate [Fig. 25.2-6(a)].
- A higher bit rate corresponds to shorter pulses [Fig. 25.2-6(b)] with broader spectra and greater dispersion. Such pulses undergo greater broadening, which leads to greater intersymbol interference (ISI).
- For a fixed optical energy per bit, a higher bit rate (shorter bit time) requires greater optical power [Fig. 25.2-6(c)], which evokes nonlinear interactions leading to nonlinear ISI.

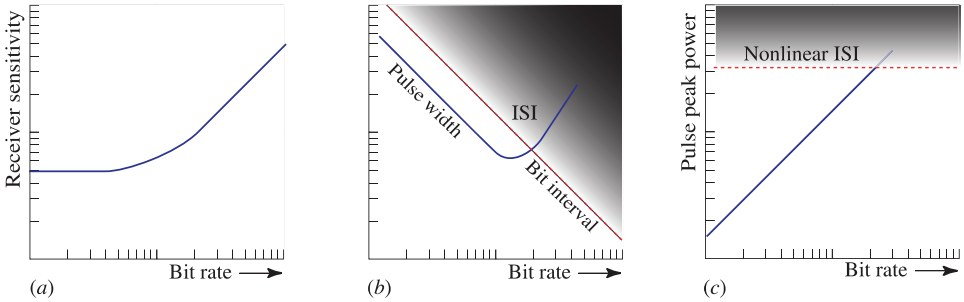


Figure 25.2-6 Effect of bit rate on (a) receiver sensitivity, (b) pulse width at the receiver, and (c) peak power. At higher bit rates, the communication system is more sensitive to attenuation, dispersion, and nonlinear effects.

Bit Error Rate

The performance of a *digital* communication system is measured by the **probability of error** per bit, which is referred to as the **bit error rate (BER)**. For an ON–OFF keying (OOK) system, such as that schematized in Fig. 25.2-1, the logic states "1" and "0" are represented, respectively, by the presence and absence of an optical pulse (see Sec. 19.6E). If p_1 is the probability of mistaking "1" for "0", and p_0 is the probability of mistaking "0" for "1", and if the two bits are equally likely to be transmitted, the BER $= \frac{1}{2}p_1 + \frac{1}{2}p_0$. A typical acceptable BER is 10^{-9}, corresponding to an average of one error every 10^9 bits. Errors occur as a result of noise in the received signal, or they arise from pulse spreading into neighboring bits, which results in intersymbol interference (ISI). Figure 25.2-7 displays an example of random realizations of the pulse corresponding to state "1", superimposed with random realizations of the signal received from possible neighboring pulses when the state is "0", This portrayal is called the **eye diagram**. The more open the "eye," the more distinguishable are the "1" and "0" states and the smaller the likelihood of error.

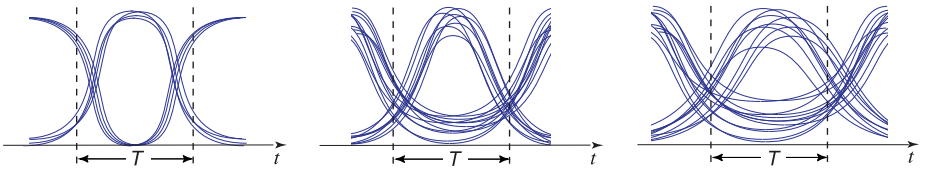


Figure 25.2-7 Closing of the eye diagram (left to right) as a result of noise and pulse broadening.

Receiver Sensitivity

The sensitivity of a digital optical receiver is defined as the minimum number of photons (or the corresponding optical energy) per bit required to guarantee that the rate of error (BER) is smaller than a prescribed value (e.g., 10^{-9}). Errors occur because of randomness in the number of photoelectrons detected during each bit and because of noise in the receiver circuit itself. The sensitivity of digital optical receivers under various conditions has been considered in Sec. 19.6E.

The simplest example of such a receiver is an ON–OFF keying (OOK) optical receiver in which the light source emits Poisson photons, the detector has unity quantum efficiency, and the receiver circuit is noise-free. An average of at least $\overline{n}_0 = 10$ photons per bit is required to achieve a BER $\leq 10^{-9}$ in such a system, so the sensitivity of this ideal optical receiver is 10 photons/bit. This signifies that state "1" should carry an average of at least 20 photons/bit, since state "0" carries zero photons. In the presence of other forms of noise, a larger number of photons per bit is required (Table 19.6-1).

A sensitivity of $\overline{n}_0$ photons per bit corresponds to an optical energy of $h\nu\overline{n}_0$ per bit and an optical power of $P_r = (h\nu\overline{n}_0)/(1/B_0)$ per bit, so that

$$P_r = h\nu\overline{n}_0 B_0, \tag{25.2-1}$$

which is proportional to the bit rate B_0. As the bit rate increases, P_r must increase commensurately to keep the number of photons/bit (and therefore the BER) constant. When circuit noise is important, the receiver sensitivity $\overline{n}_0$ depends on the receiver bandwidth (i.e., on the data rate B_0), as discussed in Sec. 19.6E. This behavior complicates the design problem so for simplicity we assume in the following analyses that the sensitivity (photons per bit) of the receiver is independent of B_0.

It will become apparent in the sequel that the design of long-haul, high-bit-rate optical fiber communication links involves the selection of fibers with the lowest attenuation and/or dispersion, careful power and pulse-width budgeting, and the avoidance of deleterious nonlinear effects associated with ultra-intense pulses.

C. Attenuation- and Dispersion-Limited Systems

We now proceed to examine the performance limits imposed by *attenuation* and *dispersion* on an elementary digital, intensity-modulation, ON–OFF keying (OOK) system. For simplicity, nonlinear effects are ignored and the fiber transmission system itself is assumed to introduce no noise. We consider an optical fiber link operated as a digital communication system at a data rate of B_0 b/s over a distance of L (km). The source has power P_s (mW), wavelength λ_o (nm), and spectral width σ_λ (nm). The fiber has attenuation coefficient α (dB/km) and chromatic dispersion coefficient D_λ (ps/km-nm). The receiver has a sensitivity of $\overline{n}_0$ (photons per bit), corresponding to a power sensitivity $P_r = (hc_o/\lambda_o)\overline{n}_0 B_0$ (mW); these values must be attained for the system to operate at an acceptable error rate.

The performance limits are established by determining the maximum distance L over which the link can transmit B_0 b/s without exceeding the prescribed bit error rate. Clearly, L decreases with increasing B_0. Alternatively, we may determine the maximum bit rate B_0 that a link of length L can transmit with an error rate not exceeding the allowable limit. The maximum bit-rate–distance product LB_0 thus serves as a metric that describes the capability of the link. We shall determine the typical dependence of L on B_0, and derive expressions for the maximum bit-rate–distance product LB_0 for various types of fibers.

Two conditions must be satisfied for acceptable operation of the link:

1. The received power must be at least equal to the receiver power sensitivity P_r. This condition is met by preparing a power budget from which the maximum fiber length is determined. A margin of 6 dB above P_r is usually specified.

2. The width of the received pulses must not significantly exceed the bit time interval $1/B_0$ to avoid overlap of adjacent pulses, which leads to intersymbol interference and increases the error rate. This condition is met by preparing a budget for the pulse spreading resulting from the transmitter, the receiver, and various forms of dispersion in the fiber.

If the bit rate B_0 is fixed and the link length L is increased, two situations leading to performance degradation can occur: the received power becomes smaller than the receiver power sensitivity P_r, or the received pulses become wider than the bit time $1/B_0$. If the former situation occurs first, the link is said to be **attenuation limited**. If the latter occurs first, the link is said to be **dispersion limited**.

Attenuation-Limited Performance: Power Budget

Attenuation-limited performance is assessed by preparing a power budget. Since fiber attenuation is measured in dB units, it is convenient to also measure optical power in dB units. Using 1 mW as a reference, dBm units are defined by

$$\boxed{\mathtt{P} = 10\log_{10} P,} \qquad P \text{ in mW}, \quad \mathtt{P} \text{ in dBm}. \tag{25.2-2}$$

As examples, $P = 0.1$ mW, 1 mW, and 10 mW correspond to $\mathtt{P} = -10$ dBm, 0 dBm, and 10 dBm, respectively. In these logarithmic units, power losses are additive.

If $\mathtt{P}_s$ is the source power (dBm), α is the fiber loss (dB/km), $\mathtt{P}_c$ is the splicing and coupling loss (dB), and L is the maximum fiber length such that the power delivered to the receiver is equal to the receiver sensitivity $\mathtt{P}_r$ (dBm), then

$$\mathtt{P}_s - \mathtt{P}_c - \mathtt{P}_m - \alpha L = \mathtt{P}_r \qquad \text{(dB units)}, \tag{25.2-3}$$

where $\mathtt{P}_m$ is a safety margin. The optical power is plotted schematically in Fig. 25.2-8 as a function of the link distance from the transmitter.

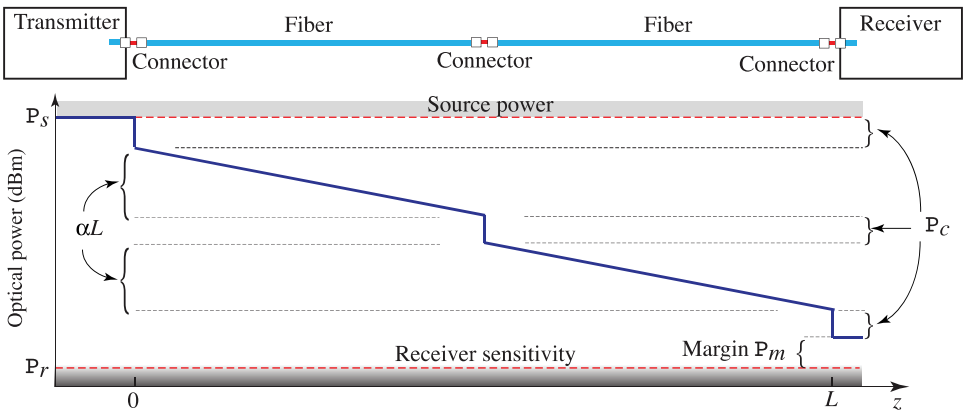


Figure 25.2-8 Power budget for an attenuation-limited fiber-optic link.

The receiver power sensitivity $\mathtt{P}_r = 10\log_{10} P_r$ (dBm) is obtained from (25.2-1):

$$\mathtt{P}_r = 10\log\left(\frac{\overline{n}_0\, h\nu B_0}{10^{-3}}\right) \quad \text{dBm}. \tag{25.2-4}$$

Thus, $\mathtt{P}_r$ increases logarithmically with B_0, and the power budget must therefore be adjusted for the value of B_0, as illustrated in Fig. 25.2-9.

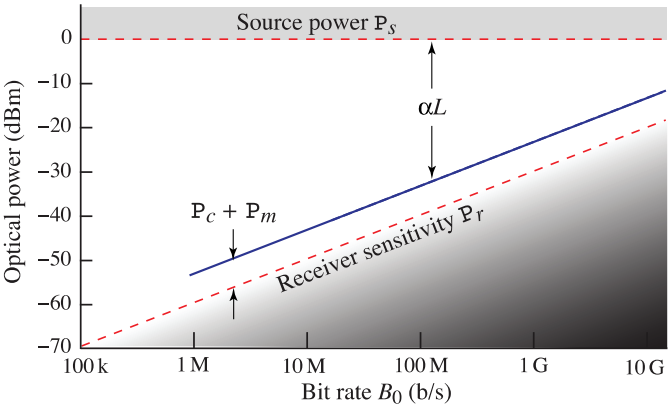


Figure 25.2-9 Power budget as a function of bit rate B_0 for an attenuation-limited fiber-optic link. As B_0 increases, the power P_r per bit required at the receiver increases (so that the energy per bit remains constant), and the maximum fiber length L decreases.

The maximum length of the link is obtained by substituting (25.2-4) into (25.2-3),

$$L = \frac{1}{\alpha}\left(P_s - P_c - P_m - 10\log\frac{\overline{n}_0\, h\nu\, B_0}{10^{-3}}\right), \tag{25.2-5}$$

which yields

$$\boxed{L = L_0 - \frac{10}{\alpha}\log B_0,} \tag{25.2-6}$$

Attenuation-Limited Fiber Length

where $L_0 = [P_s - P_c - P_m - 30 - 10\log(\overline{n}_0\, h\nu)]/\alpha$. The maximum link length decreases with increasing bit rate B_0 at a logarithmic rate with slope $-10/\alpha$. Figure 25.2-10 is a plot of this relation for the operating wavelengths 870, 1300, and 1550 nm.

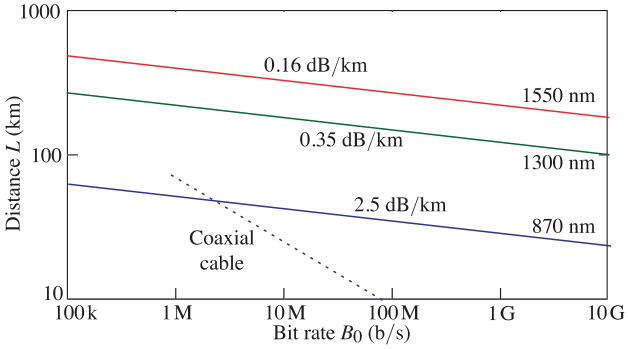


Figure 25.2-10 Maximum fiber length L as a function of bit rate B_0 under attenuation-limited conditions for a fused silica-glass fiber operating at wavelengths $\lambda_o = 870$, 1300, and 1550 nm, assuming fiber attenuation coefficients $\alpha = 2.5$, 0.35, and 0.16 dB/km, respectively. Source power $P_s = 1$ mW ($P_s = 0$ dBm); receiver sensitivity $\overline{n}_0 = 300$ photons/bit for receivers operating at 870 and 1300 nm, and $\overline{n}_0 = 1000$ for the receiver operating at 1550 nm; and $P_c = P_m = 0$. The LB_0 relation for a typical coaxial cable is shown for comparison.

Dispersion-Limited Performance: Time Budget

When a pulse representing a data bit is generated by the transmitter, propagated through the fiber, and detected by the receiver, it loses power and increases in width. The final pulse width σ_o depends on the original pulse width σ_s, the response time of the transmitter σ_{tx}, the response time of the fiber σ_τ imparted by various forms of dispersion, and the response time of the receiver σ_{rx}. The actual shape of the received pulse may be determined by convolving the original pulse profile with the impulse response functions of the transmitter, the fiber, and the receiver (assuming that all systems are linear). Moreover, if all functions are taken to be Gaussian, the square of the width of the final pulse is the sum of the squares of the widths of all the constituent functions, so that

$$\sigma_o^2 = \sigma_s^2 + \sigma_{\text{sys}}^2, \tag{25.2-7}$$

where

$$\sigma_{\text{sys}}^2 = \sigma_{\text{tx}}^2 + \sigma_\tau^2 + \sigma_{\text{rx}}^2\,, \tag{25.2-8}$$

with σ_{sys} representing the width of the response function of the entire communication system (transmitter + fiber + receiver). These relations are used in the practical design of systems even though the response functions are not actually Gaussian.

A principal design condition for the communication link ensures that the width of the received pulse does not exceed a prescribed fraction of the bit period $T = 1/B_0$ to avoid intersymbol interference. A time budget, such as that presented in Fig. 25.2-11, must be prepared to ensure that this condition is met. The choice of that fraction is arbitrary and a number of *ad hoc* values are used. As an example, some designers require that the system's response time σ_{sys} not exceed 70% of the bit period for non-return-to-zero (NRZ) pulses and 35% for return-to-zero (RZ) pulses (these modulation formats are defined in Fig. 25.3-4).

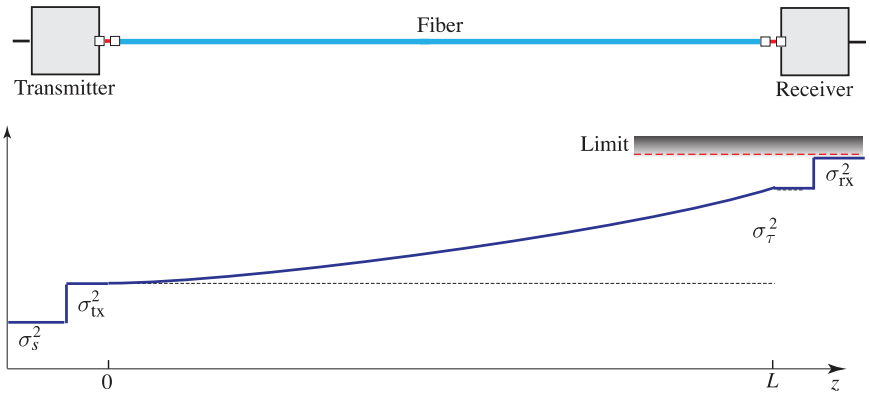


Figure 25.2-11 Budget for the pulse temporal width in a dispersion-limited fiber-optic link.

For a given receiver and transmitter, the design of the link centers around determining the maximum allowable fiber length L. Since the only length-dependent contribution to σ_{sys} is σ_τ associated with the fiber, in the following analysis we adopt a design criterion that limits the value of σ_τ to $1/4$ of the bit-time interval T, i.e.,

$$\sigma_\tau = \tfrac{1}{4}T = 1/4B_0. \tag{25.2-9}$$

The choice of the factor $1/4$ is clearly arbitrary and simply allows us to conveniently compare the different types of fibers. We now consider the distance versus bit-rate relations that arise from this condition for the various dispersion-limited cases considered in Sec. 25.1A. The results are discussed below and plotted in Fig. 25.2-12.

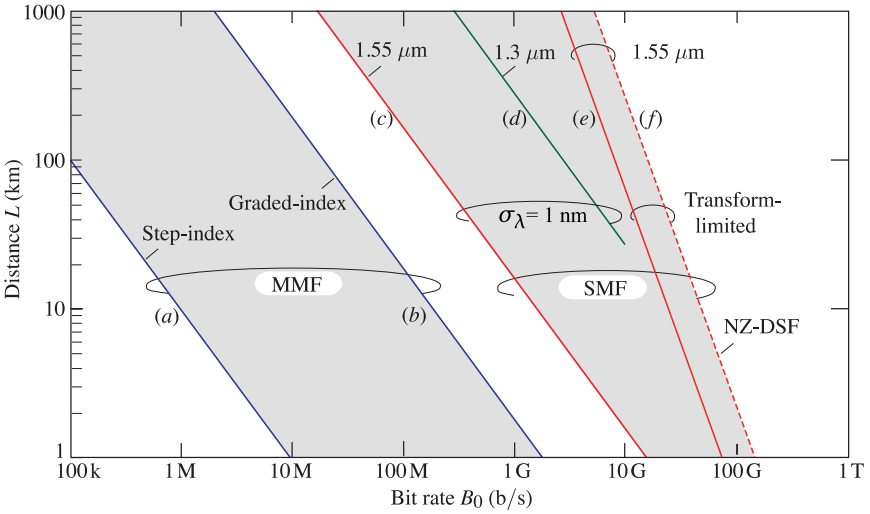


Figure 25.2-12 Dispersion-limited maximum fiber length L as a function of the bit rate B_0 for multimode fibers (MMF) and single-mode fibers (SMF). Six lines are shown (left to right): (a) MMF, step-index ($n_1 = 1.46, \Delta = 0.01$), $LB_0 = 10$ km-Mb/s; (b) MMF, graded-index with parabolic profile ($n_1 = 1.46, \Delta = 0.01$), $LB_0 = 2$ km-Gb/s; (c) SMF limited by material dispersion, operating at 1550 nm with $D_\lambda = 17$ ps/km-nm and $\sigma_\lambda = 1$ nm, $LB_0 \approx 15$ km-Gb/s; (d) SMF limited by material dispersion, operating at 1300 nm with $|D_\lambda| = 1$ ps/km-nm and $\sigma_\lambda = 1$ nm, $LB_0 = 250$ km-Gb/s; (e) SMF with transform limited pulses operating at 1550 nm with $D_\lambda = 17$ ps/km-nm; (f) same as (e) with non-zero dispersion-shifted fiber (NZ-DSF) with chromatic dispersion coefficient $D_\lambda = 4$ ps/km-nm.

- *Multimode fiber (MMF).* For multimode fiber, the width of the received pulse after propagation a distance L is dominated by modal dispersion. For step-index fibers, (25.1-2) and (25.2-9) result in the LB_0 relation

$$\boxed{LB_0 = \frac{c_1}{2\Delta},} \qquad \begin{matrix}(25.2\text{-}10)\\ \text{Step-Index MMF}\end{matrix}$$

  where $c_1 = c_o/n_1$ is the speed of light in the core material and $\Delta = (n_1 - n_2)/n_1$ is the fiber fractional index difference. In a graded-index (GRIN) fiber with an optimal (approximately parabolic) refractive-index profile, the pulse width is smaller by a factor $2/\Delta$, and LB_0 is greater by the same factor. For $n_1 = 1.46$ and $\Delta = 0.01$, the bit-rate–distance product $LB_0 \approx 10$ km-Mb/s for step-index fibers and $LB_0 \approx 2$ km-Gb/s for graded-index fibers.

- *Single-mode fiber (SMF).* Assuming that pulse broadening in a single-mode fiber results from material dispersion only (i.e., neglecting waveguide dispersion), then for a source of linewidth σ_λ, the width of the received pulse is given by (25.1-3), so that

$$\boxed{LB_0 = \frac{1}{4|D_\lambda|\sigma_\lambda},} \qquad \begin{matrix}(25.2\text{-}11)\\ \text{SMF}\end{matrix}$$

where D_λ is the dispersion coefficient of the fiber material. For operation near $\lambda_o = 1300$ nm, $|D_\lambda|$ may be as small as 1 ps/km-nm. Assuming that $\sigma_\lambda = 1$ nm (roughly the linewidth of a single-mode laser diode), the bit-rate–distance product $LB_0 \approx 250$ km-Gb/s. For operation near $\lambda_o = 1550$ nm, $D_\lambda = 17$ ps/km-nm, and for the same source spectral width ($\sigma_\lambda = 1$ nm), we have $LB_0 \approx 15$ km-Gb/s.

- *Single-mode fiber with transform-limited pulses*. To reduce chromatic dispersion, the spectral linewidth σ_λ of the source must be small. Spectral widths that are a small fraction of 1 nm may be obtained by using single-frequency lasers with external modulators. However, an extremely narrow spectral width is incompatible with an extremely short pulse because of the Fourier transform relation between the spectral and temporal distributions. As described in Sec. A.2 of Appendix A, pulses with the smallest product of temporal and spectral widths have a Gaussian profile. Such transform-limited pulses therefore suffer the least dispersion. A transform-limited Gaussian pulse of width τ_0 and complex envelope $\exp(-t^2/\tau_0^2)$ has a Gaussian spectral intensity of width (FWHM) $\sigma_\nu = 0.375/\tau_0$, as provided in (23.1-10). This corresponds to $\sigma_\lambda = |\partial\lambda_o/\partial\nu|\sigma_\nu = (\lambda_o^2/c_o)\sigma_\nu = 0.375\lambda_o^2/c_o\tau_0$. If the pulse has a width equal to half a bit period, i.e., $\tau_0 = T/2 = 1/2B_0$, then

$$\sigma_\lambda = 0.75\frac{\lambda_o^2}{c_o}B_0, \tag{25.2-12}$$

which is directly proportional to the bit rate B_0. For example, for $\lambda_o = 1550$ nm and $B_0 = 10$ Gb/s, $\sigma_\lambda = 0.06$ nm. As described in Sec. 23.3B, when a transform-limited Gaussian pulse of width τ_0 travels through a dispersive medium with dispersion coefficient D_ν, it is broadened by a factor of $\sqrt{2}$ at the characteristic distance $z_0 = \pi\tau_0^2/D_\nu$. In particular, a pulse of initial width $\tau_0 = T/2$ stretches by a time $(\sqrt{2}-1)T/2 \approx 0.21T$. We may therefore take z_0 as the maximum acceptable length L of the communication link. Using the relations $L = z_0 = \pi\tau_0^2/D_\nu$, $\tau_0 = T/2 = 1/2B_0$, and $D_\nu = D_\lambda\lambda_o^2/c_o$, we finally obtain

$$\boxed{LB_0^2 = \frac{\pi}{4}\frac{c_o}{|D_\lambda|\lambda_o^2}.} \tag{25.2-13}$$

Single-Mode Fiber Transform-Limited Pulse

The maximum link distance L is therefore inversely proportional to B_0^2 so that it decreases more rapidly with data rate than in the previous cases. Also, the product B_0L is inversely proportional to the data rate B_0. Figure 25.2-12 displays the LB_0 relation for $\lambda_o = 1550$ nm and $D_\lambda = 17$ ps/km-nm. For example, at B_0 = 10 Gb/s, L = 64 km, but at B_0 = 40 Gb/s, L decreases to 4 km. The use of transform-limited pulses therefore extends the dispersion-limited bit rate bounds substantially, although that rate decreases more rapidly with further increase of the bit rate.

- *Single-mode dispersion-compensated fiber with transform-limited pulses*. With single-mode fibers and transform-limited optical pulses, the maximum fiber distance for a given bit rate reaches its highest value, limited only by the dispersion coefficient. This coefficient can be reduced by making use of dispersion-shifted fibers (DSFs). As shown in Fig. 25.2-12, the use of a DSF to reduce D_λ from 17 ps/km-nm to 4 ps/km-nm increases the maximum link length from 64 km to 272 km at 10 Gb/s. However, DSF fibers have a slightly higher attenuation coefficient.

Combined Attenuation- and Dispersion-Limited Performance

The attenuation-limited and dispersion-limited bit-rate–distance relations portrayed in Figs. 25.2-10 and 25.2-12, respectively, are combined into Fig. 25.2-13 by selecting the smaller of the attenuation- or dispersion-limited distances.

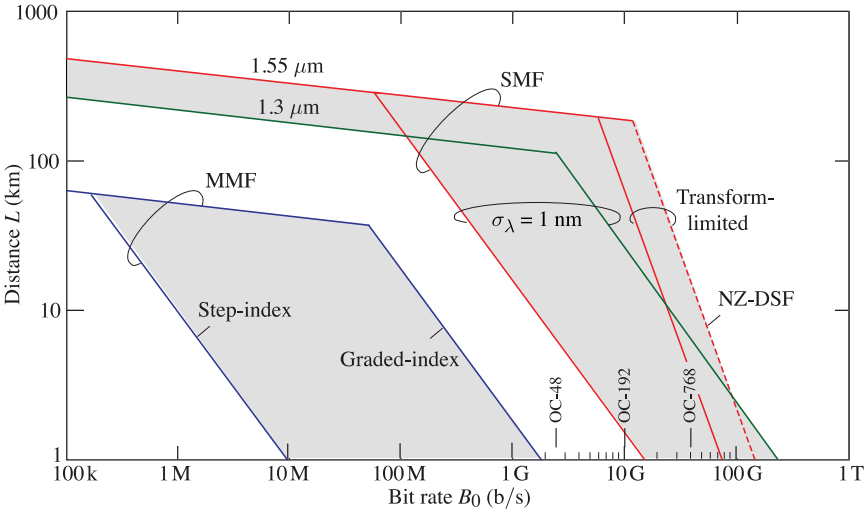


Figure 25.2-13 Maximum fiber-optic link distance L versus bit rate B_0 for a variety of situations. This graph is obtained by superposing the attenuation-limited and dispersion-limited bit-rate–distance curves presented in Figs. 25.2-10 and 25.2-12, respectively, and selecting the smaller of the attenuation- or dispersion-limited distances. Each line in the figure represents the maximum distance L of the link at each bit rate B_0 that satisfies *both* the attenuation and dispersion limits, i.e., that guarantees the reception of the required power *and* pulse width at the receiver.

The graph in Fig. 25.2-13 is the result of an exercise that summarizes the performance of the first three generations of optical fiber communication systems, as portrayed in Fig. 25.2-4: System ① at $\lambda_o = 870$ nm (multimode); System ② at 1310 nm (single-mode); and System ③ at 1550 nm (single-mode). A number of simplifying assumptions and arbitrary choices have been made to create Fig. 25.2-13, so that the values reported therein should be regarded only as indications of the order of magnitude of the relative performance of the different types of fibers. Nevertheless, a number of important conclusions can be drawn from this figure:

- At low bit rates, the fiber link is generally attenuation-limited; the length L decreases logarithmically with B_0. At high bit rates, the link is dispersion limited and L is inversely proportional to B_0 for optical pulses limited by the source linewidth, and inversely proportional to B_0^2 for transform-limited optical pulses.
- For high-data-rate and long-haul communication links, single-mode fibers are superior. The choice between operation at 1300 and 1550 nm is not obvious since, for conventional fibers, chromatic dispersion is minimal at 1300 nm while attenuation is minimal at 1550 nm. This is illustrated by the crossover of the LB_0 lines for these wavelengths.
- The use of dispersion-shifted fibers (DSFs) enables the overall chromatic dispersion coefficient at 1550 nm to be reduced, which in general makes operation at 1550 nm superior to operation at 1300 nm.

Performance of an Analog Communication System

As with digital optical fiber communication links, the performance of an analog link is limited by fiber attenuation and/or dispersion. Because of fiber attenuation, the received signal is weakened and may not be discernible in noise. Because of fiber dispersion, the transmission bandwidth $\sigma_f = 1/2\pi\sigma_\tau$ is limited so that high-frequency signal components are attenuated more than low-frequency components, resulting in signal degradation. Both of these deleterious effects increase with increasing fiber length L. The received optical power decreases exponentially with L, while the fiber bandwidth is inversely proportional to L. Nonlinear optical effects do not play a role in analog systems since the power is distributed rather than concentrated in narrow pulses.

The maximum allowable length of an analog fiber link is determined by ensuring that two conditions are met:

- The fiber attenuation must be sufficiently small so that the received power is greater than the receiver power sensitivity $\mathbb{P}_r$.
- The fiber bandwidth σ_f must be greater than the spectral width B of the transmitted signal.

As discussed in Sec. 19.6, the sensitivity of an analog optical receiver is the smallest optical power required for the signal-to-noise ratio (SNR) of the photocurrent to exceed a prescribed value, SNR_0. For an ideal receiver (with unity quantum efficiency and no circuit noise), $\text{SNR} = \overline{n} = (P/h\nu)/2B$, where B is the receiver bandwidth, P is the optical power (W), and $\overline{n}$ is the average number of photons received in a time interval $1/2B$, which is regarded as the resolution time of the system. If SNR_0 is the minimum allowed signal-to-noise ratio, the receiver sensitivity becomes $\overline{n}_0 = \text{SNR}_0$ photons per receiver resolution time, or its corresponding power

$$P_r = h\nu\overline{n}_0(2B). \tag{25.2-14}$$

This is identical to (25.2-1) for the power sensitivity of an ideal digital receiver if the resolution time of the analog system $1/2B$ is equated with the bit time of the digital system $1/B_0$.

Because of the equivalence between (25.2-14) and (25.2-1), and because of the applicability of the power budget equation (25.2-3) to analog systems as well as digital ones, the LB_0 relations set forth earlier for the ideal binary digital system are also applicable for the analog system, with B_0 replaced by $2B$, provided that the acceptable performance level of the analog system is $\text{SNR}_0 = 10$. As an example, a 1-km fiber link capable of transmitting digital data at a rate of 2 Gb/s with a BER not exceeding 10^{-9} can also be used to transmit analog data of bandwidth 1 GHz with a signal-to-noise ratio of at least 10.

In analog systems, however, the required signal-to-noise ratio is usually far greater than 10, in which case the receiver sensitivity must be far greater than 10 photons per receiver resolution time. For high-quality audio and video signals, for example, a 60-dB signal-to-noise ratio is often required. This corresponds to $\text{SNR}_0 = 10^6$, or $\overline{n}_0 = 10^6$ photons per resolution time. Certain other design considerations are also particularly important in analog systems. For example, nonlinear behavior in the light source and photodetector result in additional signal degradation and restrict the dynamic range of the transmitted waveforms.

D. Attenuation and Dispersion Compensation and Management

Attenuation Compensation

The performance of an attenuation-limited optical fiber communication system may be significantly enhanced by making use of optical fiber amplifiers placed at judicious

locations within the fiber link, as illustrated in Fig. 25.2-14. Amplifiers elevate the diminished optical power, so that the received power can remain above the receiver sensitivity for far greater fiber lengths. Indeed, optical fiber amplifiers are invaluable components in long-haul links, as exemplified by the enhanced performance offered by System ⑥ in Fig. 25.2-4. The extent to which amplification can be used is ultimately limited by the noise introduced by the amplifiers themselves. However, before this limit is reached, dispersion often takes over and the system becomes dispersion-limited. Dispersion compensation is therefore an invaluable adjunct in long-haul optical fiber communication systems that make use of optical amplifiers.

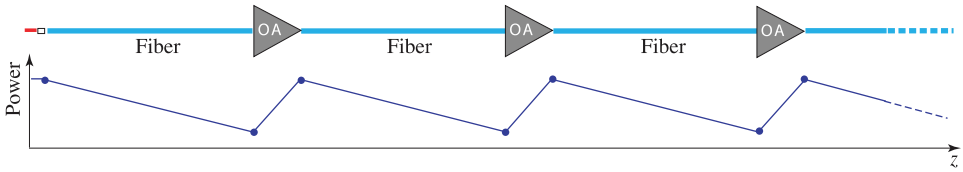


Figure 25.2-14 Compensation of attenuation by use of optical fiber amplifiers.

Dispersion Compensation

The pulse spreading introduced by propagation through an optical fiber of length L and dispersion coefficient D_λ may be reversed by making use of an auxiliary fiber, called a **dispersion-compensating fiber (DCF)**, with dispersion coefficient D'_λ of opposite sign and length L' selected such that the magnitudes of the dispersion introduced by the two fibers are equal, i.e.,

$$D'_\lambda L' = -D_\lambda L. \tag{25.2-15}$$

The pulse spreading and compression introduced by an alternating sequence of such fibers is illustrated in Fig. 25.2-15. The compensating fiber is often relatively short so its dispersion coefficient must be high. Since dispersion in conventional fibers is positive for wavelengths above 1310 nm, the dispersion-compensating fiber must have negative dispersion in this band. This can be achieved by employing dispersion-shifted fibers (DSFs).

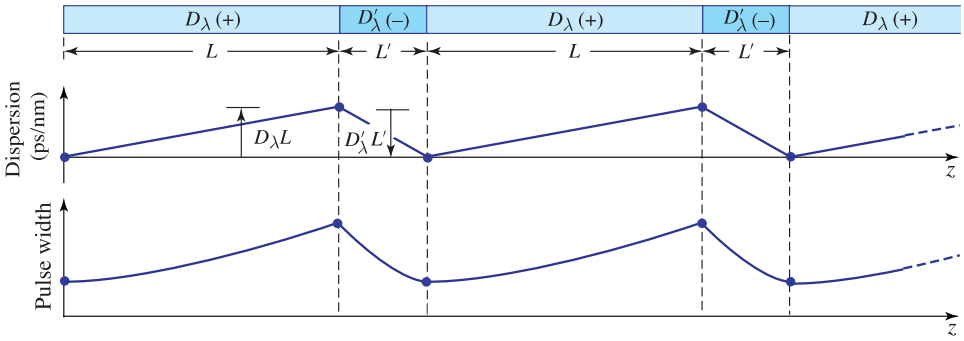


Figure 25.2-15 Dispersion compensation implemented by fiber segments of opposite dispersion.

Other optical components may be used in place of DCFs. As described in Sec. 23.2, the propagation of an optical pulse through a dispersive medium is equivalent to a quadratic chirp filter, which is a phase-only filter with a phase proportional to the square of the frequency. A fiber of length L and dispersion coefficient D_λ behaves as a quadratic chirp filter with chirp coefficient $b = D_\lambda L$. The effect of this filter

may be completely eliminated by making use of an inverse compensation filter — another quadratic chirp filter with a chirp coefficient of equal magnitude and opposite sign, $b' = -b$. The dispersion-compensating fiber plays such a role, but other optical components, such as gratings and interferometers, may also be used for this purpose (see Sec. 23.2).

The compensation filter may be placed at the transmitter end of the link, thus precompensating the dispersion that is subsequently introduced by the fiber. Alternatively, it may be placed at the receiver end, thus postcompensating the broadened pulses immediately before detection. More commonly, multiple compensation filters are placed periodically within the link, providing distributed compensation. Under linear propagation conditions, the actual locations of the compensation filters are not important. However, to avert deleterious nonlinear effects, compensation filters are often placed at locations within the fiber that obviate the presence of short pulses over extended distances.

Broadband Dispersion Compensation: Dispersion Management

For broadband communication systems, such as those that make use of wavelength-division multiplexing (WDM), the condition for dispersion compensation provided in (25.2-15) must be satisfied at all wavelengths within the spectral band; i.e., the error $e_\lambda = D_\lambda L - D'_\lambda L'$ must be zero everywhere. Since the dispersion coefficients are wavelength dependent, this condition is tricky to satisfy. Figure 25.2-16 illustrates a situation for which $e_\lambda = 0$ at wavelength λ_1 in the middle of the band, where the compensation is perfect. However, the positive value of e_λ at wavelength λ_2 corresponds to net positive dispersion while the negative value of e_λ at wavelength λ_3 corresponds to net negative dispersion.

Yet, if both D_λ and D'_λ are approximately linear functions of λ with the same slope, and if $e_\lambda = 0$ at the central wavelength λ_1, then $e_\lambda \approx 0$ everywhere. The design of dispersion-compensating filters with appropriate values of the dispersion coefficient and slope of its wavelength dependence, is known as **dispersion management**. This technique made possible the dramatic increase in system capacity offered by WDM, as evidenced by System ⑦ in Fig. 25.2-4.

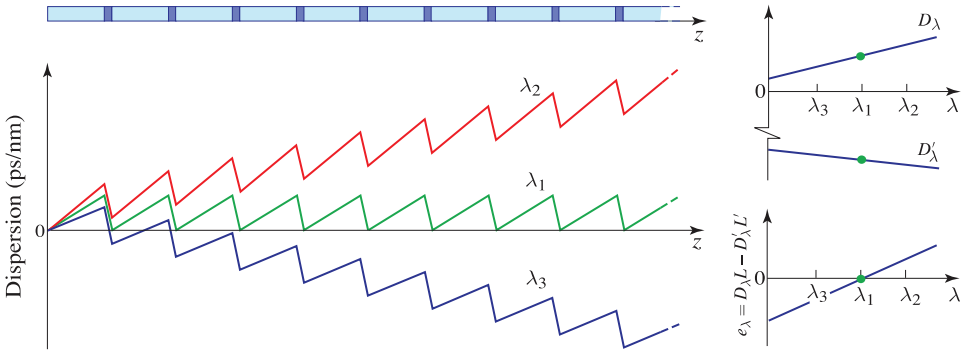


Figure 25.2-16 Perfect dispersion compensation at λ_1; and imperfect dispersion compensation, with net positive and negative dispersion at λ_2 and λ_3, respectively. The error e_λ vanishes if the slopes of D_λ and D'_λ are equal.

Electronic Dispersion Compensation

More recently, advances in digital signal processing and error control coding have made electronic dispersion compensation possible and have emerged as critical elements of optical fiber communication systems. This is particularly true for digital

coherent systems, which employ WDM and rely on optimized digital coding and electronic signal processing rather than on dispersion-compensation sections. Digital coherent communications, shown as System ⑧ in Fig. 25.2-4, is discussed in Sec. 25.4.

E. Soliton Optical Communications

The ultimate in dispersion compensation occurs naturally in optical solitons. These non-spreading pulses have an intensity that is sufficiently high so that the nonlinear optical properties of the fiber play a principal role in their formation. As described in Sec. 23.5B, optical solitons are pulses for which nonlinear dispersion (the dependence of the phase velocity on the intensity via the optical Kerr effect) completely compensates linear chromatic dispersion — the net result is that the pulse travels without altering its width or shape. Moreover, the gain provided by a fiber amplifier can be used to compensate for fiber attenuation so the pulse can maintain its peak intensity and continue to travel as a soliton.

As expressed in (23.5-13), for a pulse of width τ_0, peak intensity I_0, and free-space wavelength λ_o at its central frequency, the condition for soliton formation is

$$\frac{2\pi}{\lambda_o} n_2 I_0 = \frac{-\beta''}{\tau_0^2}, \qquad \text{(25.2-16)}$$

Soliton Condition

where n_2 is the optical Kerr coefficient and $-\beta'' = (\lambda_o^2/2\pi c_o) D_\lambda$ is proportional to the dispersion coefficient D_λ. The intensity profile of the soliton is described by $I(t) = I_0 \,\mathrm{sech}^2(t/\tau_0)$, which is a bell-shaped function with a FWHM of $1.76\,\tau_0$.

In a digital optical communication system, a soliton whose width τ_0 is much smaller than the bit interval T represents state "1", while state "0" is represented by the absence of a soliton. This is necessarily a return-to-zero (RZ) modulation format (Fig. 25.3-4). Soliton communication systems are neither attenuation-limited nor dispersion-limited. Rather, they are limited by nonlinear intersymbol interference that results from the nonlinear interaction between the tails of solitons that represent neighboring bits. When two identical solitons separated by the bit interval T travel a sufficiently long distance through the same fiber, they eventually collapse and merge into a single pulse that subsequently splits again into the original two solitons. In accordance with (23.5-32), this process is repeated periodically with period

$$L_p = \pi e^{r/2} z_0\,, \qquad \text{(25.2-17)}$$

where $r = T/\tau_0$ is the ratio of the separation to the soliton width, and $2z_0 = -\tau_0^2/\beta'' = 2\pi c_o \tau_0^2/\lambda_o^2 D_\lambda$ is the fiber dispersion distance.

The period L_p increases exponentially with the ratio r. If $r \gg 1$, i.e., if the bit interval T is much greater than the soliton width τ_0, then L_p can be made much longer than z_0. If the fiber length L is much smaller than L_p, then the interaction between neighboring bits is minimal. For a fixed ratio r, the condition $L \ll L_p$ may be written in terms of the bit rate $B_0 = 1/T$ as

$$L B_0^2 \ll \frac{\pi^2 c_o}{\lambda_o^2 D_\lambda} \frac{e^{r/2}}{r^2}. \qquad \text{(25.2-18)}$$

This places a limit on the ultimate distances and transmission rates that are permitted. It is important to note that the dispersion characteristics of existing fiber links are often not well-suited to optical-soliton systems so the technology has not been avidly pursued since the mid-1990s, as indicated by System ⑤ in Fig. 25.2-4.

EXAMPLE 25.2-1. ***Soliton Optical-Fiber Communication System.*** A soliton communication system transmits data at 10 Gb/s through a single-mode dispersion-shifted fiber at $\lambda_0 = 1550$ nm using 10-ps (FWHM) soliton pulses. At this wavelength, the dispersion coefficient D_λ= 1 ps/km-nm and the nonlinear refractive index $n_2 = 2.6 \times 10^{-20}$ m^2/W. The fiber effective cross-sectional area is $A_{\text{eff}} = 60\ \mu$m^2. We proceed to determine the source optical power and the maximum length of the link. The 10-ps FWHM pulse width corresponds to a time constant τ_0 = 10/1.76 = 5.7 ps. To satisfy the soliton condition (25.2-16), the peak intensity is $I_0 = 3.75 \times 10^8$ W/m^2, corresponding to a peak power $I_0 A_{\text{eff}} = 22.5$ mW, which must be delivered by the source. The fiber dispersion distance $2z_0 = (2\pi c_o/\lambda_o^2)\tau_0^2/D_\lambda \approx 25$ km. Since the bit interval $T = 1/B_0 = 100$ ps, the ratio $r = T/\tau_0 = 17.6$, and the interaction period provided in (25.2-17) is $L_p \approx 2.1 \times 10^4 z_0$. The fiber length must be much shorter than this length. In this example, (25.2-18) provides $LB_0^2 \ll 26$ m-Tb^2/s^2.

25.3 MODULATION AND MULTIPLEXING

A. Modulation

One way of classifying an optical communication system is in terms of the particular optical variable (intensity, frequency, or phase) that is modulated by the signal to be transmitted. Modulation takes two principal forms: field modulation and intensity modulation. Once the modulation variable is chosen, any of the conventional modulation formats (analog, pulse, or digital) can be implemented.

Field modulation. The field of a monochromatic optical wave serves as a sinusoidal carrier of very high frequency (e.g., 200 THz at $\lambda_o = 1500$ nm). As a means of carrying information, amplitude modulation (AM), phase modulation (PM), and frequency modulation (FM) rely on modulation of the amplitude, phase, and frequency of the field, respectively, in proportionality to the signal (Fig. 25.3-1). Because of the high frequency of the optical carrier, a very wide spectral band is available for the modulation so that a great deal of information can be carried.

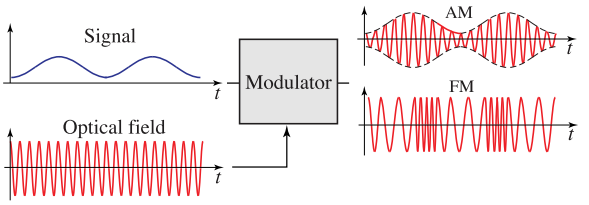


Figure 25.3-1 Amplitude modulation (AM) and frequency modulation (FM) of the optical field.

Though modulation of the optical field is an obvious extension of conventional radiowave and microwave communication techniques to the optical band, it is somewhat difficult to implement for several reasons:

- It requires a source whose amplitude, frequency, and phase are stable and free of fluctuations, e.g., a highly coherent laser.
- Direct modulation of the phase or frequency of a laser is not straightforward; an external modulator that makes use of the electro-optic effect, for example, may be required.
- Because of the assumed high degree of coherence of the source, the use of single-mode fibers is often required; multimode fibers may engender substantial modal noise unless special measures are taken to separate the modes by use of multiplexers/demultiplexers.

- A mechanism for monitoring and controlling the polarization of the field is required unless a polarization-maintaining fiber is employed.
- The receiver must be capable of measuring the magnitude and phase of the optical field; this is often accomplished by making use of a heterodyne or homodyne detection system.

Because of the requirement of a coherent field, optical communication systems that make use of field modulation are termed **coherent communication systems**. These systems are discussed in Sec. 25.4.

Intensity modulation. In an intensity modulation (IM) system, the optical intensity (or power) is imparted to the signal, or a coded version thereof, in a proportional manner, as illustrated in Fig. 25.3-2. The majority of commercial optical fiber communication systems in current use employ intensity modulation because of its relative simplicity and low cost. The optical power of an LED or laser-diode source may be modulated by simply varying the injected drive current. The fiber may be single-mode or multimode. The received optical power impinges on a p–i–n photodiode or an avalanche photodiode (APD) in what is called a **direct-detection receiver**. The high-frequency optical field oscillations play no role in the modulation and demodulation processes; it is the optical *power* that is modulated at the transmitter and demodulated at the receiver. Nevertheless, as we shall see in Sec. 25.3C, the wavelength of the light can serve as a marker that allows many different signals to travel through a single optical link via a process known as wavelength-division multiplexing (WDM).

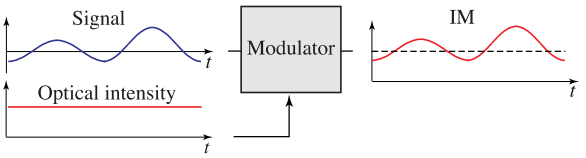


Figure 25.3-2 Intensity modulation (IM).

Digital modulation. An analog signal may be converted into a digital signal by periodically sampling it at an appropriate rate. The resulting samples are then quantized to a discrete finite number of levels, each of which is binary coded and transmitted in the form of a sequence of binary bits, "1's" and "0's", represented by pulses transmitted within the time interval between two adjacent samples (Fig. 25.3-3). This scheme is known as **pulse code modulation** (PCM).

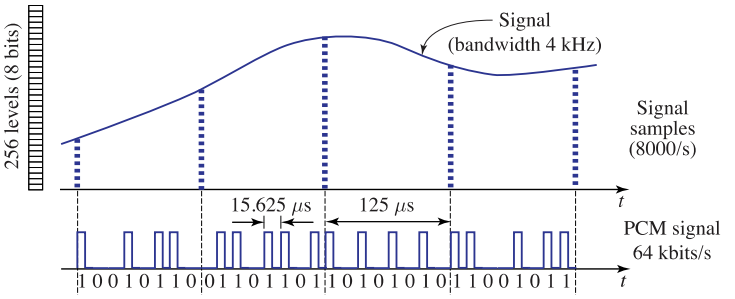


Figure 25.3-3 An example of pulse code modulation (PCM). A 4-kHz voice signal is sampled at a rate of 8×10^3 samples per second. Each sample is quantized to $2^8 = 256$ levels and represented by 8 bits. The original analog signal is thus converted into a sequence of bits transmitted at a rate of 64 kb/s.

In a binary coding system such as that illustrated above, the two states of each bit may be represented by two values of the optical amplitude, phase, or frequency, in which case the modulation scheme is known as **amplitude shift keying** (ASK), **frequency shift keying** (FSK), and **phase shift keying** (PSK), respectively. ASK modulation is also known as **ON–OFF keying** (OOK) when each bit is represented by the presence or absence of a pulse of light. These modulation schemes are illustrated in Fig. 25.3-4. Systems based on optical intensity modulation typically use OOK. However, it is also possible to modulate the optical intensity with a harmonic function that serves as a subcarrier whose frequency or phase is modulated using FSK or PSK. Field modulation with PSK is widely used in wireless networking over short distances, e.g. in wireless local area networks (WLANs), in radio frequency identification (RFID), and in Bluetooth technology.

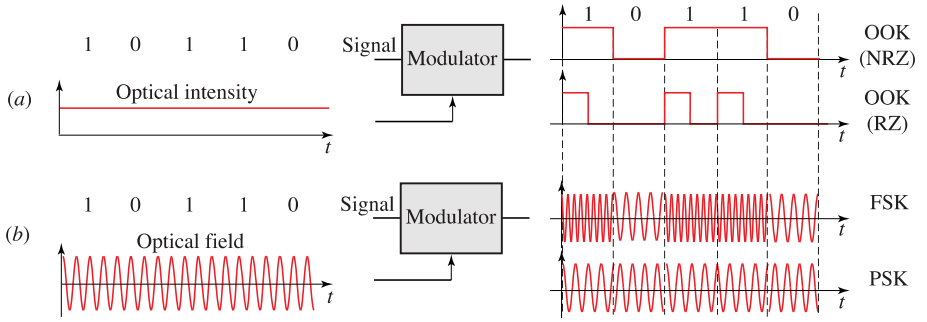


Figure 25.3-4 Examples of the binary modulation of light: (*a*) ON–OFF keying intensity modulation (OOK/IM); (*b*) frequency shift keying (FSK) and phase shift keying (PSK) field modulation.

Advances in high-speed electronics and digital signal processing have enabled **multilevel coding**, which offers higher **spectral efficiency** [(b/s)/Hz] than binary coding. In such systems, each pulse (symbol slot) represents multiple bits. As an example, in binary PSK (BPSK), the phase can assume only two values $(0, \pi)$, representing one bit, whereas in quaternary (or quadrature) PSK (QPSK), four phase values $(0, \pi/2, \pi, 3\pi/2)$ serve to encode two bits, as depicted in the complex-plane diagram depicted in Fig. 25.3-5, which is also called the quadrature diagram. QPSK transmits the same information at half the symbol rate of BPSK. For higher-order PSK, data are coded by a **constellation** of points in the quadrature diagram that are separated by equal angles. Other variations of PSK include **differential PSK** (where it is the data *change*, rather than the data, that sets the phase. The differential approach avoids any ambiguity that might be imparted by an unintended rotation of the constellation introduced in the transmission channel. Differential QPSK is denoted DQPSK.

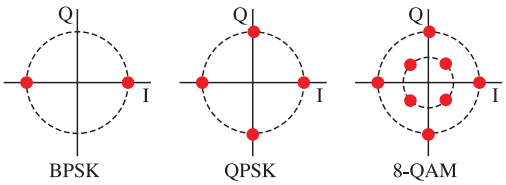


Figure 25.3-5 Constellations for BPSK (1 bit), QPSK (2 bits; also called 4-QAM), and 8-QAM (4 bits). The symbols I and Q denote the in-phase and quadrature components of the complex amplitude, which are the real and imaginary components, respectively.

Combined amplitude and phase shift keying permits a larger number of bits per symbol to be attained, as may be understood from the constellations of points in the complex plane shown in Fig. 25.3-5. The terminology **quadrature amplitude modulation** (QAM) is used to indicate that each of the quadratures of the optical field

may assume multiple amplitudes. Thus, 4-QAM is the same as QPSK since each of the two quadratures has two values of the amplitude. The constellation of the 8-QAM system, which is a 4-bit system, is also provided in Fig. 25.3-5. Constellations for 16-QAM, and so on, are constructed in the same manner.

B. Multiplexing

Multiplexing enables the transmission and retrieval of more than one signal through the same communication link, as illustrated in Fig. 25.3-6. This is accomplished by marking each signal with a distinct physical label or a code that may be identified at the receiver. We begin this section with a discussion of **electronic multiplexing**, which comprises three standard schemes: **frequency-division multiplexing** (**FDM**), **time-division multiplexing** (**TDM**), and **code-division multiplexing** (**CDM**).

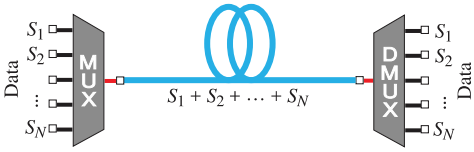


Figure 25.3-6 Transmission of N signals through the same channel by use of a multiplexer (MUX) and a demultiplexer (DMUX).

Frequency-division multiplexing (FDM). In FDM, carriers of distinct frequencies are modulated by a collection of signals. At the receiver, the individual signals are identified by making use of filters tuned to the carrier frequencies, as illustrated in Fig. 25.3-7(a).

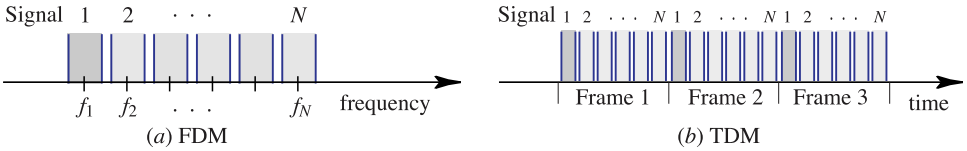


Figure 25.3-7 (a) In frequency-division multiplexing (FDM), a spectral band centered about a distinct frequency is allocated to each signal. (b) In time-division multiplexing, a sequence of time slots is allocated to each signal. The time slots of different signals are interleaved.

Time-division multiplexing (TDM). In TDM, data are transmitted in a sequence of time frames, each with a set of time slots allocated to bits or bytes of the different signals, as illustrated in Fig. 25.3-7(b). These bits must be synchronized to the same clock. At the receiver, each signal is identified by its time-slot location within the frame. An example of a hierarchical TDM system is the T-system illustrated in Fig. 25.3-8.

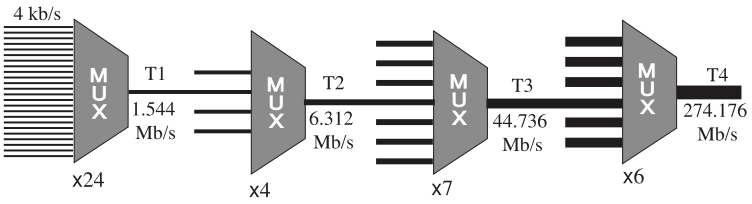


Figure 25.3-8 Hierarchy of the T-system originally developed by the Bell telephone system for transmitting voice signals via time-division multiplexing. A set of 24 4-kb/s signals are multiplexed by a time-division multiplexer that generates a T1 composite signal at 1.544 Mb/s. Four such signals are multiplexed to generate a T2 signal, and so on, as illustrated in the figure.

Code-division multiplexing (CDM). In CDM, each signal is assigned an address code (or key) in the form of a unique function of time defined within the bit period. The code can be a sequence of one/zero bits generated at a rate far higher than that of the original data. Codes of different signals must be uncorrelated (orthogonal) so that they can be separated at the receiver by use of a correlator. In the particular encoding scheme illustrated in Fig. 25.3-9, each bit "1" of the original data is replaced with the code sequence. Each receiver correlates its own code with that of the received signal, thereby locking it only to those bits associated with its own code and permitting it to disregard all other bits.

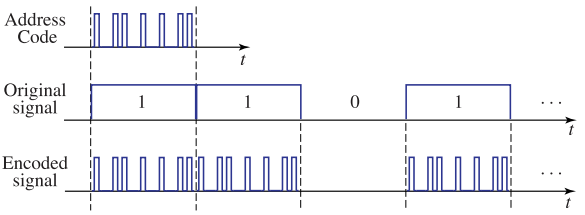


Figure 25.3-9 An example of CDM encoding.

Electronic versus optical multiplexing. As portrayed in Fig. 25.3-10, multiplexing may be electronic or optical. In **electronic multiplexing**, as discussed above, the signals are multiplexed via FDM, TDM, or CDM to generate a composite electronic signal that is used to modulate the light source via any of the optical modulation schemes discussed in Sec. 25.3A. For example, an FDM electronic signal may be generated by making use of a set of carrier frequencies, called "subcarriers," to modulate the intensity of the light source (IM). At the receiver, the light is detected and the demultiplexing is carried out by using electronic filters. In another example, a TDM electronic signal such as the T4 signal shown in Fig. 25.3-8 intensity modulates the light source; demultiplexing of the detected signal is accomplished electronically.

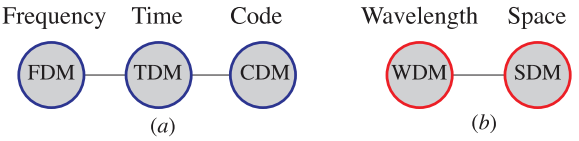


Figure 25.3-10 (*a*) Electronic multiplexing. (*b*) Optical multiplexing.

In **optical multiplexing**, on the other hand, the labels distinguishing the multiplexed signals are optical in nature. In optical FDM, for example, a collection of optical frequencies are used as the carriers of the various signals. These frequencies are separated at the receiver by means of optical filters. When the carrier frequencies in optical FDM are sufficiently widely spaced (say, greater than 20 GHz), this form of optical FDM has come to be called **wavelength-division multiplexing** (WDM), as discussed in Sec. 25.3C. Systems employing WDM are extensively used since they allow the capacity of an existing fiber network to be expanded without the necessity of deploying additional fiber. Another form of optical multiplexing is **space-division multiplexing** (SDM), discussed in Sec. 25.3D, in which the labels that distinguish the multiplexed signals are associated with the spatial modes of optical fibers, including polarization modes.

C. Wavelength-Division Multiplexing

A **wavelength-division multiplexing (WDM)** system makes use of a collection of light sources of different wavelengths, each of which is intensity modulated by a different electrical signal. The modulated light beams are combined and launched into a fiber using an optical multiplexer (OMUX). Demultiplexing is implemented at the receiver with the help of an optical demultiplexer (ODMUX), which serves to separate the different wavelengths and direct them to different detectors. Optical multiplexers and demultiplexers were described in Sec. 24.2A. The electrical signal associated with each wavelength is often an electronically multiplexed set of other signals, so that electronic demultiplexing is required at the receiver. The overall system schematic of such a configuration is illustrated in Fig. 25.3-11.

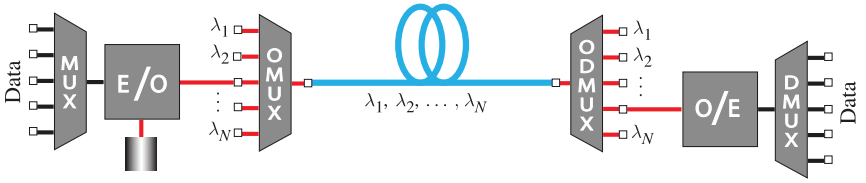


Figure 25.3-11 Wavelength-division multiplexing (WDM). A set of electronically multiplexed data-carrying signals are converted into an optical signal by modulating an optical source of a particular wavelength (E/O). Collections of such modulated optical sources at different wavelengths are optically multiplexed (OMUX) and launched into a single optical fiber. At the receiver end, the signals are optically demultiplexed (ODMUX). The optical signal at each wavelength is converted into an electrical signal by use of a detector/demodulator (O/E) and then electronically demultiplexed.

The spectral bands used in modern optical fiber communication systems are displayed in Fig. 25.3-12. WDM systems can make use of any combination of wavelengths within these bands. The spacing between the wavelengths of the different channels must be greater than the spectral widths of the modulated light in each channel, which are determined by the linewidths of the light sources as well as by the bandwidths of the data carried by the channels. The channel spacing must also be large enough to permit optical multiplexing and demultiplexing with minimal crosstalk among channels.

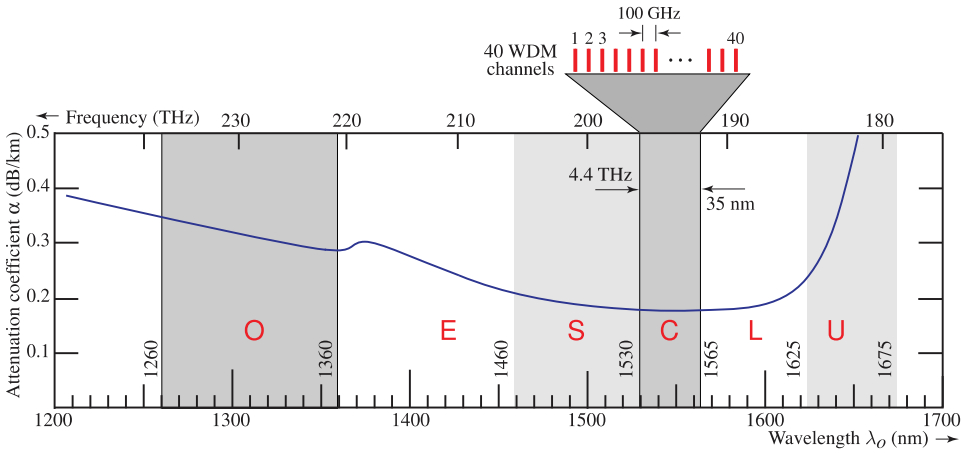


Figure 25.3-12 A 40-channel DWDM system with channel spacings of 100 GHz in the C spectral band. The curve represents the attenuation coefficient (dB/km) of silica-glass fibers with suppressed OH absorption. Fiber attenuation is minimized in the wavelength region corresponding to the C band. The bands commonly used in WDM systems are denoted O = Original, E = Extended, S = Short, C = Conventional, L = Long, and U = Ultra-long, as detailed in Fig. 25.1-2.

WDM systems are classified into two categories, coarse and dense, depending on the number of channels and the channel spacing.

- **Coarse WDM (CWDM)** systems use a few channels with widely spaced wavelengths (20 nm or more). They are typically used in short-range communications and do not make use of amplification. An example is a system with two wavelengths, one at 1310 nm and another at 1550 nm. CWDM is used in cable television networks, where different wavelengths are used for the downstream and upstream signals. The Ethernet LX-4 physical layer standard provides another example in which four wavelengths near 1310 nm are used, each carrying a 3.125-Gb/s data stream. The standardization of 100G and 400G ethernet is upon us. Metropolitan networks employ CWDM systems with a 20-nm wavelength spacing.
- **Dense WDM (DWDM)** systems have a large number of channels (generally more than 16) with closely spaced wavelengths. Such systems are used in long-haul transmission and often make use of amplification. DWDM tends to be used at a higher level (and at higher data rates) in the communications hierarchy, for example, on the internet backbone. Implementation requires the use of precision lasers to prevent wavelength drift. At a wavelength of 1550 nm in the C-band, for example, a frequency spacing of $\Delta\nu = 200$ GHz corresponds to a wavelength spacing $\Delta\lambda = (\lambda_o^2/c_o)\Delta\nu = 1.6$ nm. DWDM systems use channel spacings as small as 50 GHz, or sometimes even 20 GHz, corresponding to wavelength spacings of 0.4 nm and 0.16 nm, respectively. As illustrated in Fig. 25.3-12, the width of the C-band is 35 nm, or approximately 4.4 THz. This can hold 40 channels with a 100-GHz (0.8 nm) spacing. More channels may be accommodated by expanding beyond the C-band and reducing the channel spacing. As an example, making use of the C+L-bands, whose combined width is ≈ 9 THz, and channel spacings of 20.3 GHz, will accommodate 441 channels.

WDM for wideband multimode fibers at 800–900 nm accommodates short-haul applications such as data communications at rates exceeding 100 Gb/s. The MMFs use GRIN profiles optimized to reduce modal and chromatic dispersion, thereby doubling the 10 Gb/s rate of the OM4 fiber. VCSELs with few emission lines in the 800–900-nm band are used as sources for few-channel WDM systems.

D. Space-Division Multiplexing

The data rate carried by a fiber-optic cable may be increased by making use of a fiber bundle (FB) within the same cable or by making use of a **multicore fiber (MCF)** comprising multiple, well-separated cores within the same fiber cladding, each providing an independent communication channel [Fig. 25.3-13(a)]. Cables can contain tens or hundreds of fibers. However, such configurations require multiple systems for transmission, amplification, and reception, as well as multiple splices that require a great deal of mechanical precision in the event of a cable break. Moreover, they typically need to be deployed from scratch.

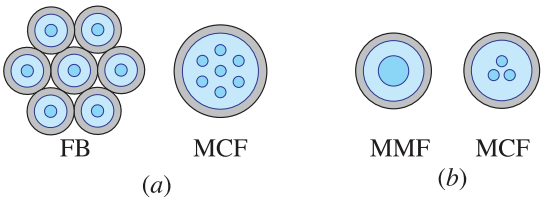


Figure 25.3-13 (a) Multiple communication channels in a fiber bundle (FB) or in a multicore fiber (MCF) with well-separated cores. (b) Space-division multiplexing (SDM) using spatial modes of a multimode fiber (MMF) or modes of a MCF with coupled cores.

It is worth considering whether a more compact configuration might be envisioned while reducing the downside. It turns out that data rates can indeed be increased by implementing **space-division multiplexing (SDM)** using either the spatial modes of one multimode fiber (MMF) or the coupled-core modes of a multicore fiber (MCF) [Fig. 25.3-13(*b*)]. The spatial modes of a MMF can provide independent communication channels, but implementing such a system requires the use of multiplexing and demultiplexing since the modes occupy a common physical volume. Similarly, a MCF whose cores are in close physical proximity can be used for data transmission, but this too requires the use of multiplexing and demultiplexing since the set of coupled cores support supermodes that occupy a common region of space. We proceed to discuss the implementation of SDM in a multimode fiber.

SDM in Multimode Fibers

As described in Sec. 10.2A, a step-index MMF of core radius a and numerical aperture NA has $M \approx \frac{1}{2}V^2$ modes, where $V = (2\pi a/\lambda_o)\,\mathrm{NA}$ is the fiber V parameter. As is evident from Fig. 10.2-5, a fiber with $V = 10$ supports approximately 50 modes. The optical field is generally expressed as a weighted superposition of these modes

$$U(r,\phi,z) = \sum_{q=1}^{M} a_q\, u_q(r,\phi)\, e^{-j\beta_q z}, \tag{25.3-1}$$

where a_q, β_q, and $u_q(r,\phi)$ are, respectively, the complex amplitude, propagation constant, and transverse spatial distribution of the qth mode. For simplicity, we use the integer $q = 1, 2, \ldots, M$ in lieu of the mode double indices (l, m). The amplitudes $\{a_q\}$, which carry the transmitted information in a SDM system, are imparted to the MMF by a multiplexer at the transmitter $(z = 0)$ and extracted from the total received optical field by a demultiplexer at the receiver terminus $(z = L)$, as illustrated in Fig. 25.3-14.

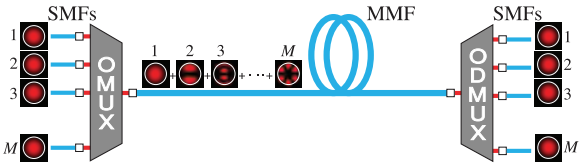


Figure 25.3-14 Space-division multiplexing (SDM) in a multimode fiber (MMF) using an optical multiplexer (OMUX) and an optical demultiplexer (ODMUX). The signals on M single-mode optical fibers (SMFs) are multiplexed into M modes of the MMF.

Since the modes are orthonormal, i.e., $\iint u_q^*(r,\phi)\, u_{q'}(r,\phi)\, r dr d\phi = \delta_{qq'}$, the amplitudes of the individual modes may be calculated from the total field at $z = L$ by use of the projections (see Appendix C)

$$a_q\, e^{-j\beta_q L} = \iint U^*(r,\phi,L)\, u_q(r,\phi)\, r dr d\phi\,. \tag{25.3-2}$$

Therefore, if the total optical field $U(r,\phi,L)$ were measured at the terminus, in both its magnitude and phase, the complex amplitudes $\{a_q\}$ of the modes would be determined. However, such a measurement requires a fast coherent wavefront sensor, which is not available, so we resort to less direct approaches.

SDM multiplexers and demultiplexers. The optical multiplexer (OMUX) and demultiplexer (ODMUX) for SDM may be implemented by selective optical couplers, assisted by computational signal-processing tools, as described by the following examples:

- **Mode conversion.** An optical system that implements the OMUX depicted in Fig. 25.3-14 converts the incoming SMF optical beams, all of which have the spatial profile of the fundamental mode and amplitudes $\{a_q\}$, into beams whose spatial profiles are proportional to those of the MMF modes $u_q(r, \phi)$. As illustrated in Fig. 25.3-15(a), the beams are optically combined into a single beam corresponding to the superposition in (25.3-1) and coupled into the MMF using a single input coupler. Since each spatial profile is excited by only the matching input profile, the modes of the MMF acquire amplitudes proportional to $\{a_q\}$. The key element of this multiplexer is the mode converter, which is implemented by means a phase plate, a spatial light modulator, or a hologram (see Sec. 4.5). Though it offers a didactic example, this type of multiplexer introduces high optical loss and crosstalk.
- **Multiple directional couplers with matching propagation constants.** A directional coupler couples the mode of a SMF to a mode of a MMF if the propagation constants are matched (see Sec. 9.4B), even if the modal spatial profiles differ. An optical multiplexer/demultiplexer based on this principle may be implemented by using a bank of directional couplers of different dimensions designed to selectively launch a particular mode of the MMF for each of the incoming/outgoing SMFs, as illustrated in Fig. 25.3-15(b).
- **Single fused coupler.** A single fiber coupler with M identical SMFs at its input, and one MMF supporting M modes at its output, may be used as a multiplexer/demultiplexer. Each of the SMFs carries light with the profile of its fundamental mode, but is positioned in relation to the output MMF such that it excites a small group of MMF modes in accordance with a coupling matrix. An example of a fused coupler is the **photonic lantern** described in Sec. 10.2D and displayed in Fig. 25.3-15(c). Since the fused coupler is reciprocal it can be used in reverse as a demultiplexer. Each output SMF then receives light from a group of modes of the MMF in accordance with another coupling matrix.

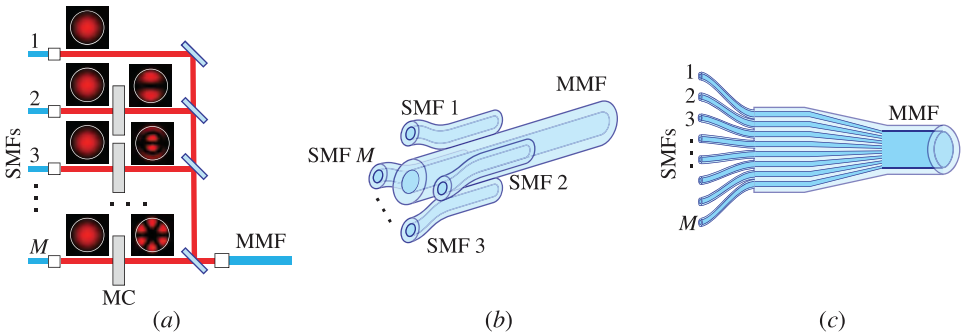


Figure 25.3-15 Optical multiplexer (OMUX) examples for space-division multiplexing (SDM). (a) OMUX using mode converters (MC) and beam combining by means of beam splitters. (b) OMUX using multiple directional couplers with propagation constants matched to the MMF modes. (c) Photonic lantern OMUX. The multiplexers in (b) and (c) can be used in reverse as demultiplexers.

Crosstalk and MIMO systems. Since multiplexers and demultiplexers are not perfect, the complex amplitudes of the modes detected by the receivers do not have one-to-one correspondences with those at the transmitter, and are instead described by coupling matrices at both ends. Additional mode coupling is introduced by the fiber itself as a result of slight variations along its length caused by fabrication errors or by micro and macro fiber bending. Small distortions of the circular symmetry of the fiber cross-section also result in coupling between the degenerate modes, which have equal propagation constants under ideal conditions. Mode coupling is a challenging obstacle to implementing SDM although it can be overcome.

In the absence of nonlinear optical effects, the modal amplitudes $\{b_q\}$ at the receiver output are related to the modal amplitudes $\{a_q\}$ at the transmitter input by the linear relation

$$b_q = \sum_{p=1}^{M} \mathcal{C}_{qp} a_p, \quad q = 1, 2, \ldots, M, \tag{25.3-3}$$

where the $\mathcal{C}_{qp}$ are the elements of an $M \times M$ coupling matrix. This matrix is in turn the product of three matrices: the coupling matrix of the multiplexer, the modal coupling matrix of the fiber, and the coupling matrix of the demultiplexer. While the multiplexer and demultiplexer coupling matrices can be measured, the modal coupling matrix of the fiber has intrinsic randomness and can only be described statistically. Moreover, elements of the overall coupling matrix are frequency-dependent because of modal and material dispersion, so that they represent a dynamical system that includes differential modal delay. Nevertheless, these effects may be compensated by making use of signal-processing tools similar to those employed in wireless systems, which are subject to multiple paths among transmitters and receivers, and are known as multiple-input multiple-output (MIMO) systems.

SDM in multicore fibers. MCF SDM is implemented by making use of a similar approach. If each of the cores supports a few modes, then the system may be described in terms of a matrix representing a combination of intra-core mode coupling and inter-core coupling. For strong core coupling, the overall system may be also be described in terms of supermodes (see Sec. 10.2D).

25.4 COHERENT OPTICAL COMMUNICATIONS

Coherent optical communication systems make use of *field* modulation (amplitude, phase, or frequency) rather than intensity modulation. They employ coherent light sources, single-mode or multimode fibers, and heterodyne or homodyne optical receivers. Coherent optical fiber communication systems were initially pursued in the late 1980s because of their superior sensitivity but were largely bypassed in the early 1990s because of the invention of the erbium-doped fiber amplifier, which enabled the sensitivity of direct-detection systems to come within a few dB of that of coherent systems.

The resurgence of interest in coherent optical fiber systems, fostered by advances in high-speed electronics and digital signal processing, has led to increased spectral efficiency and enhanced transmission capacity. We proceed to examine the principles of operation of such systems and to compare their performance with that of direct-detection systems. We also touch on the requirements for components employed in coherent systems. Coherent systems also find application in free-space optical communications.

As discussed in Sec. 19.1B, photodetectors are responsive to the photon flux and, as such, are insensitive to optical phase. Nevertheless, it is possible to measure the complex amplitude (both magnitude and phase) of a signal optical field by mixing it with a coherent reference optical field of stable phase, called a **local oscillator (LO)**. As illustrated in Fig. 25.4-1, the two waves are superposed (mixed) before impinging on the photodetector. As a consequence of interference of the two fields, information pertaining to both the amplitude and phase of the signal field is registered in the detected electric current.

This detection technique is known as **coherent detection**, in contrast to **direct detection** (**DD**), which was considered earlier in this chapter. Coherent optical detection is also called **optical mixing**, **photomixing**, or **light beating** (see Sec. 2.6B). The coherent optical receiver is the optical equivalent of the superheterodyne radio receiver. When the signal and local-oscillator waves have different frequencies (ω_s and ω_L, respectively), the technique is called **optical heterodyning** whereas it is referred to as **optical homodyning** when $\omega_s = \omega_L$.

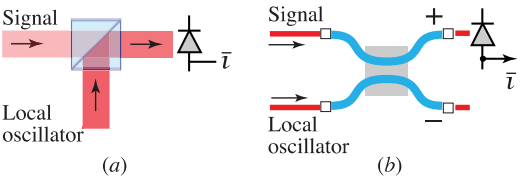


Figure 25.4-1 Coherent optical detection. A signal wave of frequency ω_s is mixed with a local oscillator wave of frequency ω_L using (a) a beamsplitter, or (b) an optical coupler. The detector photocurrent varies at the difference frequency $\omega_I = \omega_s - \omega_L$.

Heterodyne Receiver

Let $\mathcal{E}_s = \text{Re}\{E_s \exp(j\omega_s t)\}$ be the signal optical field, with $E_s = |E_s| \exp(j\varphi_s)$ its complex amplitude and ω_s its angular frequency. The magnitude $|E_s|$ (or the phase φ_s) is modulated by the signal at a rate much slower than that of ω_s. The local oscillator field is similarly described by $\mathcal{E}_L$, E_L, ω_L, and φ_L. The two fields are mixed using a beamsplitter or an optical coupler, as depicted in Fig. 25.4-1, so that the total field is the sum of the constituent fields: $\mathcal{E} = \mathcal{E}_s + \mathcal{E}_L$. If the incident fields are perfectly parallel plane waves and have the same polarization, the spatial dependence need not be carried along in the calculations. Taking the absolute square of the sum of the two complex waves then leads to

$$|E_s e^{j\omega_s t} + E_L e^{j\omega_L t}|^2 = |E_s|^2 + |E_L|^2 + 2|E_s||E_L| \cos(\omega_I t + \varphi_s - \varphi_L), \quad (25.4\text{-}1)$$

where $\omega_I = \omega_s - \omega_L$ is the difference frequency, also called the **intermediate frequency**. Since the intensities I_s, I_L, and I are proportional to the absolute-squared values of the complex amplitudes, we arrive at

$$I = I_s + I_L + 2\sqrt{I_s I_L} \cos(\omega_I t + \varphi_s - \varphi_L), \quad (25.4\text{-}2)$$

which accords with (2.6-12). From a photon-optics point of view, this process can be understood in terms of the detection of polychromatic (two-frequency) photons (see Prob. 13.1-11).

The optical power P at the photodetector is the integral of the intensity over the detector area, so that

$$P = P_s + P_L + 2\sqrt{P_s P_L} \cos(\omega_I t + \varphi_s - \varphi_L), \quad (25.4\text{-}3)$$

where P_s and P_L are the powers of the signal and the LO beams, respectively. Misalignment between the directions of the two waves washes out the interference term

[the third term of (25.4-3)], since the phase $\varphi_s - \varphi_L$ then varies sinusoidally with position across the area of the detector. As is readily understood from Fig. 2.5-4, this undesirable result can be avoided by keeping the angle θ between the wavefronts sufficiently small, such that $\theta \ll \lambda/a$, where a is the size of the photodetector aperture.

If the signal and local oscillator beams are sufficiently close in frequency, their difference ω_I will be many orders of magnitude smaller than the individual frequencies ω_s and ω_L. The superposed light is then quasi-monochromatic and the total photon flux $\Phi = P/h\overline{\nu}$ is proportional to the optical power, where $\overline{\nu} = \overline{\omega}/2\pi$ and $\overline{\omega} = \frac{1}{2}(\omega_s + \omega_L)$. In accordance with (19.1-4), the photocurrent i generated in a photodetector is proportional to the incident photon flux Φ via $i = \eta e \Phi$, where e is the electron charge and η the detector quantum efficiency. The mean photocurrent is hence $\overline{i} = (\eta e/h\overline{\nu})P$, which provides

$$\overline{i} = \overline{i}_s + \overline{i}_L + 2\sqrt{\overline{i}_s \overline{i}_L}\cos\big(\omega_I t + \varphi_s - \varphi_L\big). \tag{25.4-4}$$

Photomixing Current

Here $\overline{i}_s = \eta e P_s/h\overline{\nu}$ and $\overline{i}_L = \eta e P_L/h\overline{\nu}$ are the photocurrents generated by the signal and LO individually.

The local oscillator is usually made much stronger than the signal, in which case the first term in (25.4-4) can be neglected. The second term is constant and the useful information is carried by the third term, which oscillates at the difference frequency ω_I. With knowledge of $\overline{i}_L$ and φ_L, the amplitude and phase of this term can be determined, and $\overline{i}_s$ and φ_s estimated, from which the intensity and phase (and hence the complex amplitude) of the measured optical signal can be inferred. The information-containing signal variables $\overline{i}_s$ or φ_s are usually slowly varying functions of time in comparison with ω_I, so they act as slow modulations of the amplitude and phase of the harmonic function $2\sqrt{\overline{i}_L}\cos(\omega_I t - \varphi_L)$, respectively. The amplitude- and phase-modulated current can be demodulated by drawing on the conventional techniques used in AM and FM radio receivers.

Balanced Homodyne Receiver

The homodyne system is a special case of the heterodyne system for which $\omega_s = \omega_L$ and $\omega_I = 0$. The demodulation process is different for homodyning than for heterodyning, however. For the homodyne system, a phase-locked loop is used to lock the phase of the LO so that $\varphi_L = 0$, whereupon (25.4-4) yields

$$\overline{i} = \overline{i}_s + \overline{i}_L + 2\sqrt{\overline{i}_s \overline{i}_L}\cos\varphi_s. \tag{25.4-5}$$

A **balanced homodyne receiver**, also called a **balanced mixer**, is a coherent receiver designed to cancel the first two terms of (25.4-5). As illustrated in Fig. 25.4-2, the optical fields of the signal and LO are mixed at a beamsplitter (or directional coupler).

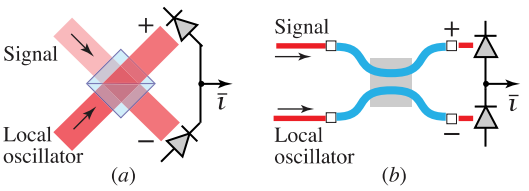


Figure 25.4-2 The balanced homodyne receiver detects the intensity of the sum of the signal and local oscillator fields in one arm (+) and the intensity of their difference in the other (−). The resulting photocurrents are then subtracted electronically, yielding a net current that contains the signal information. (*a*) Beam-splitter implementation. (*b*) Optical-coupler implementation.

Since the phase difference between the waves at the two output ports of the beamsplitter differs by π (see Sec. 2.5A), one output branch contains the sum of the two fields whereas the other branch contains their difference. The detected currents $\bar{\imath}_{\pm} = \bar{\imath}_s + \bar{\imath}_L \pm 2\sqrt{\bar{\imath}_s \bar{\imath}_L}\cos\varphi_s$ are then electronically subtracted, which gives rise to

$$\bar{\imath} = 4\sqrt{\bar{\imath}_s \bar{\imath}_L}\cos\varphi_s. \tag{25.4-6}$$

The balanced homodyne mixer thus exhibits two advantages over the heterodyne system: (1) Random fluctuations in the intensity of the local oscillator [the second term of (25.4-5)] are canceled; and (2) the information-carrying signal [the third term] is multiplied by two.

A **phase-diversity balanced homodyne receiver** comprises a pair of balanced homodyne receivers that use local oscillators with phases φ_L and $\varphi_L - \pi/2$, which generate electric currents $\bar{\imath} = 4\sqrt{\bar{\imath}_s \bar{\imath}_L}\cos(\varphi_s - \varphi_L)$ and $\bar{\imath} = 4\sqrt{\bar{\imath}_s \bar{\imath}_L}\sin(\varphi_s - \varphi_L)$, respectively. With the phase-lock suppression of φ_L, the mixers thus provide the in-phase and quadrature components of the complex field, $\mathrm{I} \propto \sqrt{\bar{\imath}_s}\cos\varphi_s$ and $\mathrm{Q} \propto \sqrt{\bar{\imath}_s}\sin\varphi_s$, respectively.

Advantages and Disadvantages of Coherent Receivers

In comparison with their direct-detection counterparts, coherent receivers have the following advantages:

- They are capable of measuring the complex optical field, including its phase and frequency, thereby enabling field-modulation communication systems with their attendant increase in spectral efficiency.
- By making use of a strong local oscillator field, coherent receivers offer inherently noiseless conversion gain that effectively amplifies the signal above the circuit noise, as will become apparent subsequently.
- Coherent receivers offer a 3-dB advantage in signal-to-noise ratio over even noiseless direct-detection receivers, as will be shown shortly.
- Coherent receivers are insensitive to unwanted background light since the local oscillator does not mix with it.
- Coherent receivers offer one of the few ways of attaining photon-noise-limited detection in the infrared region of the spectrum, where background noise is prevalent.
- Access to the complex optical field permits the use of electronic equalization to compensate for signal impairments introduced by the communication channel, such as chromatic dispersion and polarization mode dispersion, which result in pulse broadening in optical fibers (see Sec. 10.3B). Moreover, nonlinear distortions introduced in fibers can be compensated by digital signal processing.
- The principal disadvantage of the coherent receiver centers on its more stringent requirements. Coherent systems require a stable, low-noise source of narrow bandwidth; a stable local oscillator; an optical mixer in which the superposed fields must be precisely aligned; and circuitry for phase locking and ancillary functions.

Coherent Communication Systems

The schematic diagram of a basic coherent optical fiber communication system that makes use of phase modulation and a balanced homodyne receiver is displayed in Fig. 25.4-3. An essential condition for proper mixing of the local oscillator and received optical field is that they be locked in phase, parallel, and have the same polarization. This places stringent requirements on the components of the system. The lasers must be single-frequency and have minimal phase and intensity fluctuations. The local

oscillator must be phase-locked to the received optical field by means of a phase-locked loop that adaptively adjusts its phase and frequency. The fiber should generally be single-mode (to avoid modal noise) and polarization-maintaining (or the receiver should contain an adaptive polarization-compensation system).

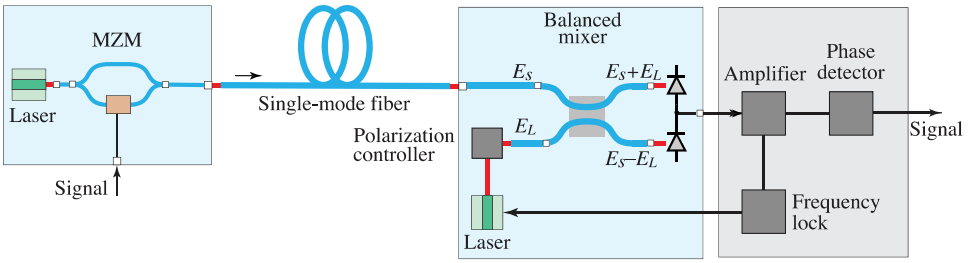


Figure 25.4-3 Coherent optical fiber communication system. The signal is phase modulated using a Mach–Zehnder modulator (MZM). The balanced mixer uses a tunable DFB laser and a phase-locked loop.

A more sophisticated optical fiber communication system is portrayed in Fig. 25.4-4. This system relies on quadrature-PSK (QPSK or 4-QAM) coding (see Sec. 25.3A) and a phase-diversity balanced homodyne receiver. At the transmitter (Tx), light from a laser is split into two branches containing Mach–Zehnder phase modulators that can introduce phase shifts of 0 or π. The upper branch represents the in-phase component I while the lower branch, which includes an additional phase shift of $\pi/2$, represents the quadrature component Q. The phase of the transmitted field can thus take on one of four values $(0, \pi/2, \pi, 3\pi/2)$, representing two bits per symbol. At the receiver (Rx), the laser serving as a local oscillator is split into two branches that feed two balanced mixers. The phase of the Q branch is shifted by $\pi/2$. The signals generated by the upper and lower mixers are hence proportional to $|E_s + E_L|^2 - |E_s - E_L|^2$ and $|E_s + jE_L|^2 - |E_s - jE_L|^2$, corresponding to detected currents proportional to $\cos\varphi_s$ and $\sin\varphi_s$, respectively, so that the I and Q (cosine and sine) components of the signal field are recovered. The system employs a phase-locked loop (not shown) that maintains the LO phase φ_L at zero.

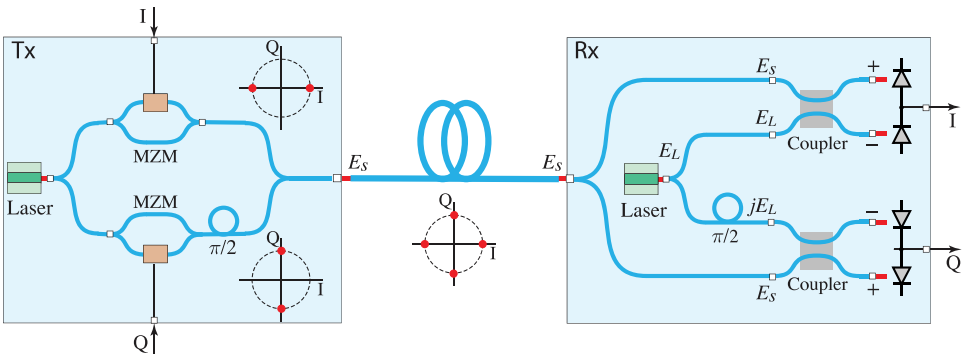


Figure 25.4-4 QPSK coherent optical fiber communication system. The transmitter (Tx) employs two Mach–Zehnder modulators (MZMs) to phase modulate the in-phase (I) and quadrature (Q) components of the complex field. The receiver (Rx) uses two balanced homodyne mixers and a laser local oscillator to recover I and Q.

A coherent optical fiber communication system implemented in the late 1980s typically operated at $\lambda_o = 1550$ nm with a bit rate below 1 Gb/s. The vast strides made over the years in high-speed electronics and digital signal processing, along with the use of spectrally efficient coding, have led to bit rates that are now many orders of magnitude greater. A system in current use might make use of 16-QAM or 64-QAM coding, a per-channel bit rate of 100 Gb/s, and hundreds of channels across the C+L-bands to yield an overall bit rate of tens of Tb/s (System ⑧ in Fig. 25.2-4). Experiments have been carried out that make use of more advanced coding, such as 2048-QAM, with the potential of achieving far higher overall bit rates.

Performance of Analog Coherent Communication Systems

Heterodyne detection is useful whenever the phase of an optical field is to be measured. It turns out that heterodyne detection is also useful for measuring optical intensity because it provides a form of amplification when the local oscillator is strong. This amplification is known as **conversion gain** since it converts some of the local-oscillator power into gain for the signal, as will become apparent below. Coherent detection thus offers an alternative to both optical amplification (Secs. 15.3 and 18.2) and avalanche-photodiode gain (Sec. 19.4). Indeed, heterodyne detection exhibits a signal-to-noise ratio advantage relative to direct detection, as we now demonstrate.

The mean photocurrent $\overline{i}$ generated in the photodiode of an optical receiver is accompanied by noise whose variance comprises two terms:

$$\sigma_i^2 = 2e\overline{i}B + \sigma_r^2, \tag{25.4-7}$$

where B is the receiver bandwidth. The first term represents the photocurrent shot noise [see (19.6-8)] while the second term represents current noise contributed by the receiver circuitry [see Sec. 19.6C]. When heterodyning is used and the local oscillator is sufficiently strong, such that $\overline{i}_L \gg \overline{i}_s$ and $2e\overline{i}_L B \gg \sigma_r^2$, the photomixing current and receiver noise variance, set forth in (25.4-4) and (25.4-7), respectively, can be approximately written as

$$\overline{i} \approx \overline{i}_L + 2\sqrt{\overline{i}_s \overline{i}_L}\cos[\omega_I t + (\varphi_s - \varphi_L)] \tag{25.4-8a}$$

$$\sigma_i^2 \approx 2e\overline{i}_L B. \tag{25.4-8b}$$

In the case of amplitude modulation, the signal is represented by the RMS value of the sinusoidal waveform in (25.4-8a), so the phase is of no significance. The electrical signal power is therefore $\frac{1}{2}[2\sqrt{\overline{i}_s\overline{i}_L}]^2 = 2\overline{i}_s\overline{i}_L$ while the noise power is $\sigma_i^2 = 2e\overline{i}_L B$, so the power signal-to-noise ratio can be written as

$$\text{SNR} = \frac{2\overline{i}_s\overline{i}_L}{2e\overline{i}_L B} = \frac{\overline{i}_s}{eB}. \tag{25.4-9}$$

If $\overline{m} = \overline{i}/2Be$ is the mean number of photoelectrons observed in the receiver resolution time $T = 1/2B$ (as derived in Sec. 19.6A), then (25.4-9) becomes

$$\boxed{\text{SNR} = 2\overline{m}\,.} \tag{25.4-10}$$

Signal-to-Noise Ratio
Heterodyne Receiver

The presence of the strong local oscillator has thus rendered the SNR independent of both the magnitude of the LO and the presence of circuit noise. The conversion gain has served to effectively amplify the signal above the circuit noise.

By way of comparison, the SNR of a direct-detection photodiode receiver with the same signal current $\bar{\imath}_s$ is given by (see Sec. 19.6)

$$\text{SNR} = \frac{\bar{\imath}_s^2}{2e\bar{\imath}_s B + \sigma_r^2} = \frac{\overline{m}^2}{\overline{m} + \sigma_q^2}, \tag{25.4-11}$$

which accords with (19.6-40), where $\sigma_q^2 = (\sigma_r/2Be)^2$ is the circuit-noise parameter defined in (19.6-34). For large signal current or small circuit noise ($\overline{m} \gg \sigma_q^2$), the direct-detection result in (25.4-11) reduces to SNR $= \overline{m}$.

The principal advantage of the coherent-detection system is apparent. The heterodyne system, with SNR $= 2\overline{m}$, offers a factor of 2 (or 3-dB) advantage over the direct-direction system. For weak light (or large circuit noise) the advantage is even greater: the SNR of the direct-detection system is further reduced by circuit noise to $\overline{m}/(1+\sigma_q^2/\overline{m})$, whereas the SNR of the heterodyne system remains at $2\overline{m}$. Moreover, an avalanche photodiode incorporated into a direct-detection system does not help matters. When the APD gain is sufficiently large so that it overcomes circuit noise, in accordance with (19.6-39) we obtain

$$\text{SNR} = \overline{m}/F, \tag{25.4-12}$$

where F is the APD excess noise factor ($F > 1$). Even a noiseless APD receiver ($F = 1$) provides a result that is a factor of 2 inferior to that of the heterodyne receiver.

Performance of Digital Coherent Communication Systems

We now proceed to examine in turn the performance and sensitivity of digital coherent communication systems that makes use of amplitude and phase modulation.

ON–OFF keying (OOK) homodyne system. Consider an ON–OFF keying (OOK) system that transmits data at a rate B_0 b/s and uses a homodyne receiver. The logic states "1" and "0" are represented by the presence and absence of the signal $\bar{\imath}_s$ during the bit time $T = 1/B_0$, respectively. Assuming that the local oscillator is strong, and that $\varphi_s = \varphi_L = 0$ and $\omega_I = \omega_s - \omega_L = 0$, the measured photocurrent exhibits the following means $\mu_{1,0}$ and variances $\sigma_{1,0}^2$, as provided by (25.4-8a) and (25.4-8b):

$$\begin{aligned} &\text{mean } \mu_1 \approx \bar{\imath}_L + 2\sqrt{\bar{\imath}_s \bar{\imath}_L}, \quad &&\text{variance } \sigma_1^2 \approx 2e\bar{\imath}_L B \quad &&\text{for state "1"} \\ &\text{mean } \mu_0 \approx \bar{\imath}_L, &&\text{variance } \sigma_0^2 \approx 2e\bar{\imath}_L B &&\text{for state "0".} \end{aligned} \tag{25.4-13}$$

The receiver bandwidth $B = B_0/2$ since the bit time $T = 1/B_0$ is the sampling time $1/2B$ for a signal of bandwidth B. Figure 25.4-5(a) offers a graphical depiction of the distance $\mu_1 - \mu_0$ in relation to the RMS noise $\sigma_1 = \sigma_0 = \sigma$ for the OOK constellation diagram.

The performance of a binary communication system under the Gaussian approximation was considered in Sec. 19.6E. As provided in (19.6-58) and (19.6-59), the bit error rate is given by

$$\text{BER} \approx \tfrac{1}{2}[1 - \text{erf}(Q/\sqrt{2})], \tag{25.4-14}$$

where $Q = (\mu_1 - \mu_0)/(\sigma_1 + \sigma_0)$. Using (25.4-13) thus leads to

$$Q = \frac{\mu_1 - \mu_0}{\sigma_1 + \sigma_0} = \sqrt{\frac{\bar{\imath}_s}{2eB}} = \sqrt{\overline{m}}, \tag{25.4-15}$$

where $\overline{m} = \overline{\imath}_s/2eB$ is the mean number of detected photoelectrons for state "1". For a bit error rate BER $= 10^{-9}$, we obtain $Q \approx 6$ and therefore $\overline{m} = 36$, corresponding to a receiver sensitivity $\overline{m}_0 = \frac{1}{2}\overline{m} = 18$ photoelectrons per bit (averaged over both logic states).

Binary phase shift keying (BPSK) homodyne system. For a binary PSK (BPSK) system, logic states "1" and "0" are represented by phase shifts $\varphi_s = 0$ and π, respectively, as illustrated in Fig. 25.4-5(b). Assuming that $\varphi_L = 0$ and $\omega_I = \omega_s - \omega_L = 0$ in (25.4-8), the means and variances of the photocurrent for states "1" and "0" are, respectively,

$$\begin{aligned} &\text{mean } \mu_1 = \overline{\imath}_L + 2\sqrt{\overline{\imath}_s \overline{\imath}_L}, \quad \text{variance } \sigma_1^2 = 2e\overline{\imath}_L B \quad \text{for state "1"} \\ &\text{mean } \mu_0 = \overline{\imath}_L - 2\sqrt{\overline{\imath}_s \overline{\imath}_L}, \quad \text{variance } \sigma_0^2 = 2e\overline{\imath}_L B \quad \text{for state "0".} \end{aligned} \tag{25.4-16}$$

In this case, therefore, we obtain

$$Q = \frac{\mu_1 - \mu_0}{\sigma_1 + \sigma_0} = 2\sqrt{\frac{\overline{\imath}_s}{2eB}} = 2\sqrt{\overline{m}} \tag{25.4-17}$$

which is a factor of 2 greater than that in the OOK case. This is also evident by comparing the constellation diagrams in Figs. 25.4-5(a) and (b). Again, a BER $= 10^{-9}$ leads to $Q \approx 6$, but now $\overline{m} = 9$. Since *each* of the two logic states of the bit must carry an average of nine photoelectrons in this case, the average number of photoelectrons per bit is $\overline{m}_0 = \overline{m} = 9$. We conclude that the receiver sensitivity is 9 photoelectrons/bit and that the BPSK homodyne receiver is hence twice as sensitive as the OOK homodyne receiver — it requires half the number of photoelectrons to achieve the same BER.

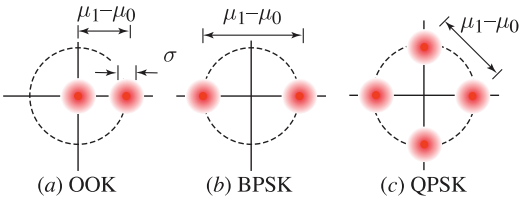


Figure 25.4-5 Constellation diagrams for OOK, BPSK, and QPSK homodyne systems. The red circle represents noise.

Quadrature phase shift keying (QPSK) homodyne system. As a final example, we consider the sensitivity of a homodyne quadrature-PSK (QPSK) system, and demonstrate that it turns out to be 9 photoelectrons per bit as well. This can be established by comparing the constellation diagrams presented in Figs. 25.4-5(b) and (c). The distance $\mu_1 - \mu_0$ between the nearest points in the QPSK constellation is smaller than that between the nearest points in the BPSK constellation by a factor of $1/\sqrt{2}$, while the noise σ is the same. Again, a BER $= 10^{-9}$ leads to $Q = 6$, but in this case $Q = \sqrt{2\overline{m}}$ so that $\overline{m} = 18$. In QPSK, however, each symbol corresponds to two bits so that the number of photoelectrons per bit $\overline{m}_0 = \overline{m}/2 = 9$, which is the same result as that for the BPSK homodyne system.

Comparison of heterodyne- and homodyne-system performance. The heterodyne digital receiver requires a factor of two more photons per bit than the homodyne receiver. This may be understood by comparing the signal currents for OOK keying in both cases. For homodyning we obtain $\mu_1 - \mu_0 = 2\sqrt{\bar{\imath}_s \bar{\imath}_L}$, as provided in (25.4-13), whereas for heterodyning the result is a factor of $\sqrt{2}$ smaller, namely $\mu_1 - \mu_0 = \sqrt{2\,\bar{\imath}_s \bar{\imath}_L}$. The origin of the distinction is the cosinusoidal factor in (25.4-8a), which is constant for homodyning but oscillates at the intermediate frequency ω_I for heterodyning, with an RMS value of $1/2$. Since the noise variances are the same ($\sigma_1^2 = \sigma_0^2$), and since $Q \approx 6$ for a BER $= 10^{-9}$, we have $6 = \sqrt{\overline{m}/2}$ so that $\overline{m}_0 = \overline{m}/2 = 36$. This factor-of-two penalty for heterodyning carries over to phase shift keying, as indicated in Table 25.4-1.

Tabulation of receiver sensitivities. Table 25.4-1 provides a comparison of the receiver sensitivities (photons per bit) for several optical receivers and modulation formats under ideal conditions. Though it appears that the direct-detection OOK system has approximately the same sensitivity as the best coherent system (homodyne PSK), namely ≈ 9 photons per bit, the conversion gain provided by the strong local oscillator in homodyning has the salutary effect of minimizing the role of circuit noise. The performance of direct-detection systems, in contrast, is often limited by circuit noise in practice. While the use of an avalanche photodiode in a direct-detection receiver can mitigate the role of circuit noise, the intrinsic APD gain noise increases the receiver sensitivity from 10 photons per bit to at least $10F$ photons per bit, where F is the APD excess noise factor. Direct detection could in principle offer performance comparable to that of coherent detection were noiseless APDs ($F = 1$) available. It is useful to compare the results tabulated in Table 25.4-1 with those presented in Table 19.6-1 for OOK direct-detection receivers in the presence of circuit noise and gain noise.

Table 25.4-1 Receiver sensitivity (number of photons per bit) for various receivers and modulation formats for an ideal detector. Homodyning is superior to heterodyning and PSK is superior to OOK.

	Direct Detection	Homodyne	Heterodyne
OOK	10	18	36
PSK (BPSK & QPSK)	—	9	18
FSK	—	—	36

25.5 FIBER-OPTIC NETWORKS

A communication network comprises a set of communication links connecting multiple users (terminals) distributed within some geographical area. Messages or data may be passed from one terminal to another by transmission through one or several links along paths controlled by routers and switches. A **local-area network** (**LAN**), for example, connects terminals such as computers, printers, video monitors, and copy machines in a circumscribed region such as a building, campus, or manufacturing plant. Larger networks include the telephone network, the global Telex network, and the Internet. The network may make use electrical cables, optical fibers, or satellite links. Fiber-optic networks rely on fiber-optic links together with electronic and/or optical routers and switches (see Chapter 24).

A. Network Topologies and Multiple Access

In its simplest configuration, a network containing N nodes is constructed by making use of a dedicated point-to-point link between each node and every other node. However, this configuration requires $N(N-1)$ duplex (bidirectional) point-to-point links, and therefore entails the use of $2N(N-1)$ transmitters and $2N(N-1)$ receivers. Topologies that make use of fewer point-to-point links, and fewer transmitters and receivers, are available; they include the star, ring, and bus topologies depicted in Fig. 25.5-1, along with a system for accessing the shared links. In these networks, only N transmitters and N receivers are necessary.

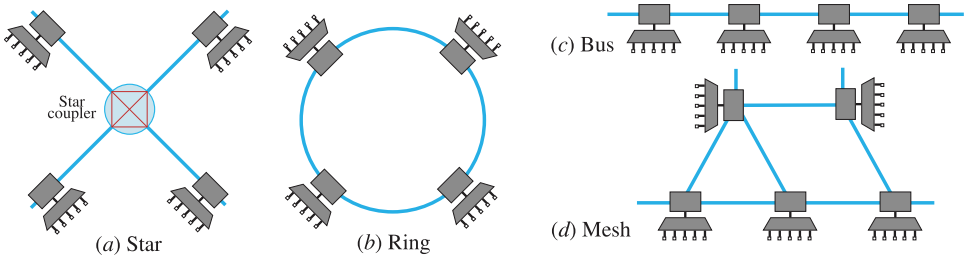


Figure 25.5-1 Network topologies: (a) star; (b) ring; (c) bus; (d) mesh.

In the star network, each node is connected to every other node via the star coupler residing at the center of the network; the power transmitted by any given node is equally distributed among all other nodes. In the ring and bus networks, the fiber passes through the nodes so that data may be extracted from, or added to, the optical signal at any node. The mesh network is a more general configuration. Since the light transmitted by any node travels different distances to different nodes, the receivers must be able to accommodate the received power over a broad range of levels. Stated differently, each receiver must have a large dynamic range. Several networks of the same or different topologies are often interconnected to create a larger network, as schematized in Fig. 25.5-2.

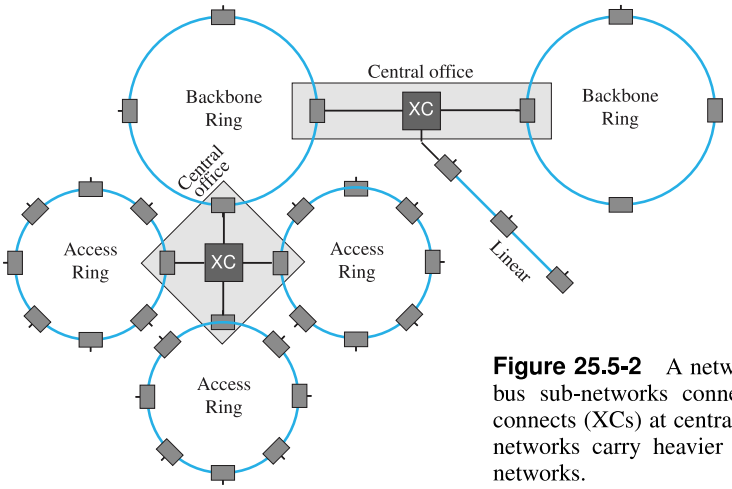


Figure 25.5-2 A network containing ring and bus sub-networks connected by digital cross-connects (XCs) at central offices. Backbone ring networks carry heavier traffic and feed access networks.

Interface

The interface between the terminal and the fiber network at each node includes a receiver, a transmitter, and an electronic add–drop multiplexer (ADM), as portrayed in Fig. 25.5-3(*a*). The receiver detects the optical signal, and the ADM extracts data and adds new data that modulate a source and transmits a new optical signal through another fiber. This interface is said to be *opaque* since the light is detected and regenerated at each node. A *transparent* interface is coupled to the fiber network optically, as illustrated in Fig. 25.5-3(*b*) (optical directional couplers are described in Secs. 24.1 and 24.3). An optical interface to a *bidirectional* (duplex) fiber uses two directional couplers to transmit and receive in either direction, as shown in Fig. 25.5-3(*c*).

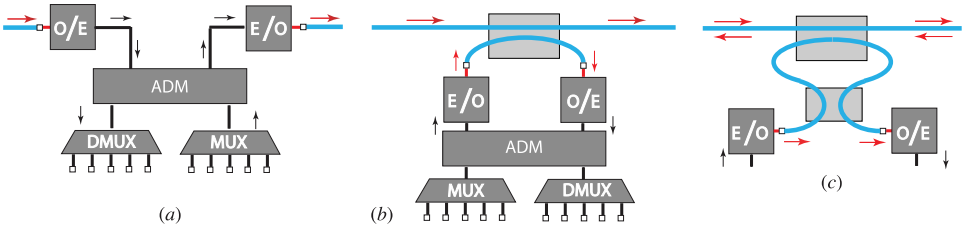


Figure 25.5-3 Interfaces between a node and the fiber network. (*a*) Opaque interface. The signal is converted from optical to electronic (O/E) and the ADM extracts data and adds new data, which is used to generate a new optical signal (E/O). (*b*) Optically coupled (transparent) interface using a directional coupler. (*c*) Optically coupled interface to a duplex fiber using two directional couplers.

Multiple Access

The signals transmitted by the network nodes share the same fiber (the **medium**). To avoid confusion, a scheme for **multiple access** or **medium access** is necessary. Time-domain, frequency-domain, and code-domain multiple access systems are in use:

- **Time-division multiple access (TDMA)** is similar to time-division multiplexing (TDM), which is used in conventional point-to-point communication systems (Sec. 25.3B). The nodes send their data through the shared medium during interleaved time slots. Buffers may be used to store data until the appropriate time. Since it is not possible to synchronize the timing of all nodes, guard times separating consecutive slots are necessary.
- **Frequency-division multiple access (FDMA)** is similar to frequency-division multiplexing (FDM) (Sec. 25.3B). Here, the nodes send their data through the shared medium in preassigned spectral bands, and there is no need to synchronize the bit clocks of the input signals. In optical networks, FDMA is called **wavelength-division multiple access** (**WDMA**), which is the counterpart of wavelength-division multiplexing (WDM).
- **Code-division multiple access (CDMA)** is similar to code-division multiplexing (CDM) (Sec. 25.3B). In CDMA, each node is preassigned a unique address code. Data transmitted by a node is encoded with the address code of the destination node. Each node correlates its own address code with the incoming signal. This locks it to only those bits associated with its own address, and it disregards all other bits. The data arrive in the form of a sequence of packets, each with the address of its destination (Sec. 24.3F).

Synchronous Optical Network (SONET)

SONET [and its international version, the Synchronous Digital Hierarchy (SDH)] is a TDM standard used for transmission over optical fibers. It addresses the difficulty of time-division multiplexing for signals with slightly different clock rates by embedding

these signals within time frames of longer duration. The payload (the signal bits) are allowed to float within the frames, but the frames are perfectly synchronous. SONET provides a hierarchy of multiplexed signals in which the basic unit, known as the STS-1 signal or the optical carrier-1 (**OC-1**), transports data at 51.84 Mb/s. Combining N such signals generates the **OC-N** signal, which has a rate N times greater, as summarized in Table 25.5-1. For example, OC-192 and OC-768 operate at approximately 10 Gb/s and 40 Gb/s, respectively. An example of operation at different data transport rates is illustrated in Example 25.5-1.

Table 25.5-1 Transmission rates (Mb/s) in the STS hierarchy used by the SONET network.

OC-1	OC-3	OC-12	OC-24	OC-48	OC-192	OC-768	OC-1920
51.84	155.52	622.08	1 244.16	2 488.32	9 953.28	39 813.12	99 532.80

EXAMPLE 25.5-1. ***Ring Network.***
An example of a fiber-optic 4-node ring network operating at different data rates is illustrated in Fig. 25.5-4. Each of the 4 nodes transmits data to the other 3 nodes at either the OC-12 (622 Mb/s) or the OC-24 (1.24 Gb/s) rate, as shown. The fiber segment connecting nodes 1 and 2 carries the heaviest traffic at a combined rate OC-12 + OC-12 + OC-24 = OC-48 (2.5 Gb/s). The traffic on the $2 \leftrightarrow 3$ and $3 \leftrightarrow 4$ segments is at the lighter OC-24 rate (1.24 Gb/s).

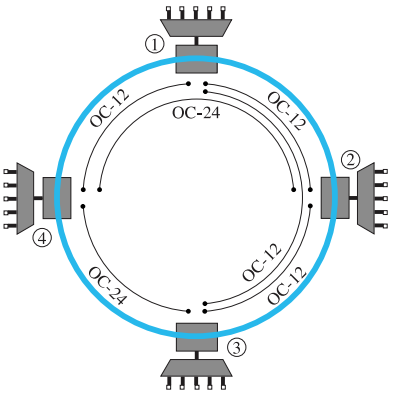


Figure 25.5-4 A 4-node ring network.

B. Wavelength-Division Multiplexing Networks

A wavelength-division multiplexing (WDM) fiber-optic network uses coarse or dense WDM for communication along its links and WDMA for medium access. The nodes are connected in a particular topology (e.g., star, ring, bus, or mesh), and each node transmits into one or several wavelength channels and receives from one or several wavelength channels. The existence of multiple wavelength channels for each physical connection adds another dimension to the network and offers additional flexibility, but this comes at the expense of additional complexity.

Broadcast-and-Select WDM Network

The simplest WDM network is the **broadcast-and-select network**. Each node transmits at a unique fixed wavelength and *broadcasts* its transmission to all other nodes via passive optical couplers. The receiver at each node *selects* the particular wavelength addressed to it by means of a tunable filter. An example is provided by the 5-node network displayed in Fig. 25.5-5(a): nodes $1, 2, \cdots, 5$ transmit at wavelengths $\lambda_1, \lambda_2, \cdots, \lambda_5$, respectively. An optical star coupler broadcasts each transmission to all other nodes. In the state shown, for example, node 1 is tuned to channel λ_5; nodes 2,

3, and 4 are tuned to channel λ_1; and node 5 is tuned to channel λ_2. As illustrated in the equivalent connection diagram in Fig. 25.5-5(b), node 2 transmits to node 5, node 5 transmits to node 1, and node 1 multicasts its transmission to nodes 2, 3, and 4.

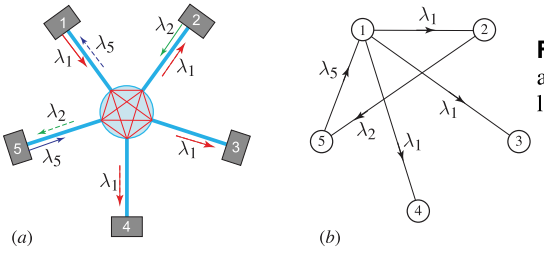


Figure 25.5-5 (a) A WDM broadcast-and-select network and (b) its equivalent logical connections.

In another example, shown in Fig. 25.5-6(a), the receiver of each node is tuned to the wavelength transmitted by its next neighbor. This network, which has a *star* physical topology, is thus equivalent to a *ring* logical topology, as illustrated in Fig. 25.5-6(b).

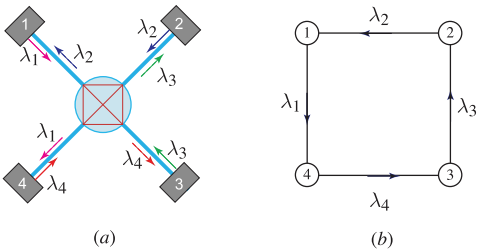


Figure 25.5-6 A WDM network in the star physical topology (a) is equivalent to the ring logical topology (b).

The network changes its state, i.e., the wavelengths to which each node is tuned, as desired. Dynamic coordination is required to avoid conflict and collisions.

Multi-Hop Broadcast-and-Select WDM Network

The requirement that each of the nodes in the broadcast-and-select network be capable of selectively detecting any of the wavelengths transmitted by the other nodes can be demanding. This requirement is alleviated in a multi-hop network, in which each node is allocated two different wavelength channels for transmission and only two different channels for reception. At any time, a node may transmit at one of its two allocated wavelengths and may receive by tuning to one its two allocated wavelengths. The channels are allocated to the nodes in such a way that a node may access any other node by following either a single-hop (i.e., direct) connection or a two-hop connection via an intermediate node. In the network shown in Fig. 25.5-7(a), for example, node 2 can transmit to node 1 directly via channel λ_3. Though node 1 cannot transmit to node 2 directly, since they share no common wavelength, this transmission can occur in two hops: node 1 transmits to node 3 on the λ_1 channel, and node 3 subsequently transmits to node 2 on the λ_6 channel [Fig. 25.5-7(b)]. This configuration is therefore called a **multi-hop broadcast-and-select network**. The logical topology of the network is displayed in Fig. 25.5-7(c).

However, networks with a large number of nodes are not well served by the broadcast-and-select, single-hop, or multi-hop configurations. Since the power transmitted by each node must reach all other nodes, the system becomes inefficient when the number of nodes is large. Also the number of channels used, which must equal or exceed the number of nodes, becomes prohibitive for large networks.

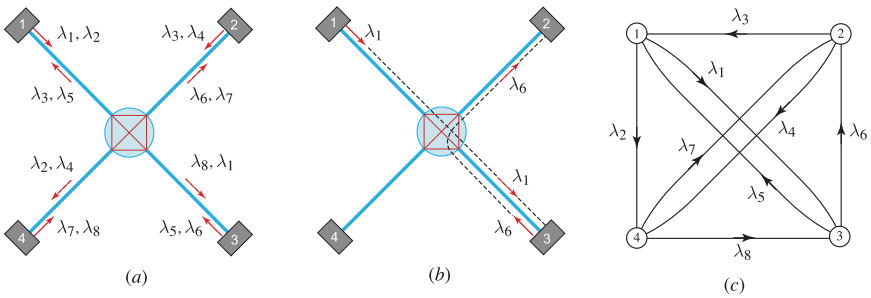


Figure 25.5-7 (a) A WDM multi-hop broadcast-and-select network. (b) A two-hop connection from node 1 to node 2 via node 3. (c) Logical topology of the network.

Wavelength-Routed Networks

In a **wavelength-routed network**, a pair of nodes communicates by use of one of the wavelength channels following some connection path. Another pair of nodes may use the same wavelength channel if their connection path does not share a common link with the path of the first pair. For example, in the network shown in Fig. 25.5-8(a), nodes 1 and 2 communicate on channel λ_1, and so do nodes 2 and 3. However, nodes 1 and 3 must use a different wavelength λ_2 if they use the path connecting them via node 2. Similarly, nodes 4 and 1 communicate via a third channel λ_3 since their path contains links that use the λ_1 and λ_2 channels.

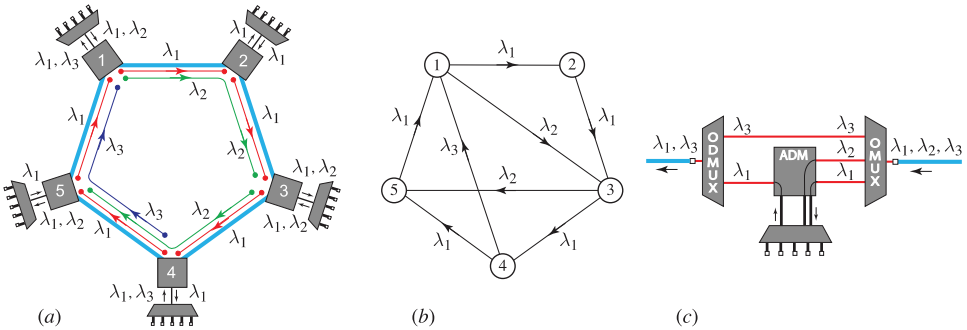


Figure 25.5-8 (a) A 5-node, 3-channel wavelength-routed ring network. (b) Logical topology of the network. (c) An optical add–drop multiplexer (OADM) used at node 5.

In this network, each link carries one or more wavelengths (but not necessarily all of the wavelengths, as is the case in the broadcast-and-select network). For example, the link between nodes 4 and 5 carries traffic at three wavelength channels, but each of the other four links carry only two channels. Also, each node transmits and receives data at one or more wavelengths. For example, node 5 receives data from node 4 at λ_1 and from node 3 at λ_2; it transmits data to node 1 at λ_1; data carried by channel λ_3 pass through this node without being detected. The logical connections for this network are shown in Fig. 25.5-8(b).

The key component in a wavelength-routed WDM network is the optical add–drop multiplexer (OADM) (Sec. 24.2A). Each node has an OADM that extracts (drops) data from certain wavelength channels on the incoming fiber, adds data to certain channels on the outgoing fiber, and lets data on certain channels of the incoming fiber pass through to the outgoing fiber without change. An OADM comprises an optical demultiplexer (ODMUX), an add–drop multiplexer (ADM), and an optical multiplexer (OMUX). As an example, the OADM used at node 5 of the network shown

in Fig. 25.5-8(a) is detailed in Fig. 25.5-8(c). Agile networks use ROADMs, which are reconfigurable OADMs (Example 24.3-3).

An optical transport network employing a ROADM is illustrated in Fig. 25.5-9. Data packets from many Internet Protocol (IP) packet routers are connected to an optical client interface (e.g., at a rate of 100 Gb/s over short-reach optical links of roughly 40 km) and are combined in small groups (one or more) and subsequently multiplexed into a WDM signal, which is transmitted over a single-fiber optical line traversing thousands of kilometers without intermediate electronic processing. The fiber may, for example, carry 100 wavelength channels, each operating at a rate of 200 Gb/s, on a 50-GHz optical frequency grid, for an overall bit rate of 20 Tb/s. The optical signal is then demultiplexed in a reverse fashion at the other of the line, and ultimately feeds other IP packet routers. Along the way, certain channels are dynamically directed to other optical lines in the network by use of ROADMs.

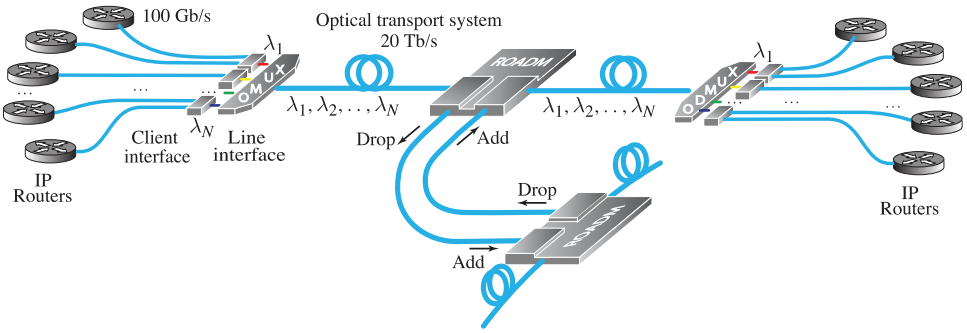


Figure 25.5-9 A WDM optical transport system linking two sets of IP routers and exchanging data with other optical lines in the network via ROADMs that add and drop information.

Wavelength-routed networks with configurations other than the ring configuration have nodes with multiple incoming and outgoing fibers. At these nodes, more complex routers are required. For example, a node with two incoming and two outgoing fibers, as shown in Fig. 25.5-10, employs an **optical cross-connect** (**OXC**) that receives data from selected incoming fibers/channels, adds data to selected outgoing fibers/channels, and routes data on selected incoming channels to selected outgoing channels. The OXC uses multidimensional space–wavelength switches and ADMs (see Sec. 24.3D). A wavelength-routed network also uses a hub node, which makes use of a server to process data at all wavelength channels.

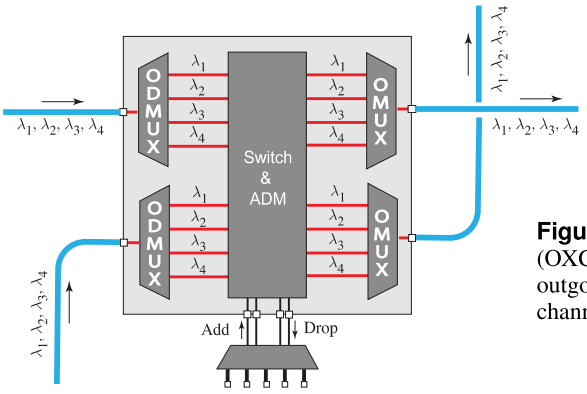


Figure 25.5-10 An optical cross-connect (OXC) at a node with two incoming and two outgoing fibers, each with four wavelength channels.

EXAMPLE 25.5-2. ***WDM Upgrade of a Ring Network***

A 4-node wavelength-routed WDM ring network operates on 3 channels with wavelengths λ_1, λ_2, and λ_3 at the rates shown in Fig. 25.5-11. This network is an upgraded version of the network considered in Example 25.5-1. Nodes 1 and 3 access wavelengths λ_1 and λ_2; node 4 accesses wavelengths λ_1 and λ_3; and node 2 accesses all three wavelengths. In the upgraded network, the nodes communicate at twice the rates of the original network, but the highest rate in any of the WDM channels does not exceed that of the original network. The fiber segment connecting nodes 1 and 2 carries the heaviest traffic at a combined OC-96 rate (5 Gb/s), but the highest rate at any given wavelength is OC-48 (2.5 Gb/s).

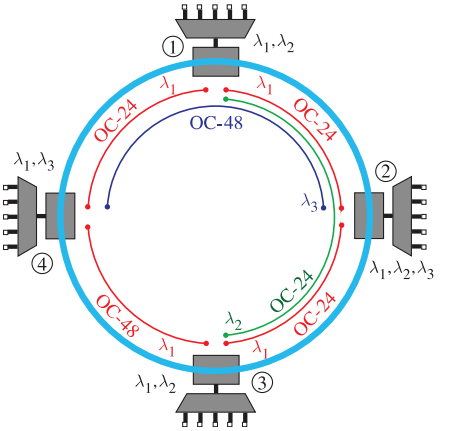


Figure 25.5-11 Schematic of a 4-node, 3-channel WDM ring network.

READING LIST

Optical Fiber Communications

See also the reading lists in Chapters 9, 10, 15–19, 23, and 24.

H. Venghaus and N. Grote, eds., *Fibre Optic Communication: Key Devices*, Springer-Verlag, 2nd ed. 2017.

E. Agrell, M. Karlsson, A. R. Chraplyvy, D. J. Richardson, P. M. Krummrich, P. Winzer, K. Roberts, J. K. Fischer, S. J. Savory, B. J. Eggleton, M. Secondini, F. R. Kschischang, A. Lord, J. Prat, I. Tomkos, J. E. Bowers, S. Srinivasan, M. Brandt-Pearce, and N. Gisin, Roadmap of Optical Communications, *Journal of Optics*, vol. 18, 063002, 2016.

R. Noé, *Essentials of Modern Optical Fiber Communication*, Springer-Verlag, 2nd ed. 2016.

P. J. Winzer, C. J. Chang-Hasnain, A. E. Willner, R. C. Alferness, R. W. Tkach, and T. G. Giallorenzi, eds., *A Third of a Century of Lightwave Technology: January 1983–April 2016*, IEEE–OSA, 2016.

F. Mitschke, *Fiber Optics: Physics and Technology*, Springer-Verlag, 2nd ed. 2016.

J. Chesnoy, ed., *Undersea Fiber Communication Systems*, Academic Press/Elsevier, 2nd paperback ed. 2016.

T. L. Singal, *Optical Fiber Communications: Principles and Applications*, Cambridge University Press, 2016.

L. N. Binh, Optical Fiber Communication Systems with MATLAB and SIMULINK Models, CRC Press/Taylor & Francis, 2nd ed. 2015.

S. Bottacchi, *Theory and Design of Terabit Optical Fiber Transmission Systems*, Cambridge University Press, 2014.

S. Kumar and M. J. Deen, *Fiber Optic Communications: Fundamentals and Applications*, Wiley, 2014.

I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-B: Systems and Networks*, Academic Press/Elsevier, 6th ed. 2013.

G. P. Agrawal, *Nonlinear Fiber Optics*, Academic Press/Elsevier, 5th ed. 2013.

G. P. Agrawal, *Fiber-Optic Communication Systems*, Wiley, 4th ed. 2010.

G. Keiser, *Optical Fiber Communications*, McGraw–Hill, 4th ed. 2010.

J. M. Senior, *Optical Fiber Communications: Principles and Practice*, Prentice Hall/Pearson, 3rd ed. 2009.

C. K. Kao, *Optical Fiber Systems: Technology, Design, and Applications*, McGraw–Hill, 1982.

Photonic Integrated Circuits

See also the reading list on silicon photonics in Chapter 18.

H. Radamson, E. Simoen, J. Luo, and C. Zhao, *CMOS Past, Present and Future*, Elsevier-Woodhead, 2018.

R. A. Soref, D. Buca, and S.-Q. Yu, Group IV Photonics: Driving Integrated Optoelectronics, *Optics & Photonics News*, vol. 27, no. 1, pp. 32–39, 2016.

D. Thomson, A. Zilkie, J. E. Bowers, T. Komljenovic, G. T. Reed, L. Vivien, D. Marris-Morini, E. Cassan, L. Virot, J.-M. Fédéli, J.-M. Hartmann, J. H. Schmid, D.-X. Xu, F. Boeuf, P. O'Brien, G. Z. Mashanovich, and M. Nedeljkovic, Roadmap on Silicon Photonics, *Journal of Optics*, vol. 18, 073003, 2016.

L. Chrostowski and M. Hochberg, *Silicon Photonics Design: From Devices to Systems*, Cambridge University Press, 2015.

K. Bergman, L. P. Carloni, A. Biberman, J. Chan, and G. Hendry, *Photonic Network-on-Chip Design*, Springer-Verlag, 2014.

R. Nagarajan, C. Doerr, and F. Kish, Semiconductor Photonic Integrated Circuit Transmitters and Receivers, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

L. A. Coldren, S. W. Corzine, and M. L. Mašanović, *Diode Lasers and Photonic Integrated Circuits*, Wiley, 2nd ed. 2012.

D. Dai, J. Bauters, and J. E. Bowers, Passive Technologies for Future Large-Scale Photonic Integrated Circuits on Silicon: Polarization Handling, Light Non-Reciprocity and Loss Reduction, *Light: Science & Applications* (2012) **1**, e1; doi:10.1038/lsa.2012.

B. Razavi, *Design of Integrated Circuits for Optical Communications*, Wiley, 2nd ed. 2012.

H. Zimmermann, *Integrated Silicon Optoelectronics*, Springer-Verlag, 2nd ed. 2010.

R. G. Hunsperger, *Integrated Optics: Theory and Technology*, Springer-Verlag, 1982, 6th ed. 2010.

R. A. Soref and J. P. Lorenzo, Single-Crystal Silicon: A New Material for 1.3 and 1.6 μm Integrated-Optical Components, *Electronics Letters*, vol. 21, pp. 953–954, 1985.

S. E. Miller, Integrated Optics: An Introduction, *Bell System Technical Journal*, vol. 48, pp. 2059–2069, 1969.

Modulation and Multiplexing

T. Hayashi, Multi-Core Optical Fibers, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

Y. Awaji, K. Saitoh, and S. Matsuo, Transmission Systems Using Multicore Fibers, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-B: Systems and Networks*, Academic Press/Elsevier, 6th ed. 2013.

D. W. Peckham, Y. Sun, A. McCurdy, and R. Lingle, Jr., Few-Mode Fiber Technology for Spatial Multiplexing, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

K.-P. Ho and J. M. Kahn, Mode Coupling and its Impact on Spatially Multiplexed Systems, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-B: Systems and Networks*, Academic Press/Elsevier, 6th ed. 2013.

P. J. Winzer, R. Ryf, and S. Randel, Spatial Multiplexing Using Multiple-Input Multiple-Output Signal Processing, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-B: Systems and Networks*, Academic Press/Elsevier, 6th ed. 2013.

L. K. Oxenløwe, A. Clausen, M. Galili, H. C. H. Mulvad, H. Ji, H. Hu, and E. Palushani, Ultra-High-Speed Optical Time Division Multiplexing, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-A: Components and Subsystems*, Academic Press/Elsevier, 6th ed. 2013.

Coherent Optical Detection and Communications

K. Kikuchi, Fundamentals of Coherent Optical Fiber Communications, *Journal of Lightwave Technology*, vol. 34, pp. 157–179, 2016.

X. Zhou and C. Xie, eds., *Enabling Technologies for High Spectral-Efficiency Coherent Optical Communication Networks*, Wiley, 2016.

D. J. Geisler, C. M. Schieler, T. M. Yarnall, M. L. Stevens, B. S. Robinson, and S. A. Hamilton, Demonstration of a Variable Data-Rate Free-Space Optical Communication Architecture Using Efficient Coherent Techniques, *Optical Engineering*, vol. 55, 111605, 2016.

M. Mazurczyk, Spectral Shaping in Long Haul Optical Coherent Systems With High Spectral Efficiency, *Journal of Lightwave Technology*, vol. 32, pp. 2915–2924, 2014.

T. J. Xia and G. A. Wellbrock, Commercial 100-Gbit/s Coherent Transmission Systems, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-B: Systems and Networks*, Academic Press/Elsevier, 6th ed. 2013.

P. Bayvel, C. Behrens, and D. S. Millar, Digital Signal Processing (DSP) and its Application in Optical Communication Systems, in I. P. Kaminow, T. Li, and A. E. Willner, eds., *Optical Fiber Telecommunications VI-B: Systems and Networks*, Academic Press/Elsevier, 6th ed. 2013.

V. V. Protopopov, *Laser Heterodyning*, Springer-Verlag, 2009.

T. Okoshi and K. Kikuchi, *Coherent Optical Fiber Communications*, Kluwer, 1988.

M. C. Teich, Laser Heterodyning, *Journal of Modern Optics* (*Optica Acta*), vol. 32, pp. 1015–1021, 1985.

M. C. Teich, Coherent Detection in the Infrared, in R. K. Willardson and A. C. Beer, eds., *Semiconductors and Semimetals*, Volume 5, *Infrared Detectors*, Academic Press, pp. 361–407, 1970.

A. Javan, E. A. Ballik, and W. L. Bond, Frequency Characteristics of a Continuous-Wave He–Ne Optical Maser, *Journal of the Optical Society of America*, vol. 52, pp. 96–98, 1962.

B. J. McMurtry and A. E. Siegman, Photomixing Experiments with a Ruby Optical Maser and a Traveling-Wave Microwave Phototube, *Applied Optics*, vol. 1, pp. 51–53, 1962.

Fiber-Optic Networks

M. Tornatore, G.-K. Chang, and G. Ellinas, eds., *Fiber-Wireless Convergence in Next-Generation Communication Networks: Systems, Architectures, and Management*, Springer-Verlag, 2017.

X. Zhou and C. Xie, eds., *Enabling Technologies for High Spectral-Efficiency Coherent Optical Communication Networks*, Wiley, 2016.

V. López and L. Velasco, eds., *Elastic Optical Networks: Architectures, Technologies, and Control*, Springer-Verlag, 2016.

J. M. Simmons, *Optical Network Design and Planning*, Springer-Verlag, 2nd ed. 2014.

C. Kachris, K. Bergman, and I. Tomkos, eds., *Optical Interconnects for Future Data Center Networks*, Springer-Verlag, 2013.

M. Cvijetic and I. B. Djordjevic, *Advanced Optical Communication Systems and Networks*, Artech, 2013.

D. Hood and E. Trojer, *Gigabit-Capable Passive Optical Networks*, Wiley, 2012.

M. Toy, *Networks and Services: Carrier Ethernet, PBT, MPLS-TP, and VPLS*, Wiley, 2012.

N. Antoniades, G. Ellinas, and I. Roudas, eds., *WDM Systems and Networks: Modeling, Simulation, Design and Engineering*, Springer-Verlag, 2012.

T. E. Stern, G. Ellinas, and K. Bala, *Multiwavelength Optical Networks: Architectures, Design, and Control*, Cambridge University Press, 2nd ed. 2009.

B. Mukherjee, *Optical WDM Networks*, Springer-Verlag, 2006.

P. R. Prucnal, ed., *Optical Code Division Multiple Access: Fundamentals and Applications*, CRC Press/Taylor & Francis, 2006.

E. Desurvire, *Global Telecommunications: Broadband Access, Optical Components and Networks, and Cryptography*, Wiley, 2004.

E. Desurvire, *Global Telecommunications: Signaling Principles, Protocols, and Wireless Systems*, Wiley, 2004.

P. E. Green, Jr., *Fiber Optic Networks*, Prentice Hall, 1992.

Seminal and Historical

I. P. Kaminow, Optical Integrated Circuits: A Personal Perspective, *Journal of Lightwave Technology*, vol. 26, pp. 994-1004, 2008.

A. Hasegawa, ed., Massive WDM and TDM Soliton Transmission Systems, Kluwer, 2002.

R. M. Gagliardi and S. Karp, *Optical Communications*, Wiley, 1976, 2nd ed. 1995.

D. L. Begley, ed., *Selected Papers on Free-Space Laser Communications II*, SPIE Optical Engineering Press (Milestone Series Volume 100), 1994.

E. G. Rawson, ed., *Selected Papers on Fiber Optic Local Area Networks*, SPIE Optical Engineering Press (Milestone Series Volume 91), 1994.

L. D. Hutcheson and S. C. Mettler, eds., *Selected Papers on Fiber Optic Communications*, SPIE Optical Engineering Press (Milestone Series Volume 88), 1993.

C. K. Kao, *Optical Fibre*, Institution of Electrical Engineers, 1988.

E. Desurvire, J. R. Simpson, and P. C. Becker, High-Gain Erbium-Doped Traveling-Wave Fiber Amplifier, *Optics Letters*, vol. 12, pp. 888–890, 1987.

R. J. Mears, L. Reekie, I. M. Jauncey, and D. N. Payne, Low-Noise Erbium-Doped Fibre Amplifier Operating at 1.54 μm, *Electronics Letters*, vol. 23, pp. 1026–1028, 1987.

B. J. Ainslie and C. R. Day, A Review of Single-Mode Fibers with Modified Dispersion Characteristics, *Journal of Lightwave Technology*, vol. LT-4, pp. 967–979, 1986.

S. D. Personick, *Fiber Optics: Technology and Applications*, Plenum, 1985.

S. B. Poole, D. N. Payne, and M. E. Fermann, Fabrication of Low Loss Optical Fibers Containing Rare Earth Ions, *Electronics Letters*, vol. 21, pp. 737–738, 1985.

T. Li, Advances in Optical Fiber Communications: An Historical Perspective, *IEEE Journal on Selected Areas in Communications*, vol. SAC-1, pp. 356–372, 1983.

J. E. Midwinter, *Optical Fibers for Transmission*, Wiley, 1979; Krieger, reissued 1992.

D. B. Keck, R. D. Maurer, and P. C. Schultz, On the Ultimate Lower Limit of Attenuation in Glass Optical Waveguides, *Applied Physics Letters*, vol. 22, pp. 307–309, 1973.

O. E. DeLange, Wide-Band Optical Communication Systems: Part II–Frequency-Division Multiplexing, *Proceedings of the IEEE*, vol. 58, pp. 1683–1690, 1970.

M. C. Teich, Infrared Heterodyne Detection, *Proceedings of the IEEE*, vol. 56, pp. 37–46, 1968.

S. E. Miller, Communication by Laser, *Scientific American*, vol. 214, no. 1, pp. 19–27, 1966.

K. C. Kao and G. A. Hockham, Dielectric-Fibre Surface Waveguides for Optical Frequencies, *IEE Proceedings*, vol. 113, pp. 1151–1158, 1966.

C. J. Koester and E. Snitzer, Amplification in a Fiber Laser, *Applied Optics*, vol. 3, pp. 1182–1186, 1964.

Popular

J. Hecht, Great Leaps of Light, *IEEE Spectrum*, vol. 53, no. 2, pp. 28–53, 2016.

P. J. Winzer, Scaling Optical Fiber Networks: Challenges and Solutions, *Optics & Photonics News*, vol. 26, no. 3, pp. 28–35, 2015.

J. Hecht, *Understanding Fiber Optics*, Laser Light Press, 5th ed. 2015.

J. Hecht, Recycled Fiber Optics: How Old Ideas Drove New Technology, *Optics & Photonics News*, vol. 23, no. 2, pp. 22–29, 2012.

J. Hecht, *City of Light: The Story of Fiber Optics*, Oxford University Press, 1999, paperback ed. 2004.

PROBLEMS

25.1-1 **Optical Fiber Communication Systems.** Discuss the validity of each of the following statements and indicate the conditions under which your conclusion is applicable.

(a) The wavelength $\lambda_o = 1300$ nm is preferred to $\lambda_o = 870$ nm for all optical fiber communication systems.

(b) The wavelength $\lambda_o = 1550$ nm is preferred to $\lambda_o = 1300$ nm for all optical fiber communication systems.
(c) Single-mode fibers are superior to multimode fibers because they have lower attenuation coefficients.
(d) There is no pulse spreading at $\lambda_o \approx 1312$ nm in silica-glass fibers.
(e) Compound semiconductor devices are required for optical fiber communication systems.
(f) APDs are noisier than p–i–n photodiodes and are therefore not useful for optical fiber communication systems.

25.1-2 **Components for Optical Fiber Communication Systems.** The design of an optical fiber communication system involves many choices of fibers, sources, amplifiers, and detectors, some of which are displayed in Fig. 25.2-3. Suggest appropriate choices for each of the applications listed below. Though more than one answer may be correct, some choices may be incompatible.

(a) A transoceanic cable carrying data at a 2.5 Gb/s rate with 100-km repeater spacings.
(b) A 1-m cable transmitting analog data from a sensor at 1 kHz.
(c) A link for a computer local-area network operating at 500 Mb/s.
(d) A 1-km data link operating at 100 Mb/s with $\pm 50^\circ$ C temperature variations.

25.2-1 **Performance of a Plastic Fiber Link.** A short-distance, low-data-rate communication system uses plastic fiber with an attenuation coefficient 0.5 dB/m, an LED that generates 1 mW at a wavelength of 870 nm, and a photodiode with receiver sensitivity -20 dBm. Assuming a power loss of 3 dB each at the input and output couplers, determine the maximum length of the link. Assume that the data rate is sufficiently low that dispersion effects play no role.

25.2-2 **Maximum Length of an Attenuation-Limited System.** An optical fiber communication link is designed for operation at 10 Mb/s. The source is a 100-μW LED operating at 870 nm and the fiber has an attenuation coefficient of 3.5 dB/km. The fiber consists of 1-km segments and each connector between segments introduces a loss of 1 dB. The input and output couplers each introduce a loss of 2 dB and the safety margin is 6 dB. Two receivers are available: a Si p–i–n photodiode receiver with a sensitivity of 5000 photons per bit, and a Si APD receiver with a sensitivity of 125 photons per bit. Determine the receiver sensitivity P_r (dBm units) and the maximum length of the link for each receiver.

25.2-3 **Maximum Data Rate for an Attenuation-Limited System.** A 50-km optical fiber link is operated at a wavelength of 1550 nm. The source is a 2-mW InGaAsP laser and the fiber has an attenuation coefficient of 0.2 dB/km. Connectors and couplers introduce a total loss of 8 dB and the safety margin is 6 dB. The receiver is an InGaAs APD with a sensitivity of 1000 photons per bit at a bit error rate of 10^{-9}. Determine the maximum data rate that can be used assuming that the system is attenuation-limited. If the required bit error rate is 10^{-11} instead, what is the maximum data rate?

25.2-4 **Maximum Length of an Analog Link.** An analog optical fiber communication link uses intensity modulation to transmit data at a bandwidth of $B = 10$ MHz with a signal-to-noise ratio of 40 dB. The source is a $\lambda_o = 870$ nm light-emitting diode that produces an average power of 100 μW with a maximum modulation index of 0.5. The fiber is a multimode step-index fiber with an attenuation coefficient of 2.5 dB/km. The detector is an avalanche photodiode with mean gain $\overline{G} = 100$, excess noise factor $F = 5$, and responsivity $\mathrm{R} = 0.5$ A/W (excluding the gain). Using the theory presented in Sec. 19.6D, assume that circuit noise is negligible and calculate the optical power sensitivity of the receiver. Calculate the attenuation-limited maximum length L of the fiber.

25.2-5 **Time Budget for a Dispersion-Limited System.** A 100-km single-mode fiber link operates at a wavelength of 1550 nm. The source is an InGaAsP laser diode of spectral width 0.2 nm and response time 20 ps. The fiber has a dispersion coefficient of 17 ps/km-nm. The receiver uses an InGaAs APD and has a response time of 0.1 ns. Determine the maximum data rate based on the criterion that the response time of the fiber does not exceed 25% of the bit duration. Also, determine the maximum data rate using the criterion that the response time of the overall system does not exceed 70% of the bit duration. If a dispersion-shifted fiber is used instead, so that the dispersion coefficient is reduced to 1 ps/km-nm, what are the maximum data rates under the two criteria set forth above?

25.3-1 **Number of WDM Channels.** Determine the number of WDM channels that fit in the C-band (1530–1565 nm) and in the O-band (1260–1360 nm) if the channel spacing is 75 GHz.

25.5-1 **Number of Nodes in a Broadcast-and-Select WDM Network.** The maximum number of nodes N that can be used by a broadcast-and-select WDM network is often limited by the available optical power. Determine N for a local area network using an optical star coupler connected to each of the nodes by a fiber of 2-km length, 0.3 dB/km attenuation coefficient, and 1 dB of connector loss. The star coupler distributes the power equally among its outputs and introduces an additional loss of 3 dB. Each node uses a 1-mW optical source, the receiver sensitivity is -35 dBm, and a 5-dB safety margin is assumed.

25.5-2 **Wavelength-Routed WDM Ring Network.** Consider a 4-node, 6-channel WDM network. Each node uses an add–drop multiplexer to transmit or receive at any of three different wavelengths assigned to it, but allows the other three wavelengths to pass through. For example, node 1 may add or drop data at channels λ_1, λ_2, or λ_3, but passes through data at λ_4, λ_5, and λ_6. Allocate sets of three add–drop channels to each of the nodes 2, 3, and 4, in such a way that any node on the ring may communicate with any of the other nodes. The idea is that each node must have one add–drop channel in common with each of the other three nodes, but this channel must *not* be in common with nodes in-between.

APPENDIX

A

FOURIER TRANSFORM

This appendix provides a brief review of the Fourier transform, and its properties, for functions of one and two variables.

A.1 ONE-DIMENSIONAL FOURIER TRANSFORM

The harmonic function $F\exp(j2\pi\nu t)$ plays an important role in science and engineering. It has frequency ν and complex amplitude F. Its real part $|F|\cos(2\pi\nu t+\arg\{F\})$ is a cosine function with amplitude $|F|$ and phase $\arg\{F\}$. The variable t usually represents time; the frequency ν has units of cycles/s or Hz. The harmonic function is regarded as a building block from which other functions may be obtained by a simple superposition.

In accordance with the Fourier theorem, a complex-valued function $f(t)$, satisfying some rather unrestrictive conditions, may be decomposed as a superposition integral of harmonic functions of different frequencies and complex amplitudes,

$$f(t)=\int_{-\infty}^{\infty}F(\nu)\exp(j2\pi\nu t)\,d\nu. \tag{A.1-1}$$

Inverse Fourier Transform

The component with frequency ν has a complex amplitude $F(\nu)$ given by

$$F(\nu)=\int_{-\infty}^{\infty}f(t)\exp(-j2\pi\nu t)\,dt. \tag{A.1-2}$$

Fourier Transform

$F(\nu)$ is termed the **Fourier transform** of $f(t)$, and $f(t)$ is the **inverse Fourier transform** of $F(\nu)$. The functions $f(t)$ and $F(\nu)$ form a **Fourier transform pair**; if one is known, the other may be determined.

In this book we adopt the convention that $\exp(j2\pi\nu t)$ is a harmonic function with positive frequency, whereas $\exp(-j2\pi\nu t)$ represents negative frequency. The opposite convention is used by some authors, who define the Fourier transform in (A.1-2) with a positive sign in the exponent, and use a negative sign in the exponent of the inverse Fourier transform (A.1-1).

In communication theory, the functions $f(t)$ and $F(\nu)$ represent a signal, with $f(t)$ its time-domain representation and $F(\nu)$ its frequency-domain representation. The absolute-squared value $|f(t)|^2$ is called the **signal power**, and $|F(\nu)|^2$ is the energy spectral density. If $|F(\nu)|^2$ extends over a wide frequency range, the signal is said to have a wide bandwidth.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.
©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Properties of the Fourier Transform

Some important properties of the Fourier transform are provided below. These properties can be proved by direct application of the definitions (A.1-1) and (A.1-2) (see any of the books in the Reading List).

- *Linearity.* The Fourier transform of the sum of two functions is the sum of their Fourier transforms.
- *Scaling.* If $f(t)$ has a Fourier transform $F(\nu)$, and τ is a real scaling factor, then $f(t/\tau)$ has a Fourier transform $|\tau|F(\tau\nu)$. This means that if $f(t)$ is scaled by a factor τ, its Fourier transform is scaled by a factor $1/\tau$. Thus, if $\tau > 1$, then $f(t/\tau)$ is a stretched version of $f(t)$, whereas $F(\tau\nu)$ is a compressed version of $F(\nu)$. The Fourier transform of $f(-t)$ is $F(-\nu)$.
- *Time Translation.* If $f(t)$ has a Fourier transform $F(\nu)$, the Fourier transform of $f(t-\tau)$ is $\exp(-j2\pi\nu\tau)F(\nu)$. Thus, delay by time τ is equivalent to multiplication of the Fourier transform by a phase factor $\exp(-j2\pi\nu\tau)$.
- *Frequency Translation.* If $F(\nu)$ is the Fourier transform of $f(t)$, the Fourier transform of $f(t)\exp(j2\pi\nu_0 t)$ is $F(\nu-\nu_0)$. Thus, multiplication by a harmonic function of frequency ν_0 is equivalent to shifting the Fourier transform to a higher frequency ν_0.
- *Symmetry.* If $f(t)$ is real, then $F(\nu)$ has Hermitian symmetry, i.e., $F(-\nu) = F^*(\nu)$. If $f(t)$ is real and symmetric, then $F(\nu)$ is also real and symmetric.
- *Convolution Theorem.* If the Fourier transforms of $f_1(t)$ and $f_2(t)$ are $F_1(\nu)$ and $F_2(\nu)$, respectively, the inverse Fourier transform of the product

$$F(\nu) = F_1(\nu)F_2(\nu) \tag{A.1-3}$$

  is

$$\boxed{f(t) = \int_{-\infty}^{\infty} f_1(\tau)f_2(t-\tau)\,d\tau.} \tag{A.1-4}$$

Convolution

  The operation defined in (A.1-4) is known as the **convolution** of $f_1(t)$ with $f_2(t)$. Convolution in the time domain is therefore equivalent to multiplication in the Fourier domain.
- *Correlation Theorem.* The **correlation** between two complex functions is defined as

$$\boxed{f(t) = \int_{-\infty}^{\infty} f_1^*(\tau)f_2(t+\tau)\,d\tau.} \tag{A.1-5}$$

Correlation

  The Fourier transforms of $f_1(t)$, $f_2(t)$, and $f(t)$ are related by

$$F(\nu) = F_1^*(\nu)F_2(\nu). \tag{A.1-6}$$

  If $f_2(t) = f_1(t)$, (A.1-5) is called the **autocorrelation**.
- *Parseval's Theorem.* The signal energy, which is the integral of the signal power $|f(t)|^2$, equals the integral of the energy spectral density $|F(\nu)|^2$, so that

$$\boxed{\int_{-\infty}^{\infty} |f(t)|^2\,dt = \int_{-\infty}^{\infty} |F(\nu)|^2\,d\nu.} \tag{A.1-7}$$

Parseval's Theorem

Examples

The Fourier transforms of some important functions are provided in Table A.1-1. The Fourier transforms of many other functions are readily obtained by making use of the properties of linearity, scaling, delay, and frequency translation. The functions used in Table A.1-1 have the following definitions:

- $\text{rect}(t) \equiv 1$ for $|t| \leq \frac{1}{2}$, and $= 0$ elsewhere, i.e., it is a pulse of unit height and unit width centered about $t = 0$.
- $\delta(t)$ is the impulse function (also called the Dirac delta function), which is defined as $\delta(t) \equiv \lim_{\alpha \to \infty} \alpha\, \text{rect}(\alpha t)$. It is the limit of a rectangular pulse of unit area as its width approaches zero so that its height approaches infinity.
- $\text{sinc}(t) \equiv \sin(\pi t)/(\pi t)$ is a symmetric function with a peak value of unity at $t = 0$ and with zeros at $t = \pm 1, \pm 2, \ldots$.

Table A.1-1 Selected functions and their Fourier transforms.

Function	$f(t)$	$F(\nu)$
Uniform	1	$\delta(\nu)$
Impulse	$\delta(t)$	1
Rectangular	$\text{rect}\,(t)$	$\text{sinc}\,(\nu)$
Exponential [a]	$\exp(-\lvert t\rvert)$	$\frac{2}{1+(2\pi\nu)^2}$
Gaussian	$\exp(-\pi t^2)$	$\exp(-\pi\nu^2)$
Hyperbolic secant	$\text{sech}\,(\pi t)$	$\text{sech}\,(\pi\nu)$
Chirp [b]	$\exp(j\pi t^2)$	$e^{j\pi/4}\exp(-j\pi\nu^2)$
$M = 2S + 1$ Impulses	$\sum_{m=-S}^{S} \delta(t-m)$	$\frac{\sin(M\pi\nu)}{\sin(\pi\nu)}$
Comb	$\sum_{m=-\infty}^{\infty} \delta(t-m)$	$\sum_{m=-\infty}^{\infty} \delta(\nu-m)$

[a] The double-sided exponential function is shown. The Fourier transform of the single-sided exponential, $f(t) = \exp(-t)$ with $t \geq 0$, is $F(\nu) = 1/[1 + j2\pi\nu]$. Its magnitude is $1/\sqrt{1+(2\pi\nu)^2}$.

[b] The functions $\cos(\pi t^2)$ and $\cos(\pi\nu^2)$ are shown. The function $\sin(\pi t^2)$ is shown in Fig. 4.3-6.

A.2 TIME DURATION AND SPECTRAL WIDTH

It is often useful to have a measure of the width of a function. The width of a function of time $f(t)$ is its time duration and the width of its Fourier transform $F(\nu)$ is its spectral width (or bandwidth). Since there is no unique definition for the width, a plethora of definitions are in use. *All definitions, however, share the property that the spectral width is inversely proportional to the temporal width, in accordance with the scaling property of the Fourier transform.* The following definitions are used at different places in this book.

The Root-Mean-Square Width

The *root-mean-square (RMS) width* σ_t of a nonnegative real function $f(t)$ is defined by

$$\sigma_t^2 = \frac{\int_{-\infty}^{\infty}(t-\bar{t})^2 f(t)\,dt}{\int_{-\infty}^{\infty} f(t)\,dt}, \quad \text{where} \quad \bar{t} = \frac{\int_{-\infty}^{\infty} t f(t)\,dt}{\int_{-\infty}^{\infty} f(t)\,dt}. \tag{A.2-1}$$

If $f(t)$ represents a mass distribution (t representing position), then $\bar{t}$ represents the centroid and σ_t the radius of gyration. If $f(t)$ is a probability density function, these quantities represent the mean and standard deviation, respectively. As an example, the *Gaussian function* $f(t) = \exp(-t^2/2\sigma_t^2)$ has an RMS width σ_t. Its Fourier transform is given by $F(\nu) = (1/\sqrt{2\pi}\,\sigma_\nu)\exp(-\nu^2/2\sigma_\nu^2)$, where

$$\sigma_\nu = \frac{1}{2\pi\sigma_t} \tag{A.2-2}$$

is the RMS spectral width.

This definition is not appropriate for functions with negative or complex values. For such functions the RMS width of the absolute-squared value $|f(t)|^2$ is used,

$$\sigma_t^2 = \frac{\int_{-\infty}^{\infty}(t-\bar{t})^2|f(t)|^2\,dt}{\int_{-\infty}^{\infty}|f(t)|^2\,dt}, \quad \text{where} \quad \bar{t} = \frac{\int_{-\infty}^{\infty} t|f(t)|^2\,dt}{\int_{-\infty}^{\infty}|f(t)|^2\,dt}.$$

We call this version of σ_t the *power-RMS width*.

With the help of the Schwarz inequality, it can be shown that the product of the power RMS widths of an arbitrary function $f(t)$ and its Fourier transform $F(\nu)$ must be equal to or greater than $1/4\pi$,

$$\boxed{\sigma_t\sigma_\nu \geq \frac{1}{4\pi},} \tag{A.2-3}$$

Duration–Bandwidth Reciprocity Relation

where the spectral width σ_ν is defined by

$$\sigma_\nu^2 = \frac{\int_{-\infty}^{\infty}(\nu-\bar{\nu})^2|F(\nu)|^2\,d\nu}{\int_{-\infty}^{\infty}|F(\nu)|^2\,d\nu}, \quad \text{where} \quad \bar{\nu} = \frac{\int_{-\infty}^{\infty}\nu|F(\nu)|^2\,d\nu}{\int_{-\infty}^{\infty}|F(\nu)|^2\,d\nu}.$$

Thus the time duration and the spectral width cannot simultaneously be made arbitrarily small. The *Gaussian function* $f(t) = \exp(-t^2/4\sigma_t^2)$, for example, has a

power-RMS width σ_t. Its Fourier transform is also a Gaussian function, $F(\nu) = (1/2\sqrt{\pi}\,\sigma_\nu)\exp(-\nu^2/4\sigma_\nu^2)$, with power-RMS width

$$\sigma_\nu = \frac{1}{4\pi\sigma_t}. \tag{A.2-4}$$

Since $\sigma_t\sigma_\nu = 1/4\pi$, the Gaussian function has the minimum permissible value of the duration–bandwidth product. In terms of the angular frequency $\omega = 2\pi\nu$,

$$\sigma_t\sigma_\omega \geq \tfrac{1}{2}. \tag{A.2-5}$$

If the variables t and ω, which usually describe time and angular frequency (rad/s), are replaced with the position variable x and the spatial angular frequency k (rad/m), respectively, then (A.2-5) becomes

$$\sigma_x\sigma_k \geq \tfrac{1}{2}. \tag{A.2-6}$$

In quantum mechanics, the position x of a particle is described by the wavefunction $\psi(x)$, and the wavenumber k is described by a function $\phi(k)$, which is the Fourier transform of $\psi(x)$. The uncertainties of x and k are the RMS widths of the probability densities $|\psi(x)|^2$ and $|\phi(k)|^2$, respectively, so that σ_x and σ_k are interpreted as the uncertainties of position and wavenumber. Since the particle momentum is $p = \hbar k$ (where $\hbar = h/2\pi$ and h is Planck's constant), the position–momentum uncertainty product satisfies the inequality

$$\boxed{\sigma_x\sigma_p \geq \frac{\hbar}{2},} \tag{A.2-7}$$

Heisenberg Uncertainty Relation

which is known as the **Heisenberg position–momentum uncertainty relation**.

The Power-Equivalent Width

The power-equivalent width of a signal $f(t)$ is the signal energy divided by the peak signal power. If $f(t)$ has its peak value at $t = 0$, for example, then the power-equivalent width is

$$\tau = \int_{-\infty}^{\infty} \frac{|f(t)|^2}{|f(0)|^2}\,dt. \tag{A.2-8}$$

The *double-sided exponential function* $f(t) = \exp(-|t|/\tau)$, for example, has a power-equivalent width τ, as does the Gaussian function $f(t) = \exp(-\pi t^2/2\tau^2)$. This definition is used in Sec. 12.1, where the coherence time of light is defined as the power-equivalent width of the complex degree of temporal coherence.

The power-equivalent spectral width is similarly defined by

$$\mathcal{B} = \int_{-\infty}^{\infty} \frac{|F(\nu)|^2}{|F(0)|^2}\,d\nu. \tag{A.2-9}$$

If $f(t)$ is real, so that $|F(\nu)|^2$ is symmetric, and if it has its peak value at $\nu = 0$, the power-equivalent spectral width is usually defined as the positive-frequency width,

$$B = \int_{0}^{\infty} \frac{|F(\nu)|^2}{|F(0)|^2}\,d\nu. \tag{A.2-10}$$

In the case when $F(\nu) = \tau/(1 + j2\pi\nu\tau)$, for example, we have

$$B = \frac{1}{4\tau}. \tag{A.2-11}$$

This definition is used in Sec. 19.6A to describe the bandwidth of photodetector circuits susceptible to photon and circuit noise (see also Prob. 19.6-5).

Using Parseval's theorem (A.1-7), together with the relation $F(0) = \int_{-\infty}^{\infty} f(t)\,dt$, (A.2-10) may be written in the form

$$B = \frac{1}{2T}, \tag{A.2-12}$$

where

$$T = \frac{\left[\int_{-\infty}^{\infty} f(t)\,dt\right]^2}{\int_{-\infty}^{\infty} f^2(t)\,dt} \tag{A.2-13}$$

is yet another definition of the time duration [the square of the area under $f(t)$ divided by the area under $f^2(t)$]. In this case, the duration–bandwidth product $BT = 1/2$.

The 1/e-, Half-Maximum, and 3-dB Widths

Another type of measure of the width of a function is its duration at a prescribed fraction of its maximum value ($1/\sqrt{2}$, $1/2$, $1/e$, or $1/e^2$, are examples). Either the half-width or the full width on both sides of the peak may be used. Two commonly encountered measures are the full-width at half-maximum (FWHM) and the half-width at $1/\sqrt{2}$-maximum, called the 3-dB width. The following are three important examples:

- The *exponential function* $f(t) = \exp(-t/\tau)$ for $t \geq 0$ and $f(t) = 0$ for $t < 0$, which describes the response of a number of electrical and optical systems, has a $1/e$-maximum width $\Delta t_{1/e} = \tau$. The magnitude of its Fourier transform $F(\nu) = \tau/(1 + j2\pi\nu\tau)$ has a 3-dB width (half-width at $1/\sqrt{2}$-maximum)

$$\Delta\nu_{\text{3-dB}} = \frac{1}{2\pi\tau}. \tag{A.2-14}$$

- The *double-sided exponential function* $f(t) = \exp(-|t|/\tau)$ has a half-width at $1/e$-maximum $\Delta t_{1/e} = \tau$. Its Fourier transform $F(\nu) = 2\tau/[1 + (2\pi\nu\tau)^2]$, known as the *Lorentzian distribution*, has a full-width at half-maximum

$$\Delta\nu_{\text{FWHM}} = \frac{1}{\pi\tau}, \tag{A.2-15}$$

  and is usually written in the form $F(\nu) = (\Delta\nu/2\pi)/[\nu^2 + (\Delta\nu/2)^2]$, where $\Delta\nu = \Delta\nu_{\text{FWHM}}$. The Lorentzian distribution describes the spectrum of certain light emissions (see Sec. 14.3D).
- The *Gaussian function* $f(t) = \exp(-t^2/2\tau^2)$ has a full-width at $1/e$-maximum $\Delta t_{1/e} = 2\sqrt{2}\,\tau$. Its Fourier transform $F(\nu) = \sqrt{2\pi}\,\tau \exp(-2\pi^2\tau^2\nu^2)$ has a full-width at $1/e$-maximum

$$\Delta\nu_{1/e} = \frac{\sqrt{2}}{\pi\tau} \tag{A.2-16}$$

and a full-width at half-maximum

$$\Delta\nu_{\text{FWHM}} = \frac{\sqrt{2\ln 2}}{\pi\tau}, \tag{A.2-17}$$

so that

$$\Delta\nu_{\text{FWHM}} = \sqrt{\ln 2}\,\Delta\nu_{1/e} = 0.833\,\Delta\nu_{1/e}\,. \tag{A.2-18}$$

The Gaussian function is also used to describe the spectrum of certain light emissions (see Sec. 14.3D), as well as to describe the spatial distribution of light beams (see Sec. 3.1).

A.3 TWO-DIMENSIONAL FOURIER TRANSFORM

We now consider a function of two variables $f(x, y)$. If x and y represent the coordinates of a point in a two-dimensional space, then $f(x, y)$ represents a spatial pattern (e.g., the optical field in a given plane). The harmonic function $F\exp[-j2\pi(\nu_x x + \nu_y y)]$ is regarded as a building block from which other functions may be composed by superposition. The variables ν_x and ν_y represent spatial frequencies in the x and y directions, respectively. Since x and y have units of length (mm), ν_x and ν_y have units of cycles/mm, or lines/mm. Examples of two-dimensional harmonic functions are illustrated in Fig. A.3-1.

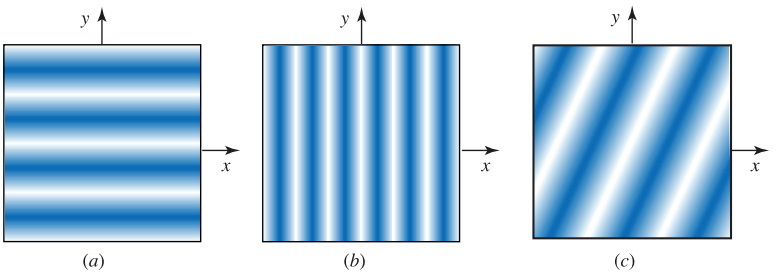


Figure A.3-1 The real part, $|F|\cos[2\pi\nu_x x + 2\pi\nu_y y + \arg\{F\}]$, of a two-dimensional harmonic function: (*a*) $\nu_x = 0$; (*b*) $\nu_y = 0$; (*c*) arbitrary ν_x and ν_y. For this illustration we have assumed that $\arg\{F\} = 0$ so that the white and dark regions represent positive and negative values of the function, respectively.

The Fourier theorem may be generalized to functions of two variables. A function $f(x, y)$ may be decomposed as a superposition integral of harmonic functions of x and y,

$$f(x,y) = \int_{-\infty}^{\infty}\!\!\int F(\nu_x,\nu_y)\exp\left[-j2\pi(\nu_x x + \nu_y y)\right]\,d\nu_x\,d\nu_y\,, \tag{A.3-1}$$

Inverse Fourier Transform

where the coefficients $F(\nu_x, \nu_y)$ are determined by use of the two-dimensional Fourier transform

$$F(\nu_x, \nu_y) = \iint_{-\infty}^{\infty} f(x, y) \exp\left[j2\pi(\nu_x x + \nu_y y)\right] \, dx \, dy. \tag{A.3-2}$$

Fourier Transform

Our definitions of the two- and one-dimensional Fourier transforms, (A.3-2) and (A.1-2), respectively, differ in the signs of their exponents. The choice of these signs are, of course, arbitrary, as long as opposite signs are used in the Fourier and inverse Fourier transforms. In this book we have adopted the convention that $\exp(j2\pi\nu t)$ has positive temporal frequency ν, while $\exp[-j2\pi(\nu_x x + \nu_y y)]$ has positive spatial frequencies ν_x and ν_y. We have elected to use different signs in the spatial (two-dimensional) and temporal (one-dimensional) cases to simplify the notation used in Chapter 4 (Fourier optics), in which the traveling wave $\exp(+j2\pi\nu t)\exp[-j(k_x x + k_y y + k_z z)]$ has temporal and spatial dependences of opposite sign.

Properties

Many properties of the two-dimensional Fourier transform are clear generalizations of those of the one-dimensional Fourier transform, but others are unique to the two-dimensional case:

- *Convolution Theorem.* The two-dimensional convolution of two functions, $f_1(x, y)$ and $f_2(x, y)$, with Fourier transforms $F_1(\nu_x, \nu_y)$ and $F_2(\nu_x, \nu_y)$, respectively, is written as

$$f(x, y) = \iint_{-\infty}^{\infty} f_1(x', y') f_2(x - x', y - y') \, dx' \, dy'. \tag{A.3-3}$$

  The Fourier transform of the convolution $f(x, y)$ is

$$F(\nu_x, \nu_y) = F_1(\nu_x, \nu_y) F_2(\nu_x, \nu_y), \tag{A.3-4}$$

  so that convolution in the spatial domain is equivalent to multiplication in the Fourier domain, as in the one-dimensional case.
- *Separable Functions.* If $f(x, y) = f_x(x) f_y(y)$ is the product of one function of x and another of y, then its two-dimensional Fourier transform is a product of one function of ν_x and another of ν_y. The two-dimensional Fourier transform of $f(x, y)$ is then related to the product of the one-dimensional Fourier transforms of $f_x(x)$ and $f_y(y)$ by $F(\nu_x, \nu_y) = F_x(-\nu_x) F_y(-\nu_y)$. We provide two examples: (1) the Fourier transform of $\delta(x - x_0)\,\delta(y - y_0)$, which represents an impulse located at (x_0, y_0), is the harmonic function $\exp[j2\pi(\nu_x x_0 + \nu_y y_0)]$; (2) the Fourier transform of the Gaussian function $\exp[-\pi(x^2 + y^2)]$ is the Gaussian function $\exp[-\pi(\nu_x^2 + \nu_y^2)]$.
- *Circularly Symmetric Functions.* The Fourier transform of a circularly symmetric function is also circularly symmetric. Consider, for example, the *circ function*, which is denoted by the symbol $\mathrm{circ}(x, y)$, and is given by

$$f(x, y) = \begin{cases} 1, & \sqrt{x^2 + y^2} \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{A.3-5}$$

Its Fourier transform is given by

$$F(\nu_x, \nu_y) = \frac{J_1(2\pi\nu_\rho)}{\nu_\rho}, \quad \nu_\rho = \sqrt{\nu_x^2 + \nu_y^2}, \tag{A.3-6}$$

where J_1 is the Bessel function of order 1. Both of these circularly symmetric functions are illustrated in Fig. A.3-2.

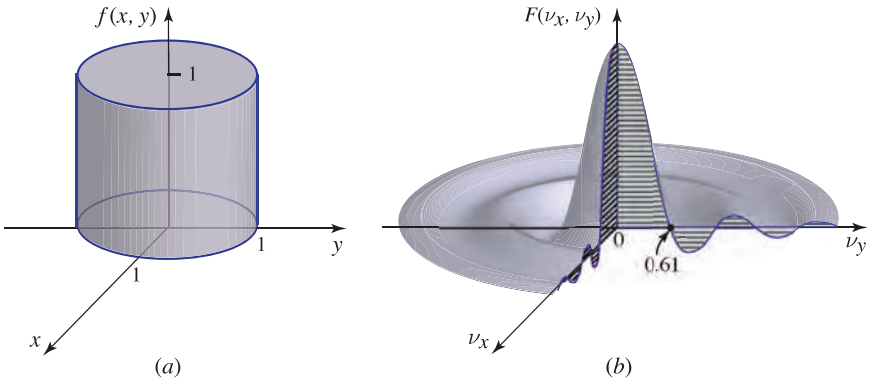


Figure A.3-2 (a) The circ function and (b) its two-dimensional Fourier transform.

READING LIST

L. Chaparro, *Signals and Systems using MATLAB*, Academic Press, 2nd ed. 2015.

G. R. Cooper and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, Oxford University Press, 3rd ed. 2007.

B. P. Lathi, *Linear Systems and Signals*, Oxford University Press, 2nd ed. 2004.

S. Haykin and B. Van Veen, *Signals and Systems*, Wiley, 2nd ed. 2003.

A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*, Prentice Hall, 1983, 2nd ed. 1997.

R. N. Bracewell, *The Fourier Transform and Its Applications*, McGraw–Hill, 3rd ed. 2000.

J. D. Gaskill, *Linear Systems, Fourier Transforms, and Optics*, Wiley, 1978.

L. E. Franks, *Signal Theory*, Prentice Hall, 1969, revised ed. 1981.
1986.

APPENDIX

B

LINEAR SYSTEMS

This appendix provides a review of the essential characteristics of one- and two-dimensional linear systems.

B.1 ONE-DIMENSIONAL LINEAR SYSTEMS

Consider a system whose input and output are the functions $f_1(t)$ and $f_2(t)$, respectively. The system is characterized by a rule that relates the output to the input. In general, the rule may take the form of a simple mathematical operation such as $f_2(t) = \log[f_1(t)]$, an integral transform, or a differential equation. An example is a harmonic oscillator that undergoes a displacement $f_2(t)$ in response to a time-varying force $f_1(t)$.

Linear Systems

A system is said to be *linear* if it satisfies the principle of superposition, i.e., if its response to the sum of any two inputs is the sum of its responses to each of the inputs separately. The output at time t is, in general, a weighted superposition of the input contributions at different times τ,

$$f_2(t) = \int_{-\infty}^{\infty} h(t;\tau) f_1(\tau)\, d\tau, \tag{B.1-1}$$

where $h(t;\tau)$ is a weighting function representing the contribution of the input at time τ to the output at time t. If the input is an impulse at time τ, so that $f_1(t) = \delta(t-\tau)$, then (B.1-1) yields $f_2(t) = h(t;\tau)$. Thus $h(t;\tau)$ is the **impulse response function** of the system (also known as the **Green's Function**).

Linear Shift-Invariant Systems

A linear system is said to be **time-invariant** or **shift-invariant** if, when its input is shifted in time, its output shifts by an equal time, but otherwise remains the same. The impulse response function is then a function of the time difference $h(t;\tau) = h(t-\tau)$. Under these conditions (B.1-1) becomes

$$\boxed{f_2(t) = \int_{-\infty}^{\infty} h(t-\tau) f_1(\tau)\, d\tau.} \tag{B.1-2}$$

The output $f_2(t)$ is then the convolution of the input $f_1(t)$ with the impulse response function $h(t)$ [see (A.1-4)]. If $f_1(t) = \delta(t)$, then $f_2(t) = h(t)$; if $f_1(t) = \delta(t-\tau)$, then $f_2(t) = h(t-\tau)$, as illustrated in Fig. B.1-1.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.
©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

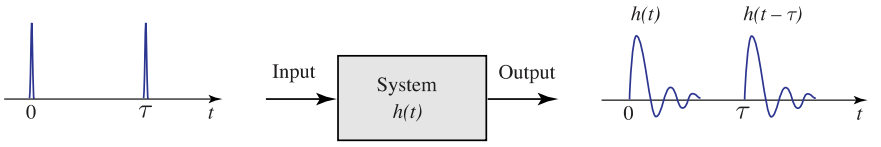


Figure B.1-1 Response of a linear shift-invariant system to impulses.

The Transfer Function

In accordance with the convolution theorem discussed in Appendix A [see (A.1-3)], the Fourier transforms $F_1(\nu)$, $F_2(\nu)$, and $H(\nu)$ of $f_1(t)$, $f_2(t)$, and $h(t)$, respectively, are related by

$$F_2(\nu) = H(\nu)F_1(\nu). \tag{B.1-3}$$

If the input $f_1(t)$ is a harmonic function $F_1(\nu)\exp(j2\pi\nu t)$, the output $f_2(t) = H(\nu)F_1(\nu)\exp(j2\pi\nu t)$ is also a harmonic function of the same frequency but with a modified complex amplitude $F_2(\nu) = F_1(\nu)H(\nu)$, as illustrated in Fig. B.1-2. The multiplicative factor $H(\nu)$ is known as the system's **transfer function**; it is the Fourier transform of the impulse response function. Equation (B.1-3) embodies the essence of the usefulness of Fourier methods in the analysis of linear shift-invariant systems. To determine the output of a system for an arbitrary input, we simply decompose the input into its harmonic components, multiply the complex amplitude of each harmonic function by the transfer function at the appropriate frequency, and superpose the resultant harmonic functions.

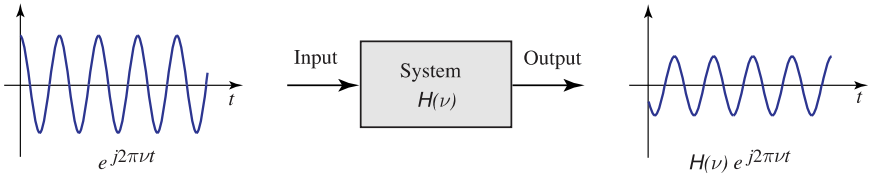


Figure B.1-2 Response of a linear shift-invariant system to a harmonic function.

Examples

- *Ideal system:* $H(\nu) = 1$ and $h(t) = \delta(t)$; the output is a replica of the input.
- *Ideal system with delay:* $H(\nu) = \exp(-j2\pi\nu\tau)$ and $h(t) = \delta(t - \tau)$; the output is a replica of the input delayed by time τ.
- *System with exponential response:* $H(\nu) = \tau/(1 + j2\pi\nu\tau)$ and $h(t) = e^{-t/\tau}$ for $t \geq 0$, and $h(t) = 0$ otherwise; this represents the response of a system described by a first-order linear differential equation, e.g., that representing an RC circuit with time constant τ. An impulse at the input results in an exponentially decaying response.
- *Chirped system:* $H(\nu) = \exp(-j\pi\nu^2)$ and $h(t) = e^{-j\pi/4}\exp(j\pi t^2)$; the system distorts the input by imparting to it a phase shift proportional to ν^2. An impulse at the input generates an output in the form of a chirped signal, i.e., a harmonic function whose instantaneous frequency (the derivative of the phase) increases linearly with time. This system describes the propagation of optical pulses through media with a frequency-dependent phase velocity; it also describes changes in the spatial distribution of light waves as they propagate through free space (see Secs. 23.3A and 4.1C, respectively).

Linear Shift-Invariant Causal Systems

The impulse response function $h(t)$ of a linear shift-invariant *causal* system must vanish for $t < 0$ since the system's response cannot begin before the application of the input. The function $h(t)$ is therefore not symmetric and its Fourier transform, the transfer function $H(\nu)$, must be complex. It can be shown[†] that if $h(t) = 0$ for $t < 0$, then the real and imaginary parts of $H(\nu)$, denoted $H'(\nu)$ and $H''(\nu)$, respectively, are related by

$$H'(\nu) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{H''(s)}{s-\nu}\,ds \tag{B.1-4}$$

$$H''(\nu) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{H'(s)}{\nu-s}\,ds, \tag{B.1-5}$$

Hilbert Transform

where the Cauchy principal values of the integrals are to be evaluated, i.e.,

$$\int_{-\infty}^{\infty} \equiv \lim_{\Delta\to 0}\left(\int_{-\infty}^{\nu-\Delta} + \int_{\nu+\Delta}^{\infty}\right), \quad \Delta > 0.$$

Functions that satisfy (B.1-4) and (B.1-5) are said to form a **Hilbert transform pair**, $H''(\nu)$ being the **Hilbert transform** of $H'(\nu)$.

If the impulse response function $h(t)$ is also real, its Fourier transform must be symmetric, $H(-\nu) = H^*(\nu)$ (see Appendix A, Sec. A.1). As a result, the real part $H'(\nu)$ then has even symmetry, and the imaginary part $H''(\nu)$ has odd symmetry. The integrals in (B.1-4) and (B.1-5) may then be rewritten as integrals over the interval $(0, \infty)$, and the resultant equations are known as the **Kramers–Kronig relations**:

$$H'(\nu) = \frac{2}{\pi}\int_{0}^{\infty} \frac{sH''(s)}{s^2-\nu^2}\,ds \tag{B.1-6}$$

$$H''(\nu) = \frac{2}{\pi}\int_{0}^{\infty} \frac{\nu H'(s)}{\nu^2-s^2}\,ds. \tag{B.1-7}$$

Kramers–Kronig Relations

In summary, the Hilbert-transform relations, or the Kramers–Kronig relations, relate the real and imaginary parts of the transfer function of a linear shift-invariant causal system, so that if one part is known at all frequencies, the other part may be determined.

Example: The Harmonic Oscillator

The linear system described by the differential equation

$$\left(\frac{d^2}{dt^2} + \zeta\frac{d}{dt} + \omega_0^2\right) f_2(t) = f_1(t) \tag{B.1-8}$$

describes a harmonic oscillator with displacement $f_2(t)$ under an applied force $f_1(t)$, where ω_0 is the resonance angular frequency and ζ is a coefficient representing damping effects. The transfer function $H(\nu)$ of this system may be obtained by substituting

[†] See, e.g., L. E. Franks, *Signal Theory*, Prentice Hall, 1969, revised ed. 1981.

$f_1(t) = \exp(j2\pi\nu t)$ and $f_2(t) = H(\nu)\exp(j2\pi\nu t)$ in (B.1-8), which yields

$$H(\nu) = \frac{1}{(2\pi)^2}\frac{1}{\nu_0^2 - \nu^2 + j\nu\Delta\nu}, \tag{B.1-9}$$

where $\nu_0 = \omega_0/2\pi$ is the resonance frequency, and $\Delta\nu = \zeta/2\pi$. The real and imaginary parts of $H(\nu)$ are therefore, respectively,

$$H'(\nu) = \frac{1}{(2\pi)^2}\frac{\nu_0^2 - \nu^2}{(\nu_0^2 - \nu^2)^2 + (\nu\Delta\nu)^2} \tag{B.1-10}$$

$$H''(\nu) = -\frac{1}{(2\pi)^2}\frac{\nu\Delta\nu}{(\nu_0^2 - \nu^2)^2 + (\nu\Delta\nu)^2}. \tag{B.1-11}$$

Since the system is causal, $H'(\nu)$ and $H''(\nu)$ satisfy the Kramers–Kronig relations. When $\nu_0 \gg \Delta\nu$, $H'(\nu)$ and $H''(\nu)$ are narrow functions centered about ν_0. For $\nu \approx \nu_0$, $(\nu_0^2 - \nu^2) \approx 2\nu_0(\nu_0 - \nu)$, whereupon (B.1-10) and (B.1-11) may be approximated by

$$H''(\nu) = -\frac{1}{(2\pi)^2}\frac{\Delta\nu/4\nu_0}{(\nu_0 - \nu)^2 + (\Delta\nu/2)^2} \tag{B.1-12}$$

$$H'(\nu) = 2\frac{\nu - \nu_0}{\Delta\nu}H''(\nu). \tag{B.1-13}$$

Equation (B.1-12) has a Lorentzian form. The transfer function of the harmonic-oscillator system is used in Secs. 5.5 and 15.1 to describe dielectric and atomic systems.

B.2 TWO-DIMENSIONAL LINEAR SYSTEMS

A two-dimensional system relates a pair of two-dimensional functions, $f_1(x, y)$ and $f_2(x, y)$, called the input and output functions. These functions may, for example, represent optical fields at two parallel planes, with (x, y) representing position variables; the system comprises the free space and optical components that lie between the two planes.

The concepts of linearity and shift invariance defined in the one-dimensional case are easily generalized to the two-dimensional case. The output $f_2(x, y)$ of a *linear* system is related to its input $f_1(x, y)$ by a superposition integral

$$f_2(x, y) = \int_{-\infty}^{\infty}\!\!\int h(x, y; x', y')f_1(x', y')\,dx'\,dy', \tag{B.2-1}$$

where $h(x, y; x', y')$ is a weighting function that represents the effect of the input at the point (x', y') on the output at the point (x, y). The function $h(x, y; x', y')$ is the **impulse response function** of the system (also known as the **point-spread function**).

The system is said to be **shift-invariant** (or **isoplanatic**) if shifting its input in some direction shifts the output by the same distance and in the same direction without otherwise altering it (see Fig. B.2-1). The impulse response function then depends on differences of position, $h(x, y; x', y') = h(x - x', y - y')$, whereupon (B.2-1) becomes

the two-dimensional convolution of $h(x, y)$ with $f_1(x, y)$:

$$f_2(x, y) = \int_{-\infty}^{\infty}\int f_1(x', y')\, h(x - x', y - y')\, dx'\, dy'. \tag{B.2-2}$$

Applying the two-dimensional convolution result provided in (A.3-4) of Appendix A yields

$$F_2(\nu_x, \nu_y) = H(\nu_x, \nu_y)\, F_1(\nu_x, \nu_y), \tag{B.2-3}$$

where $F_2(\nu_x, \nu_y)$, $H(\nu_x, \nu_y)$, and $F_1(\nu_x, \nu_y)$ are the Fourier transforms of $f_2(x, y)$, $h(x, y)$, and $f_1(x, y)$, respectively.

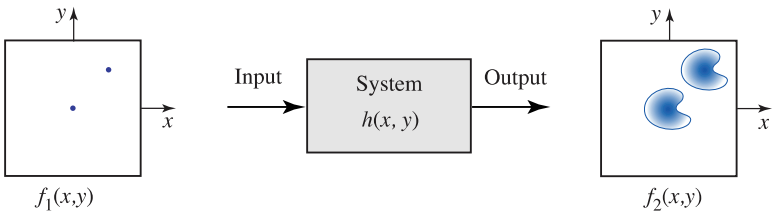


Figure B.2-1 Response of a two-dimensional linear shift-invariant system.

A harmonic input of complex amplitude $F_1(\nu_x, \nu_y)$ therefore produces a harmonic output of the same spatial frequency but with complex amplitude $F_2(\nu_x, \nu_y) = H(\nu_x, \nu_y)\, F_1(\nu_x, \nu_y)$, as illustrated in Fig. B.2-2. The multiplicative factor $H(\nu_x, \nu_y)$ is the system **transfer function**, which is the Fourier transform of its impulse response function. Either of these functions allows us to characterize the system completely and enables us to determine the output for an arbitrary input.

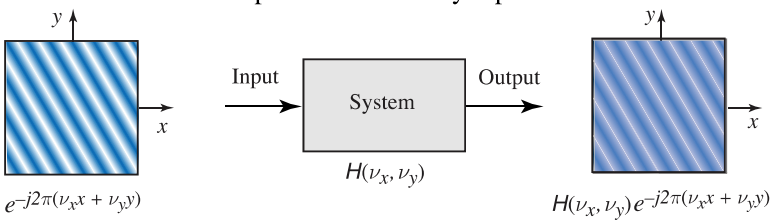


Figure B.2-2 Response of a two-dimensional linear shift-invariant system to harmonic functions.

In summary, a two-dimensional linear shift-invariant system is characterized by its impulse response function $h(x, y)$ or its transfer function $H(\nu_x, \nu_y)$. For example, a system with $h(x, y) = \text{circ}(x/\rho_s, y/\rho_s)$ smears each point of the input into a patch in the form of a circle of radius ρ_s. It has a transfer function $H(\nu_x, \nu_y) = \rho_s J_1(2\pi\rho_s\nu_\rho)/\nu_\rho$, where $\nu_\rho = \sqrt{\nu_x^2 + \nu_y^2}$, which has the shape illustrated in Fig. A.3-2(b). The system severely attenuates spatial frequencies higher than $0.61/\rho_s$ lines/mm.

READING LIST

See the reading list in Appendix A.

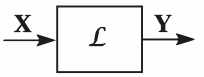
APPENDIX

C

MODES OF LINEAR SYSTEMS

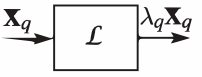
This Appendix provides a brief overview of modes of linear systems that are described explicitly by input–output relations that take the form of a matrix or integral operation, or implicitly by a linear ordinary or linear partial differential equation.

Consider first a linear system described by an explicit input–output relation characterized by a linear operator $\mathcal{L}$ that operates on an input vector $\mathbf{X}$ to generate a corresponding output vector $\mathbf{Y}$:



$$\mathbf{Y} = \mathcal{L}\mathbf{X}. \tag{C.1-1}$$

The vector $\mathbf{X}$ may be an array of complex numbers, represented by a column matrix, or a complex function of one or more variables. The **modes** of such a system are those special inputs that remain unaltered (except for a multiplicative constant) upon passage through the system. They thus obey



$$\mathcal{L}\mathbf{X}_q = \lambda_q\mathbf{X}_q, \tag{C.1-2}$$

Eigenvalue Problem

where q is an index that labels the mode. The vector $\mathbf{X}_q$ is known as the **eigenvector**, and the associated multiplicative constant λ_q, which is generally a complex number, is called the **eigenvalue**. The condition set forth in (C.1-2) is known as an **eigenvalue problem**.

Consider next a linear dynamical system whose state is described by N continuous variables constituting a vector $\mathbf{X}(t)$. The evolution of *any* of the N variables of this N-dimensional vector is, in general, dependent on all N variables. However, the same system may be described in a new coordinate system whereupon the N new variables evolve independently, so that the description of the system decomposes into N independent one-dimensional systems. These decoupled variables are the modes of the system.

Consider finally a linear system characterized implicitly by a linear partial differential equation that may be cast in the form of (C.1-2), where $\mathcal{L}$ is a differential operator and $\mathbf{X}$ is a complex function of one or several variables. In this case, the modes are simply solutions of the differential equation and the eigenvectors are called **eigenfunctions**. The notion of an input and an output is not meaningful in these circumstances.

We proceed to describe a number of applications of modal analysis in photonics. Before commencing, however, we briefly review a number of geometrical concepts from linear algebra. Associated with each pair of vectors $\mathbf{X}$ and $\mathbf{Y}$ is a complex scalar quantity $(\mathbf{X}, \mathbf{Y})$ called the **inner product**. The square root of the inner product of a vector $\mathbf{X}$ with itself, $(\mathbf{X}, \mathbf{X})$, is known as the **norm** of $\mathbf{X}$ and is a measure of its "length." The inner product of two vectors of unit norm can be thought of as the cosine of the "angle" between them. Two vectors are said to be **orthogonal** if their inner product is zero. If the vectors comprise arrays of complex numbers, $\{X_i\}$ and $\{Y_i\}$,

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.
©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

$i = 1, 2, \ldots, N$, then $(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N} X_i^* Y_i$. If, on the other hand, the vectors are complex functions $X(t)$ and $Y(t)$, then $(\mathbf{X}, \mathbf{Y}) = \int_{-\infty}^{\infty} X^*(t) Y(t) dt$.

Two classes of operators $\mathcal{L}$ that lead to solutions of the eigenvalue problem with special properties are considered in turn:

Hermitian operators. Hermitian operators are defined by the property $(\mathbf{X}, \mathcal{L}\mathbf{Y}) = (\mathcal{L}\mathbf{X}, \mathbf{Y})$, i.e., the inner product is the same no matter to which of the two vectors the operator is applied. The eigenvalues of a Hermitian operator are real and the eigenvectors are orthogonal. Further, the eigenvectors of a Hermitian operator obey the **variational principle**, which is based on a scalar $E_{\text{var}} = \frac{1}{2}(\mathbf{X}, \mathcal{L}\mathbf{X})/(\mathbf{X}, \mathbf{X})$, called the variational energy. This principle states that the eigenvector $\mathbf{X}_1$ with the lowest eigenvalue minimizes E_{var}; the eigenvector $\mathbf{X}_2$ with the next lowest eigenvalue minimizes E_{var}, subject to the condition that it is orthogonal to $\mathbf{X}_1$, and so on.

Unitary operators. Passive, lossless physical systems are described by unitary operators, which are defined by the norm-preserving property $(\mathcal{L}\mathbf{X}, \mathcal{L}\mathbf{X}) = (\mathbf{X}, \mathbf{X})$. An example is the "rotation" operation. The eigenvalues of unitary operators are unimodular, i.e., $|\lambda_q| = 1$, and therefore represent pure phase.

1) Modes of a Discrete Linear System

A discrete linear system is described by a matrix relation $\mathbf{Y} = \mathbf{M}\mathbf{X}$, where the input vector $\mathbf{X}$ is a set of N complex numbers $(X_1, X_2, \ldots, X_N)$ arranged in a column matrix, $\mathbf{M}$ is an $N \times N$ matrix that represents the linear system, and the output vector $\mathbf{Y}$ is also a column matrix of dimension N. The modes are those input vectors that remain parallel to themselves upon transmission through the system, so that the matrix equation

$$\mathbf{M}\mathbf{X}_q = \lambda_q \mathbf{X}_q \tag{C.1-3}$$

is obeyed. Thus, the modes of the system are the eigenvectors $\mathbf{X}_q$ of the matrix $\mathbf{M}$, and the scalars λ_q are the corresponding eigenvalues, which are determined by solving the algebraic equation $\det(\mathbf{M} - \lambda\mathbf{I}) = 0$, where $\mathbf{I}$ is the identity matrix. There are N such modes, labeled by the index $q = 1, 2, \ldots, N$.

The special case of binary systems ($N = 2$) is particularly important in optics. In a binary system, each vector is a pair of complex numbers (X_1, X_2) arranged in a column matrix $\mathbf{X}$. The system is characterized by a 2×2 square matrix $\mathbf{M}$ whose elements are denoted A, B, C, and D. The relation $\mathbf{Y} = \mathbf{M}\mathbf{X}$ signifies

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

The eigenvalues are determined by solving the algebraic equation $(A - \lambda)(D - \lambda) - BC = 0$ for the two eigenvalues λ_1 and λ_2.

The following are examples of optical systems described by binary linear systems:

Application: Polarization matrix optics. In polarization matrix optics (Sec. 6.1B), the vector (X_1, X_2) represents the components of the input electric field in two orthogonal directions (the Jones vector), and (Y_1, Y_2) similarly represents the output electric field. The matrix $\mathbf{M}$ is the Jones matrix of the system. In this case, the modes are the polarization states that are maintained as light is transmitted through the system.

Application: Ray matrix optics. In geometrical paraxial optics (Sec. 1.4), the position and angle of an optical ray are described by a vector (X_1, X_2), and the effect of optical components, such as lenses and mirrors, is described by a matrix $\mathbf{M}$, called the ray-transfer matrix or the $ABCD$ matrix. For a closed optical system, such as a resonator, the modes are ray positions and angles that self-reproduce after a round trip, so that they are confined within the resonator.

Application: Multilayer matrix optics. In multilayer matrix optics (Sec. 7.1A) light is reflected and refracted at each boundary, so that there are forward- and backward-traveling waves at each plane, with amplitudes described by a vector $\mathbf{X} = (X_1, X_2)$. A system containing a set of boundaries between an input and an output plane is described by a wave-transfer matrix $\mathbf{M}$. The modes of such a system are the vectors that self reproduce upon transmission through the system, so that if the system is replicated periodically, as in a 1D photonic crystal (Sec. 7.2), the propagation modes are the modes of the system $\mathbf{M}$.

2) Modes of a Continuous System Described by an Integral Operator

Linear systems represented by integral operators are discussed in Appendix B. Consider, for example, a function of time $f(t)$, such as an optical pulse or a broadband optical field, transmitted through a linear time-invariant system such as an optical filter. The system is described by the convolution operation (A.1-4):

$$g(t) = \int_{-\infty}^{\infty} h(t-\tau) f(\tau)\, d\tau. \tag{C.1-4}$$

In this system, the vectors $\mathbf{X}$ and $\mathbf{Y}$ are the functions $f(t)$ and $g(t)$, respectively, and the operator $\mathcal{L}$ is an integral operator. The modes of this system are the harmonic functions $\exp(j2\pi\nu t)$. This is evident since the input function $\exp(j2\pi\nu t)$ generates another harmonic output function $H(\nu)\exp(j2\pi\nu t)$, where $H(\nu)$ is the Fourier transform of $h(t)$. In this case, there is a continuum of modes with continuous eigenvalues $H(\nu)$. Here, the index q is the frequency ν, which takes continuous values.

Another example is a linear shift-invariant system that operates on a two-dimensional (2D) function $f(x, y)$ of the position (x, y), as described in (B.2-2):

$$g(x,y) = \iint_{-\infty}^{\infty} h(x-x', y-y') f(x', y')\, dx'\, dy'. \tag{C.1-5}$$

The eigenfunctions are 2D harmonic functions $\exp[j2\pi(\nu_x x + \nu_y y)]$, and the eigenvalues are $H(\nu_x, \nu_y)$, the 2D Fourier transform of $h(x, y)$. Again, there is a continuum of eigenfunctions, labeled by the spatial frequencies (ν_x, ν_y).

Translational symmetry and harmonic modes. It is not surprising that harmonic functions are the modes of a shift-invariant system. Because the harmonic function is invariant to time shift, i.e., it remains a harmonic function if translated in time, it is the eigenfunction of the time-invariant (stationary) linear system. Similarly, because 2D harmonic functions are invariant to translation in the plane, they are the eigenfunctions of the space-invariant (homogeneous) linear system.

If the linear system is not space-invariant, i.e., does not enjoy translational symmetry, then in the 2D case it is represented by the more general linear operation described in (B.2-1):

$$g(x,y) = \iint_{-\infty}^{\infty} h(x, y; x', y') f(x', y')\, dx'\, dy'. \tag{C.1-6}$$

The eigenfunctions, which are now not necessarily harmonic functions, are determined by solving the eigenvalue problem posed in (C.1-2), which in this case takes the form of an integral equation

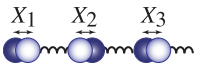
$$\iint_{-\infty}^{\infty} h(x, y; x', y')\, f_q(x', y')\, dx'\, dy' = \lambda_q\, f_q(x, y), \quad q = 1, 2, \ldots. \tag{C.1-7}$$

The functions $f_q(x, y)$ and the constants λ_q are the eigenfunctions and eigenvalues of the system, respectively, and the index q labels a discrete set of modes.

Application: Optical resonator modes. An example is provided by light traveling between the two parallel mirrors of a laser resonator (Sec. 11.2E). The distributions of the optical field in the transverse plane at the beginning and at the end of a single round trip are the input and output of the system. The modes of the resonator are those field distributions that maintain their form after one round trip. The kernel $h(x, y; x', y')$ in (C.1-7) represents propagation in free space and reflection from one of the mirrors, followed by backward free-space propagation and reflection from the other mirror. Clearly, the presence of curved mirrors, or mirrors of finite extent, makes this system shift-variant. If the mirrors are spherical and are assumed to modulate the incoming light by a phase factor that is a quadratic function of the radial distance, then the resonator modes are Hermite–Gaussian functions of x and y (see Sec. 3.3). In the presence of apertures, (C.1-7) can only be solved numerically, as considered in Sec. 11.2E.

3) Modes of a System Described by an Ordinary Differential Equation

The dynamics of certain physical systems are characterized by a set of coupled ordinary differential equations. For example, the dynamics of N coupled oscillators are described by N differential equations that are conveniently written in matrix form as



$$\ddot{\mathbf{X}} = -\mathbf{M}\mathbf{X}, \tag{C.1-8}$$

where $\mathbf{X}$ is a column matrix with components $(X_1, X_2, \ldots, X_N)$, $\ddot{\mathbf{X}} \equiv d^2\mathbf{X}/dt^2$, and $\mathbf{M}$ is an $N \times N$ matrix with time-independent coefficients, so that the system is time invariant.

Time invariance requires that the modes be harmonic functions that take the form $\exp(j\omega t)$, i.e., the vector $\mathbf{X}(t) = \mathbf{X}(0)\exp(j\omega t)$. Hence, substitution in (C.1-8) yields

$$\mathbf{M}\mathbf{X} = \omega^2\mathbf{X}. \tag{C.1-9}$$

This equation represents a discrete-system eigenvalue problem. Its eigenvalues provide the resonance frequencies $\omega_1, \omega_2, \ldots, \omega_N$ of the modes, and its eigenvectors are called the **normal modes**. All components of the eigenvector $\mathbf{X}_q$ of mode q oscillate at the same resonance frequency ω_q, without alteration of their relative amplitudes or phases. In this sense, the modes are stationary solutions that are decoupled from one another.

4) Modes of a System Described by a Partial Differential Equation

Fields and waves are described by partial differential equations such as Maxwell's equations, which characterize the dynamics of the electric and magnetic fields in a dielectric medium. Similarly, the Schrödinger equation expresses the dynamics of the wavefunction of a particle subject to a specified potential. If these physical systems are stationary, i.e., if the dielectric medium and the potential are time independent, then each mode must be a harmonic function of time that takes the form $\exp(j\omega t)$ with some frequency ω. The wave equation is then converted into the generalized Helmholtz equation

$$\nabla \times [\eta(\mathbf{r})\ \nabla \times \mathbf{H}] = \frac{\omega^2}{c_o^2}\mathbf{H}, \tag{C.1-10}$$

where $\eta(\mathbf{r}) = \epsilon_o/\epsilon(\mathbf{r})$ is the electric impermeability of the dielectric medium [see (7.0-2)]. Analogously, the Schrödinger equation (14.1-1) yields the time-independent Schrödinger equation (14.1-3),

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}), \tag{C.1-11}$$

where $V(\mathbf{r})$ is the potential and $E = \hbar\omega$.

Both of these equations take the form of an eigenvalue problem (C.1-2), where $\mathcal{L}$ is a Hermitian differential operator characterized by the function $\eta(\mathbf{r})$ or $V(\mathbf{r})$. The eigenvalues, which are real, provide the frequencies ω_q of the modes (and hence the corresponding energies E_q in the case of the Schrödinger equation). The eigenfunctions are the spatial distributions of the electromagnetic field (or the wavefunction) for each mode. Note that the field (or the wavefunction) of the qth mode evolves with time as $\exp(j\omega_q t)$ at all positions, so that each mode is stationary, as required.

Modes of fields/waves in a homogeneous medium with boundary conditions. If the dielectric medium is homogeneous, i.e., the impermeability $\eta(\mathbf{r})$ is constant, then the system is shift-invariant. To be consistent with this translational symmetry, the modes of the electromagnetic system must be harmonic functions of position, i.e., plane waves. Similarly, if the potential $V(\mathbf{r})$ is constant, then the modes are plane-wave wavefunctions, so that the particle is equally likely to be found anywhere.

In other situations, $\eta(\mathbf{r})$ and $V(\mathbf{r})$ are constant within a finite region bounded by a surface that imposes certain boundary conditions. For example, the electromagnetic modes of a cavity resonator with perfectly conducting surfaces can be determined by requiring that the parallel components of the electric field vanish at the surface. For a rectangular resonator, the modes are harmonic functions of position — standing waves oscillating in unison (see Sec. 11.3C). Similarly, the modes of a particle in a quantum box (dot) are obtained by requiring that the wavefunction vanishes at the boundaries (see Sec. 17.1G).

In yet another geometry, a homogeneous dielectric medium may be bounded in one direction, e.g., by two parallel planar mirrors. Here, the boundary conditions correspond to a discrete set of standing waves in the direction orthogonal to the mirrors (transverse direction), with traveling waves in the parallel (axial) direction, so that the modes travel in this optical waveguide as harmonic functions in the axial direction, without altering their transverse distributions (see Sec. 9.1). If β_q is the propagation constant of mode q, then the eigenvalue is the phase factor $\exp(-j\beta_q z)$.

Modes of fields/waves in a periodic medium. As is evident from the previous examples, the modes of a system described by a partial differential equation are dictated by the spatial distribution of the medium, e.g., the function $\eta(\mathbf{r})$ or $V(\mathbf{r})$. If this function is constant, the modes must be invariant to arbitrary translation. If it is periodic, then the modes must be invariant to translation by a period. This type of translational symmetry requires that the modes be Bloch waves (see Sec. 7.2A). For example, if the medium is homogeneous in the x and y directions but periodic in the z direction, a Bloch mode takes the form of a harmonic function $\exp(-jKz)$, modulated by a periodic standing wave $p_K(z)$ with period equal to that of the medium; the dependence on x and y is, of course, harmonic. For a given value of K, the frequencies of the modes and the shapes of the corresponding standing waves $p_K(z)$ depend on the shape of the periodic function $\eta(\mathbf{r})$ or $V(\mathbf{r})$. This type of translational symmetry results in a spectrum of eigenvalues (and hence frequencies ω or energies $E = \hbar\omega$) in the form of bands that are separated by bandgaps within which no modes are allowed. Thus, an electron in a periodic potential distribution exhibits the well-known band structure of solids (see Secs. 14.1D and 17.1A). Likewise, an optical field in a periodic dielectric medium, i.e., a photonic crystal, exhibits a band structure with photonic bandgaps (see Secs. 7.2 and 7.3).

READING LIST

D. C. Lay, S. R. Lay, and J. J. McDonald, *Linear Algebra and its Applications*, Pearson, 5th ed. 2015.

S. Axler, *Linear Algebra Done Right*, Springer-Verlag, 3rd ed. 2015.

S. J. Leon, *Linear Algebra with Applications*, Pearson, 9th ed. 2014.

G. Strang, *Introduction to Linear Algebra*, Wellesley Cambridge Press, 4th ed. 2009.

SYMBOLS AND UNITS

Roman Symbols and Acronyms

a = Radius of an aperture or fiber [m]; also, Radius of a spherical scattering particle [m]; also, Radius of a circle [m]; also, Distance between locations [m]; also, Lattice constant [m]; also, Length of a thin metallic rod [m]; also, Chirp parameter for an optical pulse

a_0 = Bohr radius (radius of ground state of Bohr hydrogen atom; $a_0 \approx 0.53$ Å) [m]

$\mathfrak{a}$ = Complex amplitude or magnitude of an optical wave; also, Normalized complex amplitude of an optical field ($|\mathfrak{a}|^2$ = photon-flux density)

a = Normalized field amplitude in a cavity ($|\mathit{a}|^2$ = field energy in units of photon number)

$\mathtt{a}$ = Acceleration of a carrier [$\mathrm{m \cdot s^{-2}}$]

$\mathbf{a}$ = Primitive vector defining a lattice unit cell [m]

$\boldsymbol{\mathfrak{a}}$ = Complex-amplitude vector

A = Complex envelope of a monochromatic plane wave; also, Pulse amplitude

$A(\mathbf{r})$ = Complex envelope of a monochromatic wave

$A(\nu)$ = Fourier transform of the complex envelope of an optical pulse

$\mathbf{A}$ = Complex vector envelope of a monochromatic plane wave; also, Vector potential [$\mathrm{V \cdot s \cdot m^{-1}}$]

$\mathcal{A}$ = Absorbance

$\mathcal{A}(\mathbf{r}, t)$ = Complex envelope of a polychromatic (e.g., pulsed) wave

$\mathcal{A}(t)$ = Complex envelope of an optical pulse

A = Area [$\mathrm{m^2}$]; also, Element of the ABCD ray-transfer and wave-transfer matrices $\mathbf{M}$

A_c = Coherence area [$\mathrm{m^2}$]

A_{ij} = Element of Jacobian transformation matrix

A_r = Relative atomic mass

$\mathtt{Ai}(\cdot)$ = Airy function

$\mathbf{A}$ = Jacobian transformation matrix

$\mathbb{A}$ = Einstein A coefficient [$\mathrm{s^{-1}}$]

AC = Alternating current

ACS = American Chemical Society

ADC = Analog-to-digital converter

ADM = Add–drop multiplexer

ADP = Ammonium dihydrogen phosphate

AGIL = All gas-phase iodine laser

AM = Amplitude modulation

AMLCD = Active-matrix liquid-crystal display

AMOLED = Active-matrix organic light-emitting display

AND = AND logic gate

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.
©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

AOC = Active optical cables
AOM = Acousto-optic modulator
APD = Avalanche photodiode
APS = American Physical Society
ASE = Amplified spontaneous emission
ASK = Amplitude shift keying
AWG = Arrayed waveguides

b = Radius of a circle $[\mathrm{m}]$; also, Chirp coefficient $[\mathrm{s}^2]$
$\mathfrak{b}$ = Bowing parameter
B = Magnetic flux-density complex amplitude $[\mathrm{Wb}\cdot\mathrm{m}^{-2}$ or $\mathrm{T}]$; also, Bandwidth $[\mathrm{Hz}]$; also, Bandwidth of an electrical circuit $[\mathrm{Hz}]$; also, Spatial bandwidth $[\mathrm{m}^{-1}]$; also, Spectral width supporting net gain in a laser medium $[\mathrm{Hz}]$
$\mathbf{B}$ = Magnetic flux-density complex amplitude vector $[\mathrm{Wb}\cdot\mathrm{m}^{-2}$ or $\mathrm{T}]$
B_0 = Bit rate $[\mathrm{b}\cdot\mathrm{s}^{-1}]$
$\mathcal{B}$ = Power-equivalent spectral width $[\mathrm{Hz}]$
$\boldsymbol{\mathcal{B}}$ = Magnetic flux density vector $[\mathrm{Wb}\cdot\mathrm{m}^{-2}$ or $\mathrm{T}]$
B = Element of the $ABCD$ ray-transfer and wave-transfer matrices $\mathbf{M}$
$\mathbb{B}$ = Einstein B coefficient $[\mathrm{m}^3\cdot\mathrm{J}^{-1}\cdot\mathrm{s}^{-2}]$
BBO = Beta barium borate
BER = Bit error rate
BGR = Bragg grating reflector
BPF = Bandpass filter
BPP = Bulk plasmon polariton
BPSK = Binary phase shift keying
BRF = Birefringent filter
BSO = Bismuth silicon oxide

c = Speed of light $[\mathrm{m}\cdot\mathrm{s}^{-1}]$; also, Phase velocity $[\mathrm{m}\cdot\mathrm{s}^{-1}]$
c_o = Speed of light in free space $[\mathrm{m}\cdot\mathrm{s}^{-1}]$
C = Electrical capacitance [F]
$C(\cdot)$ = Fresnel integral
$\mathcal{C}$ = Coupling coefficient in a directional coupler $[\mathrm{m}^{-1}]$
C = Element of the $ABCD$ ray-transfer and wave-transfer matrices $\mathbf{M}$
C-band = Conventional optical fiber telecommunications band (1530–1565 nm)
CAD = Computer-aided design
CAPD = Conventional avalanche photodiode
CARS = Coherent anti-Stokes Raman scattering
CATV = Cable television
CCD = Charge-coupled device
CCT = Correlated color temperature
CCW = Counterclockwise
CD = Compact-disc
CDM = Code-division multiplexing
CDMA = Code-division multiple access
CFL = Compact fluorescent lamp
CLSM = Confocal laser-scanning microscopy
CMOS = Complementary metal-oxide-semiconductor
COB = Chip-on-board light-emitting diode
COIL = Chemical oxygen–iodine laser

CPA = Chirped-pulse amplification
CRI = Color rendering index
CVD = Chemical vapor deposition
CW = Continuous-wave; also, Clockwise
CWDM = Coarse wavelength-division multiplexing

$d\mathbf{r}$ = Incremental volume [m^3]
ds = Incremental length [m]
d = Coefficient of second-order optical nonlinearity [$\mathrm{C}\cdot\mathrm{V}^{-2}$]
$\mathrm{d}_{\mathrm{eff}}$ = Effective coefficient of second-order optical nonlinearity [$\mathrm{C}\cdot\mathrm{V}^{-2}$]
d_{ijk} = Component of second-order optical nonlinearity tensor [$\mathrm{C}\cdot\mathrm{V}^{-2}$]
d_{iJ} = Component of second-order optical nonlinearity tensor (contracted indices) [$\mathrm{C}\cdot\mathrm{V}^{-2}$]
$\mathrm{d}(\omega_3;\omega_1,\omega_2)$ = Coefficient of second-order optical nonlinearity (dispersive medium) [$\mathrm{C}\cdot\mathrm{V}^{-2}$]
d = Distance, Length, Thickness [m]
d_{b} = Propagation length of a plasmon wave along its boundary [m]
d_{d} = Thickness of dielectric layer in a layered metamaterial [m]
d_{ex} = Mean distance traveled by a photon in a random laser before exiting [m]
d_{m} = Thickness of metallic layer in a layered metamaterial [m]
d_{min} = Minimum distance [m]
d_p = Penetration depth [m]
d_{p} = Distance of coupling prism from a waveguide [m]
d_{pulse} = Length of a mode-locked optical pulse [m]
d_{S} = Length along a small dimension [m]
d_{st} = Mean distance traveled by a photon in a random laser before stimulating a clone photon [m]
D = Diameter [m]; also, Electric flux-density complex amplitude [$\mathrm{C}\cdot\mathrm{m}^{-2}$]; also, Width of an optical beam [m]
D_s = Width of an acoustic beam [m]
D^* = Specific detectivity of a photodetector [$\mathrm{cm}\cdot\sqrt{\mathrm{Hz}}\cdot\mathrm{W}^{-1}$]
$\mathbf{D}$ = Electric flux-density complex amplitude vector [$\mathrm{C}\cdot\mathrm{m}^{-2}$]
D_w = Waveguide dispersion coefficient [$\mathrm{s}\cdot\mathrm{m}^{-2}$]
D_x, D_y = Lateral widths [m]
D_λ = Material dispersion coefficient [$\mathrm{s}\cdot\mathrm{m}^{-2}$]
D_ν = Material dispersion coefficient [$\mathrm{s}^2\cdot\mathrm{m}^{-1}$]
$\mathfrak{D}$ = Electric flux density vector [$\mathrm{C}\cdot\mathrm{m}^{-2}$]
D = Element of the $ABCD$ ray-transfer and wave-transfer matrices $\mathbf{M}$
DBR = Distributed Bragg reflector
DC = Direct current
DCF = Dispersion-compensating fiber
DD = Direct detection
DESY = Deutsches Elektronen-Synchrotron
DEW = Directed-energy weapon
DFB = Distributed-feedback
DFF = Dispersion-flattened fiber
DFG = Difference-frequency generation
DGD = Differential group delay
DH = Double-heterostructure
DIP = Dual-inline package
DKDP = Deuterated potassium dihydrogen phosphate
DMD = Digital micromirror device
DMUX = Demultiplexer (also abbreviated as DEMUX)

DNG = Double-negative medium
DPC = Digital photon-counting device
DPS = Double-positive medium
DPSS = Diode-pumped solid-state
DQPSK = Differential quaternary phase shift keying
DRO = Doubly resonant oscillator
DSF = Dispersion-shifted fiber
DSP = Digital signal processing
DSPP = Doubly stochastic Poisson process
DUV = Deep ultraviolet, stretching from 200 to 300 nm
DVD = Digital-video-disc
DWDM = Dense wavelength-division multiplexing
DWELL = Quantum-dot-in-well
DWELL-QDIP = Quantum-dot-in-well quantum-dot infrared photodetector

e = Magnitude of electron charge $[\mathrm{C}]$
$\widehat{\mathbf{e}}_x$ = Unit vector in the x direction
E = Electric-field complex amplitude $[\mathrm{V}\cdot\mathrm{m}^{-1}]$; also, Steady or slowly varying field $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
E_L = Local-oscillator electric-field complex amplitude $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
E_s = Signal electric-field complex amplitude $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
$\mathbf{E}$ = Electric-field complex amplitude vector $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
$\mathbf{E}_{\mathrm{i}}$ = Electric-field complex amplitude vector within a scattering sphere $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
$\mathbf{E}_{\mathrm{s}}$ = Scattered electric-field complex amplitude vector $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
$\mathbf{E}_0$ = Incident electric-field complex amplitude vector $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
$\mathfrak{E}$ = Electric field vector $[\mathrm{V}\cdot\mathrm{m}^{-1}]$
E = Energy [J]; also, Radiant energy [J]
E_A = Acceptor energy level [J]; also, Activation energy [J]
$\mathit{E}_{\mathrm{beam}}$ = Electron-beam energy in a free-electron laser $[\mathrm{GeV}]$
E_c = Energy at the bottom of the conduction band [J]
E_D = Donor energy level [J]
E_f = Fermi energy [J]
E_{fc} = Quasi-Fermi energy for the conduction band [J]
E_{fv} = Quasi-Fermi energy for the valence band [J]
E_g = Bandgap energy [J]
E_k = Kinetic energy [J]
$\mathit{E}_{\mathrm{max}}$ = Maximum kinetic energy [J]
E_{r} = Rotational energy [J]
E_v = Energy at the top of the valence band [J]
E_{v} = Luminous energy $[\mathrm{lm}\cdot\mathrm{s}]$
$\mathit{E}_{\mathrm{var}}$ = Variational energy
E_ν = Energy spectral density $[\mathrm{J}\cdot\mathrm{Hz}^{-1}]$
E-band = Extended optical fiber telecommunications band (1360–1460 nm)
EAM = Electroabsorption modulator
ECLD = External-cavity laser diode
EDFA = Erbium-doped fiber amplifier
e-e-o = Extraordinary-extraordinary-ordinary designations for waves 1, 2, and 3
EIT = Electromagnetically induced transparency
ELI = Extreme Light Infrastructure
E/O = Electronic-to-optical
e-o-e = Extraordinary-ordinary-extraordinary designations for waves 1, 2, and 3

e-o-o = Extraordinary-ordinary-ordinary designations for waves 1, 2, and 3
EQE = External quantum efficiency
EUV = Extreme-ultraviolet, stretching from 10 to 100 nm

f = Focal length of a lens $[\mathrm{m}]$; also, Frequency $[\mathrm{Hz}]$; also, Frequency of sound $[\mathrm{Hz}]$
$f(E)$ = Fermi function
f_a = Probability that absorption condition is satisfied
$f_c(E)$ = Fermi function for the conduction band
f_{col} = Collision rate $[\mathrm{s}^{-1}]$
f_e = Probability that emission condition is satisfied
f_g = Fermi inversion factor
$f_v(E)$ = Fermi function for the valence band
f = Frequency of sound $[\mathrm{Hz}]$; also, Modulation frequency $[\mathrm{Hz}]$
f = Volume fraction of a homogeneous medium occupied by scatterers (filling ratio)
F = Focal point of an optical system; also, Excess noise factor of a photodetector
F_P = Purcell factor
$F_{\#}$ = F-number of a lens
$\mathcal{F}$ = Finesse of a resonator; also, Force $[\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-2}]$
$\mathcal{F}_{\mathrm{p}}$ = Ponderomotive force $[\mathrm{kg}\cdot\mathrm{m}\cdot\mathrm{s}^{-2}]$
$F_{m\ell}$ = Elements of the matrix $\mathbf{F}$
$\mathbf{F}$ = Hermitian matrix for generalized Helmholtz equation posed as an eigenvalue problem
FB = Fiber bundle
FBG = Fiber Bragg grating
FDM = Frequency-division multiplexing
FDMA = Frequency-division multiple access
FEL = Free-electron laser
FET = Field-effect transistor
FFT = Fast Fourier transform
FIR = Far infrared, stretching from 20 to 300 μm
FM = Frequency-modulated
FON = Fiber-optic network
FPA = Focal-plane array
FPI = Fabry–Perot interferometer
FROG = Frequency-resolved optical gating
FSK = Frequency shift keying
FTIR = Frustrated total internal reflection
FUV = Far ultraviolet, stretching from 100 to 200 nm
FWHM = Full-width at half-maximum
FWM = Four-wave mixing

g = Resonator g-parameter; also, Gravitational acceleration constant at earth's surface $[\mathrm{m}\cdot\mathrm{s}^{-2}]$
$g(\mathbf{r}_1,\mathbf{r}_2)$ = Normalized mutual intensity
$g(\mathbf{r}_1,\mathbf{r}_2,\tau)$ = Complex degree of coherence
$g(\nu)$ = Lineshape function of a transition $[\mathrm{Hz}^{-1}]$
$g(\tau)$ = Complex degree of temporal coherence
g_0 = Gain factor
$g_{\nu 0}(\nu)$ = Electron–photon collisionally broadened lineshape function in a semiconductor $[\mathrm{Hz}^{-1}]$
g = Coupling coefficient in a parametric interaction $[\mathrm{m}^{-3}]$
g = Fundamental spatial frequency of a periodic structure; also, Degeneracy parameter
$\mathbf{g}$ = Primitive vector defining a reciprocal-lattice unit cell $[\mathrm{m}^{-1}]$

G = Gain of an amplifier; also, Gain of a photodetector; also, Conductance $[\Omega^{-1}]$
$G(\mathbf{r}_1, \mathbf{r}_2)$ = Mutual intensity $[\mathrm{W} \cdot \mathrm{m}^{-2}]$
$G(\mathbf{r}_1, \mathbf{r}_2, \tau)$ = Mutual coherence function $[\mathrm{W} \cdot \mathrm{m}^{-2}]$
$G(\nu)$ = Gain of an optical amplifier
$G(\tau)$ = Temporal coherence function $[\mathrm{W} \cdot \mathrm{m}^{-2}]$
$G_{\mathcal{A}}(\tau)$ = Pulse-envelope autocorrelation function
$G_I(\tau)$ = Intensity autocorrelation function $[\mathrm{J}^2 \cdot \mathrm{m}^{-4} \cdot \mathrm{s}^{-1}]$
G_R = Gain of a Raman amplifier
$\mathbf{G}$ = Coherency matrix $[\mathrm{W} \cdot \mathrm{m}^{-2}]$; also, Gyration vector of an optically active medium; also, Wavevector of a phase grating $[\mathrm{m}^{-1}]$
G = Photoionization rate in a photorefractive material
G_0 = Rate of thermal electron–hole generation in a semiconductor $[\mathrm{m}^{-3} \cdot \mathrm{s}^{-1}]$
$\mathbf{G}$ = Reciprocal-lattice vector $[\mathrm{m}^{-1}]$
$\mathbb{G}_n(\cdot)$ = Hermite–Gaussian function of order n
GB = Gain–bandwidth product [Hz]
GFP = Group-IV photonics; also, Green fluorescent protein
GR = Generation–recombination
GRIN = Graded-index
GVD = Group velocity dispersion

h = Complex round-trip amplitude attenuation factor in a resonator; also, Planck's constant $[\mathrm{J} \cdot \mathrm{s}]$
$h(t)$ = Impulse response function of a linear system
$h(x, y)$ = Impulse response function of a two-dimensional linear system
$h_D(t)$ = Photodetector impulse response function
$\hbar$ = $h/2\pi$ $[\mathrm{J} \cdot \mathrm{s}]$
H = Principal point of an optical system; also, Magnetic-field complex amplitude $[\mathrm{A} \cdot \mathrm{m}^{-1}]$
$\mathbf{H}$ = Magnetic-field complex amplitude vector $[\mathrm{A} \cdot \mathrm{m}^{-1}]$
$\mathfrak{H}$ = Magnetic field vector $[\mathrm{A} \cdot \mathrm{m}^{-1}]$
$H(\nu)$ = Transfer function of a linear system [$H(f)$ for low-frequency signals]
$H'(\nu)$ = Real part of the transfer function of a linear system
$H''(\nu)$ = Imaginary part of the transfer function of a linear system
$H(\nu_x, \nu_y)$ = Transfer function of a two-dimensional linear system
$H_e(f)$ = Envelope transfer function of a linear system
H_0 = Transfer-function magnitude
$\mathrm{H}_\ell^{(1)}(\cdot)$ = Hankel function of the first kind of order ℓ
$\mathbb{H}_n(\cdot)$ = Hermite polynomial of order n
H = Horizontal polarization
HAPLS = High-repetition-rate Advanced Petawatt Laser System
HD = High definition
HG = Hermite–Gaussian
HHG = High-harmonic generation
HNLF = Highly nonlinear fiber
HOE = Holographic optical element
HOM = Hong–Ou–Mandel
HOMO = Highest occupied molecular orbital
HVPE = Hydride vapor-phase epitaxy
HXR = Hard-X-ray

i = Electric current [A]; also, Integer; also, $\sqrt{-1}$
i_d = Dark current [A]

i_e = Electron current [A]
i_h = Hole current [A]
i_p = Photoelectric current (photocurrent) [A]
i_s = Reverse current in a semiconductor p–n diode [A]
i_{t} = Threshold current of a laser diode [A]
i_{T} = Transparency current for a laser-diode amplifier [A]
I = Optical intensity (also called Irradiance) [$\mathrm{W \cdot m^{-2}}$]
$I(t)$ = Intensity of an optical pulse [$\mathrm{W \cdot m^{-2}}$]
I_L = Local-oscillator intensity [$\mathrm{W \cdot m^{-2}}$]
I_s = Saturation optical intensity of an amplifier or absorber [$\mathrm{W \cdot m^{-2}}$]; also, Acoustic intensity [$\mathrm{W \cdot m^{-2}}$]; also, Signal intensity [$\mathrm{W \cdot m^{-2}}$]
I_{s} = Optical intensity of a scattered wave [$\mathrm{W \cdot m^{-2}}$]
I_{t} = Threshold intensity of a laser [$\mathrm{W \cdot m^{-2}}$]
I_ν = Spectral intensity [$\mathrm{W \cdot m^{-2} \cdot Hz^{-1}}$]
I_0 = Optical intensity of an incident wave [$\mathrm{W \cdot m^{-2}}$]
$I_0(\cdot)$ = Modified Bessel function of order zero
$\mathcal{I}$ = Fourier transform of intensity profile; also, Moment of inertia [$\mathrm{kg \cdot m^2}$]
$\mathbf{I}$ = Identity matrix
I = In-phase component of the field
IC = Integrated circuit
ICL = Interband cascade laser
IEEE = Institute of Electrical and Electronics Engineers
IF = Intermediate frequency
IG = Ince–Gaussian
IM = Intensity modulation
IP = Internet protocol
IR = Infrared
IRE = Institute of Radio Engineers
ISI = Intersymbol interference
ITO = Indium tin oxide
ITU = International Telecommunications Union

j = $\sqrt{-1}$; also, Integer
J = Electric current density [$\mathrm{A \cdot m^{-2}}$]
J_e = Electron current density [$\mathrm{A \cdot m^{-2}}$]
J_h = Hole current density [$\mathrm{A \cdot m^{-2}}$]
$J_\ell(\cdot)$ = Bessel function of the first kind of order ℓ
J_p = Photoelectric current density [$\mathrm{A \cdot m^{-2}}$]
J_{t} = Threshold current density of a laser diode [$\mathrm{A \cdot m^{-2}}$]
J_{T} = Transparency current density of a laser-diode amplifier [$\mathrm{A \cdot m^{-2}}$]
$\mathbf{J}$ = Jones vector
$\mathcal{J}$ = Total angular-momentum quantum number
$\boldsymbol{\mathcal{J}}$ = Electric current density vector [$\mathrm{A \cdot m^{-2}}$]

k = Wavenumber [$\mathrm{m^{-1}}$]; also, Integer; also, Spatial angular frequency [$\mathrm{rad \cdot m^{-1}}$]
k_{e} = Complex effective wavenumber of a host medium with embedded scatterers [$\mathrm{m^{-1}}$]
k_o = Free-space wavenumber [$\mathrm{m^{-1}}$]
k_r = Reflected wavenumber [$\mathrm{m^{-1}}$]; also, Upshifted wavenumber of a Bragg-reflected wave [$\mathrm{m^{-1}}$]
k_s = Downshifted wavenumber of a Bragg-reflected wave [$\mathrm{m^{-1}}$]

k_s = Wavenumber inside a scattering medium $[\mathrm{m}^{-1}]$
$k_T = \sqrt{k_x^2 + k_y^2}$ = Transverse component of the wavevector $[\mathrm{m}^{-1}]$
k_x, k_y = Wavevector components in x and y directions $[\mathrm{m}^{-1}]$; also, Spatial angular frequencies in x and y directions $[\mathrm{rad} \cdot \mathrm{m}^{-1}]$
k_0 = Central wavenumber $[\mathrm{m}^{-1}]$
$\mathbf{k}$ = Wavevector $[\mathrm{m}^{-1}]$
$\mathbf{k}_g$ = Grating wavevector $[\mathrm{m}^{-1}]$
$\mathbf{k}_r$ = Reflected wavevector $[\mathrm{m}^{-1}]$; also, Upshifted wavevector of a Bragg-reflected wave $[\mathrm{m}^{-1}]$
$\mathbf{k}_s$ = Downshifted wavevector of a Bragg-reflected wave $[\mathrm{m}^{-1}]$
k = Ionization ratio for an avalanche photodiode
k = Boltzmann's constant $[\mathrm{J} \cdot \mathrm{K}^{-1}]$
K = Undulator (magnetic-deflection) parameter in a free-electron laser
$\mathrm{K}_m(\cdot)$ = Modified Bessel function of the second kind of order m
$\mathrm{K}\alpha$ = Designation of X-ray line arising from transition from $\mathrm{n} = 2$ to $\mathrm{n} = 1$ atomic shells
$\mathcal{K}$ = Bloch wavenumber $[\mathrm{m}^{-1}]$
$\mathbf{K}$ = Bloch wavevector $[\mathrm{m}^{-1}]$
KDP = Potassium dihydrogen phosphate
KGW = Potassium gadolinium tungstate
KTP = Potassium titanyl phosphate
KYW = Potassium yttrium tungstate

l = Length $[\mathrm{m}]$; also, Integer
l_c = Coherence length $[\mathrm{m}]$
ℓ = Azimuthal quantum number
ℓ_0 = Optical pathlength of the central frequency component of a pulse
L = Length $[\mathrm{m}]$; also, Distance $[\mathrm{m}]$; also, Electrical inductance $[\mathrm{H}]$; also, Loss factor; also, Number of incoming optical beams to an acousto-optic switch; also, Integer
L_c = Coherence length in a parametric interaction $[\mathrm{m}]$
L_p = Soliton–soliton interaction period $[\mathrm{m}]$
L_v = Luminance $[\mathrm{cd} \cdot \mathrm{m}^{-2}]$
$L_0 = \pi/2\mathcal{C}$ = Coupling length (transfer distance) in a directional coupler $[\mathrm{m}]$
$\mathcal{L}$ = Linear operator
L = Orbital angular momentum quantum number
$^{2\mathcal{S}+1}\mathrm{L}_\mathcal{J}$ = Term symbol for angular-momentum quantum numbers with LS coupling
$\mathcal{L}$ = Angular momentum $[\mathrm{J} \cdot \mathrm{s}]$
$\mathbb{L}_m(\cdot)$ = Laguerre polynomial of degree m
$\mathbb{L}_m^l(\cdot)$ = Generalized Laguerre polynomial of degree m, order l, and index $(m,\, l)$
L-band = Long optical fiber telecommunications band (1565–1625 nm)
LB_0 = Bit-rate–distance product $[\mathrm{km} \cdot \mathrm{Gb} \cdot \mathrm{s}^{-1}]$
LAN = Local-area network
LANL = Los Alamos National Laboratory
LASER = Light amplification by stimulated emission of radiation
LaWS = Laser Weapon System (U.S. Navy)
LBO = Lithium triborate
LC = Liquid crystal
LCD = Liquid-crystal display
LCLS = Linac Coherent Light Source at SLAC National Accelerator Laboratory operated by Stanford University
LCP = Left-circularly polarized
LD = Laser diode

LED = Light-emitting diode
LEP = Light-emitting polymer material
LG = Laguerre–Gaussian
LHS = Left-hand side
LIGO = Laser Interferometer Gravitational-wave Observatory
LINAC = Linear accelerator
LLNL = Lawrence Livermore National Laboratory
LMA = Large mode-area
LO = Local oscillator
LP = Linearly polarized
LPE = Liquid-phase epitaxy
LSP = Localized surface plasmon (localized surface plasmon polariton)
LuAG = Lutetium aluminum garnet
LUMO = Lowest unoccupied molecular orbital
LWFA = Laser wakefield acceleration
LWI = Lasing without inversion
LWIR = Long-wavelength infrared, stretching from 8 to 14 μm

m = Mass of a particle [kg]; also, Free electron mass [kg]; also, Integer; also, Contrast or modulation depth
m_c = Effective mass of a conduction-band electron [kg]
m_p = Proton mass [kg]
m_r = Reduced mass of an electron–hole pair in a semiconductor [kg]
m_v = Effective mass of a valence-band hole [kg]
m_0 = Free electron mass [kg]
$\mathfrak{m}$ = Magnetic dipole moment $[\mathrm{A}\cdot\mathrm{m}^2]$
m = Photon number; also, Photoelectron number
m_0 = Photoelectron-number sensitivity of an optical receiver
m = Magnetic quantum number
M = Magnification in an image system; also, Number of modes; also, Magnetization density complex amplitude $[\mathrm{A}\cdot\mathrm{m}^{-1}]$; also, Number of harmonics; also, Integer
M_v = Illuminance [lx]
$\mathbf{M}$ = Magnetization density complex amplitude vector $[\mathrm{A}\cdot\mathrm{m}^{-1}]$
$\mathcal{M}$ = Figure of merit indicating strength of acousto-optic effect in a material $[\mathrm{m}^2\cdot\mathrm{W}^{-1}]$
$\boldsymbol{\mathfrak{M}}$ = Magnetization density vector $[\mathrm{A}\cdot\mathrm{m}^{-1}]$
M = Mass of an atom or molecule [kg]
$\mathsf{M}(\nu)$ = Density of modes in a resonator [$\mathrm{m}^{-3}\cdot\mathrm{Hz}^{-1}$ for 3D resonator; $\mathrm{m}^{-1}\cdot\mathrm{Hz}^{-1}$ for 1D resonator]
M_r = Reduced mass of an atom or molecule [kg]
$\mathbf{M}$ = Ray-transfer matrix; also, Wave-transfer matrix
$\mathbb{M}^2$ = Factor representing deviation of optical-beam profile from Gaussian form
MAN = Metropolitan-area network
MBE = Molecular-beam epitaxy
MC = Mode converter
MCF = Multicore fiber
MCP = Microchannel plate
MEMS = Microelectromechanical system
MI = Michelson interferometer
MIM = Metal–insulator–metal
MIMO = Multiple-input multiple-output
MIR = Mid infrared, stretching from 2 to 20 μm

MIRACL = Mid-infrared advanced chemical laser
MIS = Metal–insulator–semiconductor
MKS = Meter/kilogram/second unit system
MMF = Multimode fiber
MOCVD = Metalorganic chemical vapor deposition (same as MOVPE)
MOFA = Master-oscillator fiber-amplifier
MOPA = Master-oscillator power-amplifier
MOSFET = Metal-oxide-semiconductor field-effect transistor
MOT = Magneto-optical trap
MOVPE = Metalorganic vapor phase epitaxy (same as MOCVD)
MPM = Multiphoton microscopy
MQD = Multiquantum dot
MQW = Multiquantum well
M-ROADM = Multi-degree reconfigurable optical add–drop multiplexer
MUV = Mid ultraviolet, stretching from 200 to 300 nm
MUX = Multiplexer
MWIR = Medium-wavelength infrared, stretching from 3 to 5 μm
MZI = Mach–Zehnder interferometer
MZM = Mach–Zehnder modulator

n = Refractive index; also, Integer
$n(\mathbf{r})$ = Refractive index of an inhomogeneous medium
$n(\theta)$ = Refractive index of extraordinary wave in a uniaxial crystal
n_{b} = Effective refractive index associated with SPP at metal–dielectric boundary
n_e = Extraordinary refractive index
n_o = Ordinary refractive index
n_{p} = Refractive index of a prism
n_{s} = Refractive index of a scattering volume
n_2 = Optical Kerr coefficient (nonlinear refractive index) $[\mathrm{m}^2 \cdot \mathrm{W}^{-1}]$
$\mathfrak{n}$ = Photon-number density $[\mathrm{m}^{-3}]$
$\mathfrak{n}_s$ = Saturation photon-number density $[\mathrm{m}^{-3}]$
n = Photon number
$\overline{n}$ = Mean photon number
$\overline{n}_\nu$ = Spectral photon number $[\mathrm{Hz}^{-1}]$
n_0 = Number-state photon number; also, Photon-number sensitivity of an optical receiver
n = Principal quantum number
n = Concentration of electrons in a semiconductor $[\mathrm{m}^{-3}]$
$\mathrm{n_i}$ = Concentration of electrons/holes in an intrinsic semiconductor $[\mathrm{m}^{-3}]$
n_0 = Equilibrium concentration of electrons in a semiconductor $[\mathrm{m}^{-3}]$
N = Group index; also, Integer; also, Number of atoms; also, Number of stages; also, Number of optical cycles; also, Number of resolvable spots of a scanner; also, Order of a higher-order soliton
N_{F} = Fresnel number
N = Number density $[\mathrm{m}^{-3}]$; also, $N = N_2 - N_1 =$ Population density difference $[\mathrm{m}^{-3}]$
N_a = Atomic number density $[\mathrm{m}^{-3}]$
N_A = Number density of ionized acceptor atoms in a semiconductor $[\mathrm{m}^{-3}]$
N_D = Number density of ionized donor atoms in a semiconductor $[\mathrm{m}^{-3}]$; also, Number density of donor atoms in a photorefractive material $[\mathrm{m}^{-3}]$
N_D^+ = Number density of ionized donor atoms in a photorefractive material $[\mathrm{m}^{-3}]$
N_{s} = Number density of scatterers $[\mathrm{m}^{-3}]$

N_t = Laser threshold population difference $[\mathrm{m}^{-3}]$
N_0 = Steady-state population difference in the absence of amplifier radiation $[\mathrm{m}^{-3}]$
NA = Numerical aperture
NEA = Negative-electron-affinity
NEP = Noise-equivalent power
NIF = National Ignition Facility
NIM = Negative-index material
NIR = Near infrared, stretching from 0.760 to 2 μm
NL = Nonlinear
NLDC = Nonlinear directional coupler
NOLM = Nonlinear optical loop mirror
NRI = Negative refractive index
NRZ = Non-return-to-zero
NUV = Near ultraviolet, stretching from 300 to 390 nm
NZ-DSF = Non-zero dispersion shifted fiber

O-band = Original optical fiber telecommunications band (1260–1360 nm)
OA = Optical amplifier
OADM = Optical add–drop multiplexer
OAM = Orbital angular momentum
OC = Optical carrier
OCT = Optical coherence tomography
ODMUX = Optical demultiplexer (also abbreviated as ODEMUX)
O/E = Optical-to-electronic
o-e-e = Ordinary-extraordinary-extraordinary designations for waves 1, 2, and 3
OEIC = Optoelectronic integrated circuit
o-e-o = Ordinary-extraordinary-ordinary designations for waves 1, 2, and 3
OEO = Optical-electrical-optical
OFA = Optical fiber amplifier
OFC = Optical frequency comb; also, Optical frequency conversion
OH = Hydroxyl radical
OLED = Organic light-emitting diode
OM = Optical mode
OMUX = Optical multiplexer
o-o-e = Ordinary-ordinary-extraordinary designations for waves 1, 2, and 3
OOK = ON–OFF keying
OPA = Optical parametric amplifier
OPC = Optical phase conjugation
OPCPA = Optical parametric chirped-pulse amplification
OPD = Organic photodetector
OPO = Optical parametric oscillator
OR = OR logic gate
OSA = Optical Society of America
OTDM = Optical time-division multiplexer
OXC = Optical cross-connect

p = Probability; also, Momentum $[\mathrm{kg} \cdot \mathrm{m} \cdot \mathrm{s}^{-1}]$; also, Graded-index fiber profile parameter
$p(n)$ = Probability of n events
$p(n)$ = Photon-number distribution
$p(x, y)$ = Aperture function or pupil function

p_{ab} = Probability density for absorption (mode containing one photon) $[\mathrm{s}^{-1}]$
p_{sp} = Probability density for spontaneous emission (into one mode) $[\mathrm{s}^{-1}]$
p_{st} = Probability density for stimulated emission (mode containing one photon) $[\mathrm{s}^{-1}]$
p = Electric dipole moment $[\mathrm{C} \cdot \mathrm{m}]$
p = Normalized electric-field quadrature component
p = Elasto-optic (strain-optic) coefficient
p_{ijkl} = Component of the photoelasticity tensor
p_{IK} = Component of the photoelasticity tensor (contracted indices)
p = Concentration of holes in a semiconductor $[\mathrm{m}^{-3}]$
p_0 = Equilibrium concentration of holes in a semiconductor $[\mathrm{m}^{-3}]$
P = Electric polarization-density complex amplitude $[\mathrm{C} \cdot \mathrm{m}^{-2}]$; also, Probability of impact ionization
$P(\nu_x, \nu_y)$ = Fourier transform of the aperture function $p(x, y)$
P_{ab} = Probability density for absorption (mode containing many photons) $[\mathrm{s}^{-1}]$
$P_{\mathrm{m}}^{\ell}(\cdot)$ = Microsphere-resonator adjoint Legendre function
P_{NL} = Complex amplitude of the nonlinear component of the polarization density $[\mathrm{C} \cdot \mathrm{m}^{-2}]$
P_{sp} = Probability density for spontaneous emission (into any mode) $[\mathrm{s}^{-1}]$
P_{st} = Probability density for stimulated emission (mode containing many photons) $[\mathrm{s}^{-1}]$
$\mathbf{P}$ = Electric polarization-density complex amplitude vector $[\mathrm{C} \cdot \mathrm{m}^{-2}]$
$\boldsymbol{\mathcal{P}}$ = Electric polarization density vector $[\mathrm{C} \cdot \mathrm{m}^{-2}]$
$\mathcal{P}_{\mathrm{L}}$ = Linear component of the polarization density $[\mathrm{C} \cdot \mathrm{m}^{-2}]$
$\mathcal{P}_{\mathrm{NL}}$ = Nonlinear component of the polarization density $[\mathrm{C} \cdot \mathrm{m}^{-2}]$
P = Optical power (also called Radiant power or Radiant flux) [W]
P_{e} = Electrical power [W]
P_i = Incident optical power [W]
P_L = Local-oscillator power [W]
P_o = Output optical power [W]; also, Average output optical power [W]
P_p = Peak pulse power [W]
P_{p} = Optical pump power [W]
P_r = Received optical power [W]
P_s = Signal (optical) power [W]
P_{s} = Optical power of a scattered wave [W]
P_{t} = Threshold optical pump power [W]
P_{v} = Luminous flux [lm]
P_ν = Power spectral density $[\mathrm{W} \cdot \mathrm{Hz}^{-1}]$
P_π = Half-wave optical power in a Kerr medium [W]
$\mathbb{P}$ = Degree of polarization
P = Optical power [dBm]
P_c = Optical power loss associated with splicing and coupling [dBm]
P_m = Optical power allotted for safety margin [dBm]
P_r = Receiver optical power sensitivity [dBm]
P_s = Source optical power [dBm]
PAL-SLM = Parallel aligned spatial light modulator
PBG = Photonic bandgap
PBS = Polarizing beamsplitter
PCB = Printed circuit board
PCF = Photonic-crystal fiber
PC-LED = Phosphor-conversion LED
PCM = Pulse code modulation
PD = Photodetector

PDE = Photon detection efficiency
PET = Positron-emission tomography
PG-FROG = Polarization-gated frequency-resolved optical gating
PIC = Photonic integrated circuit
PIN = p-type–i-type–n-type photodiode
PLASER = Powder laser
PLC = Planar lightwave circuit
PLED = Polymer organic light-emitting diode
PM = Phase modulation
PMD = Polarization mode dispersion
PMT = Photomultiplier tube
P-OLED = Polymer light-emitting diode
PPLN = Periodically poled lithium niobate
PPV = Poly(p-phenylene vinylene)
PROM = Pockels readout optical modulator
PRR = Pulse repetition rate
PSK = Phase shift keying
PWFA = Plasma wakefield acceleration
PWM = Pulse-width modulation

q = Electric charge $[\mathrm{C}]$; also, Wavenumber of an acoustic wave $[\mathrm{m}^{-1}]$; also, Integer (mode index, diffraction order, quantum number); also, Spatial angular frequency $[\mathrm{rad} \cdot \mathrm{m}^{-1}]$
$q(z)$ = Complex Gaussian-beam parameter $[\mathrm{m}]$
q = Quantum defect
$\mathbf{q}$ = Wavevector of an acoustic wave $[\mathrm{m}^{-1}]$
Q = Electric charge $[\mathrm{C}]$; also, Quality factor of an optical resonator or a resonant circuit
Q_{a} = Absorption efficiency
Q_{s} = Scattering efficiency
Q = Quadrature component of the field
QAM = Quadrature amplitude modulation
QCL = Quantum cascade laser
QCSE = Quantum-confined Stark effect
QD = Quantum dot
QDIP = Quantum-dot infrared photodetector
QED = Quantum electrodynamics
QOCT = Quantum optical coherence tomography
QPM = Quasi-phase matching; also, Quadratic phase modulator
QPSK = Quaternary phase shift keying
QWIP = Quantum-well infrared photodetector

r = Radial distance in spherical and cylindrical coordinates $[\mathrm{m}]$
r_{n} = Radii of allowed electron orbits in Bohr atom $[\mathrm{m}]$
$\mathbf{r}$ = Position vector $[\mathrm{m}]$
$\widehat{\mathbf{r}}$ = Unit vector in radial direction in spherical coordinates
r = Complex amplitude reflectance; also, Complex round-trip amplitude attenuation factor in a resonator
$|\mathrm{r}|$ = Magnitude of round-trip amplitude attenuation factor in a resonator
r_{+} = Frequency-upshifted Bragg amplitude reflectance
r_{-} = Frequency-downshifted Bragg amplitude reflectance
$r(\nu)$ = Rate of photon emission/absorption from a semiconductor $[\mathrm{s}^{-1} \cdot \mathrm{m}^{-3} \cdot \mathrm{Hz}^{-1}]$

r = Linear electro-optic (Pockels) coefficient $[\mathrm{m}\cdot\mathrm{V}^{-1}]$; also, Rotational quantum number
r_{ijk} = Component of the linear electro-optic (Pockels) tensor $[\mathrm{m}\cdot\mathrm{V}^{-1}]$
r_{Ik} = Component of the linear electro-optic (Pockels) tensor (contracted indices) $[\mathrm{m}\cdot\mathrm{V}^{-1}]$
r = Electron–hole recombination coefficient $[\mathrm{m}^3\cdot\mathrm{s}^{-1}]$
r_{nr} = Nonradiative electron–hole recombination coefficient $[\mathrm{m}^3\cdot\mathrm{s}^{-1}]$
r_{r} = Radiative electron–hole recombination coefficient $[\mathrm{m}^3\cdot\mathrm{s}^{-1}]$
$\mathrm{rect}(\cdot)$ = Pulse of unit height and unit width centered about 0
R = Radius of curvature $[\mathrm{m}]$; also, Electrical resistance $[\Omega]$
$R(z)$ = Radius of curvature of a Gaussian beam $[\mathrm{m}]$
$R(\tau)$ = Field autocorrelation function
R_L = Load resistance $[\Omega]$
R_m = Distance from the focal point to the mth ring of a Fresnel zone plate $[\mathrm{m}]$
R_0 = Radius of cylinder in which a meridional ray is confined $[\mathrm{m}]$
$\mathbf{R}(\theta)$ = Jones matrix for coordinate rotation by an angle θ
$\mathcal{R}$ = Intensity or power reflectance; also, approximate reflectance of a Bragg reflector
$\mathcal{R}_e$ = Exact intensity or power reflectance of a Bragg reflector
R = Pumping rate $[\mathrm{s}^{-1}\cdot\mathrm{m}^{-3}]$; also, Recombination rate in a semiconductor $[\mathrm{s}^{-1}\cdot\mathrm{m}^{-3}]$; also, Electron–hole injection rate in a semiconductor $[\mathrm{s}^{-1}\cdot\mathrm{m}^{-3}]$
R_{t} = Laser threshold pumping rate $[\mathrm{s}^{-1}\cdot\mathrm{m}^{-3}]$
$\mathbf{R}$ = Lattice vector $[\mathrm{m}]$
R = Responsivity of a photon source $[\mathrm{W}\cdot\mathrm{A}^{-1}]$; also, Responsivity of a photon detector $[\mathrm{A}\cdot\mathrm{W}^{-1}]$
R_d = Differential responsivity of a laser diode $[\mathrm{W}\cdot\mathrm{A}^{-1}]$
$\mathbb{R}_{\mathrm{n}\ell}(r)$ = Hydrogen-atom associated Laguerre function of order ℓ and index n
$\mathbb{R}$ = Ratio of complex-envelope polarization-component magnitudes
RC = Resistor–capacitor combination
RC = Resonant-cavity
RCLED = Resonant-cavity light-emitting diode
RCP = Right-circularly polarized
REFA = Rare-earth-doped fiber amplifier
RF = Radio-frequency
RFA = Raman fiber amplifier
RFID = Radio-frequency identification
RFL = Raman fiber laser
RHS = Right-hand side
RMS = Root-mean square
ROADM = Reconfigurable optical add–drop multiplexer
RW = Ridge waveguide
Rx = Receiver
RZ = Return-to-zero

s = Length or distance $[\mathrm{m}]$; also, Scale factor
$s(x,t)$ = Strain wavefunction
s_{ij} = Component of the strain tensor
s = Photorefractivity proportionality constant for photoionization cross section
$\mathsf{s}(\mathbf{r}_1,\mathbf{r}_2,\nu)$ = Normalized cross-spectral density
s = Quadratic electro-optic (Kerr) coefficient $[\mathrm{m}^2\cdot\mathrm{V}^{-2}]$; also, Spin quantum number
s_{ijkl} = Component of the quadratic electro-optic (Kerr) tensor $[\mathrm{m}^2\cdot\mathrm{V}^{-2}]$
s_{IK} = Component of the quadratic electro-optic (Kerr) tensor (contracted indices) $[\mathrm{m}^2\cdot\mathrm{V}^{-2}]$
$\mathrm{sinc}(\cdot)$ = Symmetric function with peak value of unity at 0 [$\mathrm{sinc}(t)\equiv\sin(\pi t)/(\pi t)$]
S = Poynting-vector magnitude $[\mathrm{W}\cdot\mathrm{m}^{-2}]$; also, Transition strength (oscillator strength) $[\mathrm{m}^2\cdot\mathrm{Hz}]$

$S(\mathbf{r})$ = Complex amplitude for a radiation source $[\mathrm{V}\cdot\mathrm{m}^{-3}]$
$S(\cdot)$ = Fresnel integral
S_0 = Strain amplitude
$\mathbf{S}$ = Complex Poynting vector $[\mathrm{W}\cdot\mathrm{m}^{-2}]$
$\mathcal{S}(t)$ = Source of optical radiation created by an incident field $[\mathrm{V}\cdot\mathrm{m}^{-3}]$; also, Spin angular-momentum quantum number
$\boldsymbol{\mathcal{S}}$ = Poynting vector $[\mathrm{W}\cdot\mathrm{m}^{-2}]$
$\mathcal{S}(\mathbf{r})$ = Eikonal $[\mathrm{m}]$
$\mathcal{S}(\mathbf{r}_1,\mathbf{r}_2,\nu)$ = Cross-spectral density $[\mathrm{W}\cdot\mathrm{m}^{-2}\cdot\mathrm{Hz}^{-1}]$
$\mathcal{S}(\lambda_o)$ = Wavelength power spectral density $[\mathrm{W}\cdot\mathrm{m}^{-1}]$;
$\mathcal{S}(\nu)$ = Spectral intensity of an optical wave or pulse $[\mathrm{W}\cdot\mathrm{m}^{-2}\cdot\mathrm{Hz}^{-1}]$; also, Power spectral density $[\mathrm{W}\cdot\mathrm{m}^{-2}\cdot\mathrm{Hz}^{-1}]$
$\mathcal{S}(\nu,t)$ = Spectrogram of an optical pulse $[\mathcal{S}(\nu,t)=|\Phi(\nu,t)|^2]$
$\mathbf{S}$ = Scattering matrix
$\mathbb{S}$ = Projection of photon-spin angular momentum along the wavevector (helicity) $[\mathrm{J}\cdot\mathrm{s}]$
$\mathbb{S}_{[\cdot]}$ = Stokes parameters
S-band = Short optical fiber telecommunications band (1460–1530 nm)
SACM = Separate absorption, charge, and multiplication
SAM = Separate absorption and multiplication
SAPD = Superlattice avalanche photodiode; also, Staircase avalanche photodiode
SASE = Self-amplified spontaneous emission
SBN = Strontium barium niobate
SBS = Stimulated Brillouin scattering
SCF = Single-core fiber
SCG = Supercontinuum generation
SCIDCM = Single-carrier-injection double-carrier multiplication
SCISCM = Single-carrier-injection single-carrier multiplication
SDH = Synchronous digital hierarchy
SDL = Semiconductor disk laser
SDM = Space-division multiplexing
SEED = Self-electro-optic-effect device
SESAM = Semiconductor saturable-absorber mirror
SFG = Sum-frequency generation
SGDFB = Sampled-grating distributed-feedback
SH = Second harmonic
SHG = Second-harmonic generation
SHG-FROG = Second-harmonic generation frequency-resolved optical gating
SI = International system of units; also, Step-index
SLA = Semiconductor laser amplifier
SLAC = National Accelerator Laboratory at Stanford University
SLED = Superluminescent diode
SLM = Spatial light modulator
SMD = Surface-mounted device
SMF = Single-mode fiber
SMOLED = Small-molecule organic light-emitting diode
SNG = Single-negative medium
SNOM = Scanning near-field optical microscopy
SNR = Signal-to-noise ratio
SNSPD = Superconducting nanowire single-photon detector
SOA = Semiconductor optical amplifier

SOI = Silicon-on-insulator
SONET = Synchronous optical network
SOS = Silica-on-silicon
SPAD = Single-photon avalanche diode
SPASER = Surface-plasmon amplification by stimulated emission of radiation
SPDC = Spontaneous parametric downconversion
SPE = Single-photon emitter
SPIE = The International Society for Optical Engineering
SPM = Self-phase modulation
SPP = Surface plasmon polariton
SPR = Surface plasmon resonance
SQUID = Superconducting quantum-interference device
SQW = Single quantum well
SRO = Singly resonant oscillator
SRS = Stimulated Raman scattering
SSFS = Soliton self-frequency shift
SSPD = Superconducting single-photon detector
STS = Synchronous transport signal
SVE = Slowly varying envelope
SXR = Soft-X-ray

t = Time [s]
t_i = Ionization time [s]
t_r = Recombination time [s]
t_{sp} = Spontaneous lifetime [s]; also, Effective spontaneous lifetime [s]
t = Complex amplitude transmittance; also, Normalized time for an optical pulse
T = Temperature [°K]
$\mathbf{T}$ = Jones matrix
$\mathcal{T}$ = Intensity or power transmittance; also, Power-transfer or power-transmission ratio
T = Transit time [s]; also, Counting time [s]; also, Switching time [s]; also, Bit time interval [s]; also, Resolution time ($T = 1/2B$ where $B = \mathrm{Bandwidth}$) [s]; also, Period of a wave ($T = 1/\nu$ where $\nu = \mathrm{frequency}$) [s]; also, Transit time of sound across an optical beam
$T_F = 1/\nu_F$ = Inverse of Fabry–Perot resonator-mode frequency spacing ($T_F = 2d/c$) [s]; also, Period of a mode-locked laser pulse train [s]
$T_{\ell m}$ = Element of transmission or interconnection matrix
T_0 = Ground-state orbital period in Bohr hydrogen atom ($T_0 \approx 150$ as) [s]
T_2 = Electron–phonon collision time [s]
$\mathbf{T}$ = Transmission matrix; also, Interconnection matrix
TADF = Thermally activated delayed fluorescence
TDM = Time-division multiplexing
TDMA = Time-division multiple access
TE = Transverse electric
TEM = Transverse electromagnetic
TES = Transition-edge sensor
TFF = Thin-film filter
TFT = Thin-film transistor
TGG = Terbium gallium garnet
THG = Third-harmonic generation
TIR = Total internal reflection
TM = Transverse magnetic

TMD = Transition-metal dichalcogenide
TOAD = Terahertz optical asymmetric demultiplexer
TPD = Triphenyl diamine derivative
TPLSM = Two-photon laser scanning fluorescence microscopy
TSI = Time-slot interchange
TST = Time–space–time
TV = Television receiver
Tx = Transmitter
2PM = Two-photon microscopy
3PM = Three-photon microscopy

u = Displacement $[\mathrm{m}]$
$u(\mathbf{r}, t)$ = Wavefunction of an optical wave
$\widehat{\mathbf{u}}$ = Unit vector
u = Number of electrons in a subshell
$U(\mathbf{r})$ = Complex amplitude of a monochromatic optical wave
$U(\mathbf{r}, t)$ = Complex wavefunction of an optical wave
$U(t)$ = Complex wavefunction of an optical pulse
$U(t_1, t_2)$ = Joint temporal wavefunction for two-photon light
$U(x_1, x_2)$ = Joint spatial wavefunction for two-photon light
$U_s(x)$ = Probability amplitude of location of maximally entangled photons
$U_\mathrm{s}(\mathbf{r})$ = Complex amplitude of a scattered monochromatic optical wave
$U_0(\mathbf{r})$ = Complex amplitude of an incident monochromatic optical wave
$\mathcal{U}(x)$ = Unit step function [$\mathcal{U}(x) = 1$ if $x > 0$ and $\mathcal{U}(x) = 0$ if $x < 0$]
U-band = Ultra-long optical fiber telecommunications band (1625–1675 nm)
ULSI = Ultra-large-scale integration
UV = Ultraviolet
UVA = Ultraviolet-A band, stretching from 315 to 400 nm
UVB = Ultraviolet-B band, stretching from 280 to 315 nm
UVC = Ultraviolet-C band, stretching from 100 to 280 nm

v = Group velocity of a wave $[\mathrm{m} \cdot \mathrm{s}^{-1}]$
$v(\mathbf{r}, \nu)$ = Fourier transform of the wavefunction of an optical wave
v_s = Velocity of sound $[\mathrm{m} \cdot \mathrm{s}^{-1}]$
v = Velocity of an atom or object $[\mathrm{m} \cdot \mathrm{s}^{-1}]$; also, Drift velocity of a carrier $[\mathrm{m} \cdot \mathrm{s}^{-1}]$
v_e = Velocity of an electron $[\mathrm{m} \cdot \mathrm{s}^{-1}]$
v_h = Velocity of a hole $[\mathrm{m} \cdot \mathrm{s}^{-1}]$
$\mathbf{v}$ = Velocity vector of a charge carrier $[\mathrm{m} \cdot \mathrm{s}^{-1}]$
v = Vibrational quantum number
V = Vertex; also, Volume $[\mathrm{m}^3]$; also, Modal volume $[\mathrm{m}^3]$; also, Voltage [V]
$V(\mathbf{r}, \nu)$ = Fourier transform of the complex wavefunction of an optical wave
$V(\lambda_o)$ = Photopic luminosity function (photopic luminous efficiency function)
$V(\nu)$ = Fourier transform of the complex wavefunction of an optical pulse
$V(\nu_1, \nu_2)$ = Joint spectral wavefunction for two-photon light
$V(\nu_{x1}, \nu_{x2})$ = Joint wavevector wavefunction for two-photon light [$k_{x1} = 2\pi\nu_{x1}$ and $k_{x2} = 2\pi\nu_{x2}$]
V_B = Battery voltage [V]
V_c = Critical voltage for a liquid-crystal cell [V]
$V_s(\nu_x)$ = Fourier transform of $U_s(x)$
V_π = Half-wave voltage of an electro-optic retarder or modulator [V]

V_0 = Built-in potential difference in a p–n junction [V]; also, Switching voltage of a directional coupler [V]
$\mathcal{V}$ = Visibility
V = Verdet constant $[\mathrm{rad} \cdot \mathrm{m}^{-1} \cdot \mathrm{T}^{-1}]$
V = Fiber V parameter
$V(\mathbf{r})$ = Potential energy [J]
$\mathbb{V}$ = Abbe number of a dispersive medium
V = Vertical polarization
VCSEL = Vertical-cavity surface-emitting laser
VECSEL = Vertical external-cavity surface-emitting laser
VLSI = Very-large-scale integration
VOx = Vanadium oxide
VPE = Vapor-phase epitaxy
VUV = Vacuum ultraviolet, stretching from 10 to 200 nm

w = Width [m]; also, Radius of a thin metallic rod [m]; also, Length of an acousto-optic cell [m]
$\mathcal{W}$ = Integrated photon flux (integrated optical power in units of photon number)
w_d = Width of the absorption region in an avalanche photodiode [m]
w_m = Width of the multiplication region in an avalanche photodiode [m]
W = Time-averaged electromagnetic energy density $[\mathrm{J} \cdot \mathrm{m}^{-3}]$
$W(t)$ = Window function; also, Window function for short-time Fourier transform
$W(z)$ = Width (radius) of a Gaussian beam at an axial distance z from the beam center [m]
W_0 = Waist radius of a Gaussian beam [m]
$\mathcal{W}$ = Electromagnetic energy density $[\mathrm{J} \cdot \mathrm{m}^{-3}]$
W = Probability density for absorption of pump light $[\mathrm{s}^{-1}]$
W_i = Probability density for absorption and stimulated emission $[\mathrm{s}^{-1}]$
$\mathbb{W}$ = Ionization energy of an atom [J]; also, Photoelectric work function [J]
WAN = Wide-area network
WC = Wavelength converter
WCI = Wavelength-channel interchange
WDM = Wavelength-division multiplexing
WDMA = Wavelength-division multiple access
WG = Waveguide
WGM = Whispering-gallery mode
WGR = Waveguide grating router
WKB = Wentzel–Kramers–Brillouin
WLAN = Wireless local-area network
WOLED = White organic light-emitting diode
WPE = Wall-plug efficiency
WSS = Wavelength-selective switch

x = Position coordinate [m]; also, Displacement [m]
$\widehat{\mathbf{x}}$ = Unit vector in the x direction in Cartesian coordinates
$\mathrm{x}(t)$ = Inverse Fourier transform of the susceptibility of a dispersive medium $\chi(\nu)$
x = Normalized electric-field quadrature component
X = Normalized photon-flux density at the input to an optical amplifier
$\mathbf{X}$ = Input vector to a linear system
$\mathbf{X}_q$ = Eigenvectors associated with an eigenvalue problem
$\mathcal{X}(u)$ = Real function associated with the Hermite–Gaussian beam
$\mathcal{X}^{(2)}(\omega_1, \omega_2)$ = Second-order nonlinear susceptibility

$X(\cdot)$ = Normalized rate of change of the radial distribution in the core of a step-index fiber
XC = Cross-connect
XGM = Cross-gain modulation
XOR = Exclusive OR gate
XPM = Cross-phase modulation
XUV = Extreme ultraviolet

y = Position coordinate $[\mathrm{m}]$
$\hat{\mathbf{y}}$ = Unit vector in the y direction in Cartesian coordinates
Y = Normalized photon-flux density at the output of an optical amplifier
$\mathbf{Y}$ = Output vector from a linear system
$\mathcal{Y}(v)$ = Real function associated with the Hermite–Gaussian beam
$Y(\cdot)$ = Normalized rate of change of the radial distribution in the cladding of a step-index fiber
YAG = Yttrium aluminum garnet
YIG = Yttrium iron garnet
YLF = Yttrium lithium fluoride

z = Position coordinate (Cartesian or cylindrical coordinates) $[\mathrm{m}]$
$z_{\min}$ = Location of the minimum width for a chirped Gaussian pulse $[\mathrm{m}]$
z_{NL} = Nonlinear characteristic length of a Kerr medium $[\mathrm{m}]$
z_p = Soliton period $[\mathrm{m}]$
z_{T} = Talbot distance $[\mathrm{m}]$
z_0 = Rayleigh range of a Gaussian beam $[\mathrm{m}]$
$|z_0|$ = Dispersion length of a Gaussian pulse traveling through a dispersive medium $[\mathrm{m}]$
$\hat{\mathbf{z}}$ = Unit vector in the z direction in Cartesian coordinates
z = Normalized distance for an optical pulse
Z = Atomic number; also, Electrical-circuit impedance $[\Omega]$
$\mathcal{Z}(\cdot)$ = Real function associated with the Hermite–Gaussian beam

Greek Symbols

α = Apex angle of a prism; also, Twist coefficient of a twisted nematic liquid crystal $[\mathrm{m}^{-1}]$; also, Attenuation or absorption coefficient $[\mathrm{m}^{-1}]$; also, Intensity extinction coefficient: $\alpha = \alpha_{\mathrm{a}} + \alpha_{\mathrm{s}}$ $[\mathrm{m}^{-1}]$; also, Linewidth-enhancement factor for a laser diode
α_{a} = Absorption coefficient of a scattering material $[\mathrm{m}^{-1}]$
α_e = Electron ionization coefficient in a semiconductor $[\mathrm{m}^{-1}]$
α_h = Hole ionization coefficient in a semiconductor $[\mathrm{m}^{-1}]$
α_m = Loss coefficient of a resonator attributed to a mirror $[\mathrm{m}^{-1}]$
α_p = Mean value of p for a coherent state
α_r = Effective overall distributed loss coefficient $[\mathrm{m}^{-1}]$
α_s = Loss coefficient of a laser medium $[\mathrm{m}^{-1}]$
α_{s} = Scattering coefficient $[\mathrm{m}^{-1}]$
α_x = Mean value of x for a coherent state
α_ν = Angular dispersion coefficient $[\mathrm{Hz}^{-1}]$
α = Attenuation coefficient of an optical fiber $[\mathrm{dB/km}]$

$\beta = k_z$ = Propagation constant $[\mathrm{m}^{-1}]$; also, Phase-retardation coefficient of a twisted nematic liquid crystal $[\mathrm{m}^{-1}]$
β' = First derivative of β with respect to ω $[\mathrm{m}^{-1} \cdot \mathrm{s}]$
β'' = Second derivative of β with respect to ω $[\mathrm{m}^{-1} \cdot \mathrm{s}^2]$
$\beta(\nu)$ = Propagation constant in a dispersive medium $[\mathrm{m}^{-1}]$

$\beta_0 = \beta(\nu_0) =$ Propagation constant at the central frequency ν_0 $[\mathrm{m}^{-1}]$

$\beta =$ Spontaneous-emission coupling coefficient

$\gamma =$ Field attenuation coefficient $[\mathrm{m}^{-1}]$; also, Field extinction coefficient $[\mathrm{m}^{-1}]$; also, Lateral decay coefficient in a waveguide $[\mathrm{m}^{-1}]$; also, Lorentz factor in special relativity ($\gamma = [1 - (v/c)^2]^{-1/2}$); also, Coupling coefficient in a parametric device $[\mathrm{m}^{-1}]$; also, Nonlinear coefficient in soliton theory

$\gamma_\mathrm{b} =$ Amplitude attenuation coefficient of a traveling surface plasmon polariton $[\mathrm{m}^{-1}]$

$\gamma_B =$ Magnetogyration coefficient $[\mathrm{m}^2 \cdot \mathrm{Wb}^{-1}]$

$\gamma_m =$ Field extinction coefficient of an evanescent wave $[\mathrm{m}^{-1}]$

$\gamma_R =$ Photorefractivity recombination-rate coefficient

$\gamma =$ Gain coefficient of an optical amplifier $[\mathrm{m}^{-1}]$

$\gamma(\nu) =$ Gain coefficient of an optical amplifier $[\mathrm{m}^{-1}]$

$\gamma_m =$ Maximum gain coefficient of a quantum-well laser-diode amplifier $[\mathrm{m}^{-1}]$

$\gamma_p =$ Peak gain coefficient of a laser-diode amplifier $[\mathrm{m}^{-1}]$

$\gamma_R =$ Raman gain coefficient $[\mathrm{m}^{-1}]$

$\gamma_\beta(\nu) =$ Gain coefficient of a subset of atoms $[\mathrm{m}^{-1}]$

$\gamma_0(\nu) =$ Small-signal gain coefficient of an optical amplifier $[\mathrm{m}^{-1}]$

$\Gamma =$ Phase retardation; also, Power confinement factor in a laser diode or waveguide photodiode; also, Crystallographic symbol for irreducible Brillouin zone

$\Gamma_m =$ Power confinement factor in a waveguide

$\delta =$ Secondary-emission gain random variable

$\delta(\cdot) =$ Delta function or impulse function

$\delta x =$ Increment of x

$\delta\theta =$ Angular divergence of an optical beam

$\delta\theta_s =$ Angular divergence of an acoustic beam

$\delta\nu =$ Spectral width of a resonator mode [Hz]

$\Delta =$ Thickness of a thin optical component [m]; also, Fractional refractive-index change in an optical fiber or waveguide ($\Delta \approx (n_1 - n_2)/n_1$)

$\Delta\mathrm{n} =$ Concentration of excess electron–hole pairs $[\mathrm{m}^{-3}]$

$\Delta\mathrm{n_T} =$ Concentration of injected carriers for a semiconductor optical amplifier at transparency $[\mathrm{m}^{-3}]$

$\Delta x =$ Increment of x

$\Delta\nu =$ Spectral width or linewidth [Hz]; also, Atomic linewidth or transition linewidth [Hz]

$\Delta\nu_c = 1/\tau_c =$ Spectral width [Hz]

$\Delta\nu_\mathrm{D} =$ Doppler linewidth [Hz]

$\Delta\nu_\mathrm{FWHM} =$ Full-width-at-half-maximum spectral width [Hz]

$\Delta\nu_\mathrm{L} =$ Laser linewidth [Hz]

$\Delta\nu_s =$ Linewidth of a saturated amplifier [Hz]

$\Delta\nu_\mathrm{ST} =$ Schawlow–Townes minimum laser linewidth [Hz]

$\epsilon =$ Electric permittivity of a medium $[\mathrm{F} \cdot \mathrm{m}^{-1}]$; also, Focusing error $[\mathrm{m}^{-1}]$

$\epsilon' =$ Real part of the electric permittivity of a medium $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

$\epsilon'' =$ Imaginary part of the electric permittivity of a medium $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

$\epsilon_\mathrm{b} =$ Effective permittivity associated with SPP at metal–dielectric boundary $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

$\epsilon_\mathrm{c} =$ Effective permittivity of a conductive medium $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

$\epsilon_\mathrm{d} =$ Electric permittivity of the dielectric medium in a layered metamaterial $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

$\epsilon_\mathrm{e} =$ Effective permittivity of a host medium with embedded objects $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

$\epsilon_{ij} =$ Component of the electric permittivity tensor $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

$\epsilon_\mathrm{m} =$ Electric permittivity of the metallic medium in a layered metamaterial $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

ϵ_o = Electric permittivity of free space $[\mathrm{F} \cdot \mathrm{m}^{-1}]$
ϵ_r = Relative permittivity or dielectric constant
ϵ_s = Electric permittivity of a scattering volume $[\mathrm{F} \cdot \mathrm{m}^{-1}]$
$\boldsymbol{\epsilon}$ = Electric permittivity tensor $[\mathrm{F} \cdot \mathrm{m}^{-1}]$

$\zeta(z)$ = Excess axial phase of a Gaussian beam; also, Fraction of electron–hole pairs that successfully contribute to detector photocurrent
ζ = Damping coefficient of a harmonic oscillator $[\mathrm{s}^{-1}]$; also, Scattering rate or collision frequency ($\zeta = 1/\tau$ where τ = scattering time or collision time) $[\mathrm{s}^{-1}]$

η = Impedance of a dielectric medium $[\Omega]$; also, Photodetector quantum efficiency; also, Photon detection efficiency
η_c = Power-conversion efficiency (also called Overall efficiency and Wall-plug efficiency)
η_d = External differential quantum efficiency
η_e = Extraction efficiency; also, Transmission efficiency
η_ex = External efficiency
η_i = Internal quantum efficiency
η_o = Impedance of free space $[\Omega]$
η_o = Optical-to-optical efficiency
η_OFC = Optical frequency conversion efficiency
η_s = Differential power-conversion efficiency (also called Slope efficiency)
η_s = Optical-to-optical slope efficiency
η_SHG = Second-harmonic generation efficiency
$\upeta$ = Electric impermeability
$\upeta_{ij}$ = Component of the electric impermeability tensor
$\boldsymbol{\upeta}$ = Electric impermeability tensor

θ = Angle; also, Twist angle in a liquid crystal; also, Deflection angle of a prism
$\overline{\theta} = 90^\circ - \theta$ = Complement of angle θ
θ_a = Acceptance angle
θ_B = Brewster angle
$\theta_\mathcal{B}$ = Bragg angle
θ_c = Critical angle
$\overline{\theta}_c$ = Complementary critical angle
θ_d = Deflection angle of a prism
θ_max = Maximum angle
θ_p = Angle of incidence of a ray at the surface of a prism from its interior
θ_r = Angle of reflection
θ_s = Deviation angle of maximum spatial frequency; also, Angle subtended by a source
θ_0 = Divergence angle of a Gaussian beam
$\widehat{\boldsymbol{\theta}}$ = Unit vector in polar direction in spherical coordinates
ϑ = Threshold
$\Theta_{\ell\mathrm{m}}(\theta)$ = Hydrogen-atom associated Legendre function

κ = Elastic constant of a harmonic oscillator $[\mathrm{J} \cdot \mathrm{m}^{-2}]$

λ = Wavelength $[\mathrm{m}]$
λ_A = Long-wavelength limit $[\mathrm{m}]$; also, Activation wavelength $[\mathrm{m}]$
λ_c = Cutoff wavelength $[\mathrm{m}]$
λ_F = Wavelength spacing of adjacent Fabry–Perot resonator modes $[\mathrm{m}]$

λ_{FEL} = Wavelength of light emitted by a free-electron laser [m]
λ_g = Bandgap wavelength (long-wavelength limit) of a semiconductor [m]
λ_o = Free-space wavelength [m]
λ_p = Plasma wavelength of a metal [m]
λ_{p} = Wavelength of maximum blackbody energy density [m]; also, Peak wavelength [m]
λ_q = Eigenvalues associated with an eigenvalue problem
λ_{ZD} = Zero-dispersion wavelength of a medium [m]
λ_0 = Central wavelength [m]
λ_{dB} = de Broglie wavelength [m]
Λ = Spatial period of a grating or periodic structure [m]; also, Wavelength of acoustic wave [m]
Λ_{u} = Spatial period of undulator in a free-electron laser [m]

μ = Magnetic permeability of a medium [$\mathrm{H \cdot m^{-1}}$]; also, Carrier mobility in a semiconductor [$\mathrm{m^2 \cdot s^{-1} \cdot V^{-1}}$]; also, Mean of a random variable
μ' = Real part of the magnetic permeability of a medium [$\mathrm{H \cdot m^{-1}}$]
μ'' = Imaginary part of the magnetic permeability of a medium [$\mathrm{H \cdot m^{-1}}$]
μ_{e} = Effective magnetic permeability of a host medium with embedded objects [$\mathrm{H \cdot m^{-1}}$]
μ_e = Electron mobility [$\mathrm{m^2 \cdot s^{-1} \cdot V^{-1}}$]
μ_h = Hole mobility [$\mathrm{m^2 \cdot s^{-1} \cdot V^{-1}}$]
μ_{ij} = Component of the magnetic permeability tensor [$\mathrm{H \cdot m^{-1}}$]
μ_o = Magnetic permeability of free space [$\mathrm{H \cdot m^{-1}}$]
$\boldsymbol{\mu}$ = Magnetic permeability tensor [$\mathrm{H \cdot m^{-1}}$]

ν = Frequency [Hz]; also, Spatial frequency [$\mathrm{m^{-1}}$]
ν_A = Anti-Stokes-shifted frequency [Hz]
ν_B = Brillouin frequency [Hz]
ν_{B} = Bragg frequency [Hz]
ν_c = Cutoff frequency [Hz]
ν_F = Frequency spacing between adjacent Fabry–Perot modes (free spectral range) [Hz]
ν_i = Offset frequency for an optical frequency comb [Hz]
ν_{i} = Instantaneous frequency [Hz]
ν_p = Pump frequency [Hz]
ν_{p} = Frequency of maximum blackbody energy density [Hz]; also, Peak frequency of electroluminescence spectrum [Hz]
ν_q = Frequency of mode q [Hz]
ν_R = Raman frequency [Hz]
ν_s = Spatial bandwidth of an imaging system [$\mathrm{m^{-1}}$]; also, Signal frequency [Hz]; also, Laser frequency [Hz]
ν_S = Stokes-shifted frequency [Hz]
ν_x, ν_y = Spatial frequencies in the x and y directions [$\mathrm{m^{-1}}$]
ν_0 = Central frequency [Hz]
ν_ρ = Radial component of the spatial frequency: $\nu_\rho = \sqrt{\nu_x^2 + \nu_y^2}$ [$\mathrm{m^{-1}}$]

ξ = Coupling coefficient in four-wave mixing
$\xi_{\mathrm{sp}}(\nu)$ = Amplifier noise photon-flux density per unit length [$\mathrm{m^{-3} \cdot s^{-1}}$]

ρ = Rotatory power of an optically active medium [$\mathrm{m^{-1}}$]; also, Resistivity [$\Omega \cdot \mathrm{m}$]; also, $\rho = \sqrt{x^2 + y^2}$ = radial distance in a cylindrical coordinate system [m]
ρ_c = Coherence distance [m]
ρ_m = Radius of the center of the mth ring of a Fresnel zone plate [m]

ρ_s = Radius of the Airy disk [m]; also, Radius of the blur spot of an imaging system [m]
ϱ = Mass density of a medium [$\mathrm{kg \cdot m^{-3}}$]; also, Charge density [$\mathrm{C \cdot m^{-3}}$]; also, Retardance of a birefringent medium [m]
$\varrho(k)$ = Wavenumber density of states [$\mathrm{m^{-2}}$]
$\varrho(\nu)$ = Spectral energy density [$\mathrm{J \cdot m^{-3} \cdot Hz^{-1}}$]; also, Optical joint density of states [$\mathrm{m^{-3} \cdot Hz^{-1}}$]
$\varrho_c(E)$ = Density of states near the conduction band edge [$\mathrm{m^{-3} \cdot J^{-1}}$ in a bulk semiconductor]
$\varrho_v(E)$ = Density of states near the valence band edge [$\mathrm{m^{-3} \cdot J^{-1}}$ in a bulk semiconductor]
$\rho(\nu)$ = Normalized Lorentzian cavity mode [$\mathrm{Hz^{-1}}$]

σ = Conductivity [$\mathrm{\Omega^{-1} \cdot m^{-1}}$]
$\sigma(\nu)$ = Transition cross section [$\mathrm{m^2}$]
$\overline{\sigma}(\nu)$ = Average transition cross section [$\mathrm{m^2}$]
σ_a = Absorption cross section [$\mathrm{m^2}$]
σ_ab = Effective transition cross section for absorption [$\mathrm{m^2}$]
σ_em = Effective transition cross section for emission [$\mathrm{m^2}$]
σ_f = Transfer-function bandwidth [Hz]
σ_max = Maximum transition cross section [$\mathrm{m^2}$]
σ_q = Circuit-noise parameter
σ_r = Circuit-noise current RMS value [A]
σ_s = Scattering cross section [$\mathrm{m^2}$]
σ_x = Standard deviation of a random variable x; RMS width of a function of x
σ_x^2 = Variance of a random variable x
$\sigma_0 = \sigma(\nu_0)$ = Transition cross section at the central frequency ν_0 [$\mathrm{m^2}$]
σ_0 = Conductivity at low frequencies [$\mathrm{\Omega^{-1} \cdot m^{-1}}$]
σ_λ = Spectral width [m]
σ_τ = Temporal width [s]
$\boldsymbol{\sigma}$ = Conductivity tensor [$\mathrm{\Omega^{-1} \cdot m^{-1}}$]

τ = Lifetime [s]; also, Decay time [s]; also, Pulse width [s]; also, Relaxation time [s]; also, Scattering time [s]; also, Collision time [s]; also, Width of a function of time [s]; also, Excess-carrier electron–hole recombination lifetime in a semiconductor [s]
τ_c = Coherence time [s]
τ_col = Mean time between collisions [s]
τ_d = Delay time [s]
τ_e = Electron transit time [s]
τ_h = Hole transit time [s]
τ_m = Multiplication time in an avalanche photodiode [s]
τ_nr = Nonradiative electron–hole recombination lifetime [s]
τ_p = Resonator photon lifetime [s]
τ_pulse = Duration of a mode-locked optical pulse [s]
τ_r = Radiative electron–hole recombination lifetime [s]
τ_R = Receiver-circuit time constant [s]
τ_{RC} = RC time constant [s]
τ_s = Saturation time constant of a laser transition [s]
τ_{21} = Lifetime of a transition between energy levels 2 and 1 [s]

ϕ = Angle in a cylindrical or spherical coordinate system; also, Photon-flux density [$\mathrm{m^{-2} \cdot s^{-1}}$]
$\phi(p)$ = Particle momentum wavefunction [$\mathrm{s^{1/2} \cdot kg^{-1/2} \cdot m^{-1/2}}$]
$\phi(p)$ = Wavefunction for p quadrature component of the electric field
$\phi_s(\nu)$ = Saturation photon-flux density [$\mathrm{m^{-2} \cdot s^{-1}}$]

ϕ_ν = Spectral photon-flux density $[\mathrm{m}^{-2} \cdot \mathrm{s}^{-1} \cdot \mathrm{Hz}^{-1}]$
$\widehat{\boldsymbol{\Phi}}$ = Unit vector in azimuthal direction in spherical coordinates
φ = Phase or phase difference or phase shift
$\varphi(t)$ = Phase of the complex envelope of an optical pulse
$\varphi(\nu)$ = Phase-shift coefficient of an optical amplifier $[\mathrm{m}^{-1}]$
φ_L = Local-oscillator phase
φ_0 = Phase shift from reflection at a resonator mirror
Φ = Photon flux $[\mathrm{s}^{-1}]$
$\Phi(\nu, \tau)$ = Short-time Fourier transform; also, Wigner distribution function
$\Phi_{\mathrm{m}}(\phi)$ = Hydrogen-atom harmonic function
Φ_ν = Spectral photon flux $[\mathrm{s}^{-1} \cdot \mathrm{Hz}^{-1}]$

χ = Electric susceptibility; also, Electron affinity [J]
χ' = Real part of the electric susceptibility χ
χ'' = Imaginary part of the electric susceptibility χ
$\chi(\nu)$ = Electric susceptibility of a dispersive medium
χ_{e} = Effective electric susceptibility
χ_{ij} = Component of the electric susceptibility tensor
χ_m = Electric susceptibility of a metal at high frequencies ($\omega \gg \omega_p$)
$\chi^{(3)}$ = Coefficient of third-order optical nonlinearity $[\mathrm{C} \cdot \mathrm{m} \cdot \mathrm{V}^{-3}]$
$\chi^{(3)}_{ijkl}$ = Component of the third-order optical nonlinearity tensor $[\mathrm{C} \cdot \mathrm{m} \cdot \mathrm{V}^{-3}]$
$\chi^{(3)}_{\mathrm{I}}$ = Imaginary part of the nonlinear third-order susceptibility $\chi^{(3)}$
$\chi^{(3)}_{IK}$ = Component of the third-order optical nonlinearity tensor (contracted indices) $[\mathrm{C} \cdot \mathrm{m} \cdot \mathrm{V}^{-3}]$
$\chi^{(3)}_{\mathrm{R}}$ = Real part of the nonlinear third-order susceptibility $\chi^{(3)}$
χ = Polarization-ellipse angle of ellipticity
$\boldsymbol{\chi}$ = Electric susceptibility tensor

ψ = Normalized amplitude of an optical pulse
$\psi(\mathbf{r}, t)$ = Particle wavefunction $[\mathrm{m}^{-3/2} \cdot \mathrm{s}^{-1/2}]$
$\psi(x)$ = Particle position wavefunction $[\mathrm{m}^{-1/2}]$
$\psi(x)$ = Wavefunction for x quadrature component of the electric field
$\psi(\nu)$ = Spectral phase of an optical pulse
Ψ = Polarization-ellipse orientation of major axis
$\Psi(\mathcal{E})$ = Nonlinear polarization density $[\mathrm{C} \cdot \mathrm{m}^{-2}]$
$\Psi_e(f)$ = Envelope transfer function phase

ω = Angular frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
$\omega_{\mathcal{B}}$ = Bragg angular frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
ω_{i} = Instantaneous angular frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
ω_I = Heterodyne intermediate (angular) frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
ω_L = Local-oscillator frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
ω_p = Plasma frequency of a metal $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$; also, Pump angular frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
ω_r = Upshifted angular frequency of a Bragg-reflected wave $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
ω_s = Downshifted angular frequency of a Bragg-reflected wave $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$; also, Angular frequency below which a surface plasmon polariton can exist $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$; also, Signal frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
ω_0 = Central angular frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$; also, Localized surface plasmon resonance frequency $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$
Ω = Angular frequency of an acoustic wave $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$; also, Angular frequency of a harmonic electric signal $[\mathrm{rad} \cdot \mathrm{s}^{-1}]$; also, Solid angle [sr]

Mathematical Symbols

$\dot{=}$ = Result decreased to the nearest integer
$\dot{=}$ = Result increased to the nearest integer
$\det\{\cdot\}$ = Determinant of a matrix
$\mathrm{Tr}\{\cdot\}$ = Trace of a matrix
$\{\cdot\}^{\mathsf{T}}$ = Transpose of a matrix
$\{\cdot\}^{-1}$ = Inverse of a matrix
$\overline{x}$ = Mean of the quantity x
$\langle x \rangle$ = Ensemble average over x
d = Differential
∂ = Partial differential
∇ = Gradient operator
$\nabla\cdot$ = Divergence operator
$\nabla\times$ = Curl operator
∇^2 = Laplacian operator ($\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ in Cartesian coordinates)
∇_T^2 = Transverse Laplacian operator ($\nabla_T^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ in Cartesian coordinates)

AUTHORS

Bahaa E. A. Saleh has been Distinguished Professor and Dean of CREOL, The College of Optics and Photonics at the University of Central Florida, since 2009. He was at Boston University in 1994–2008, serving as Chair of the Department of Electrical and Computer Engineering (ECE) in 1994–2007, and becoming Professor Emeritus in 2008. He received the Ph.D. degree from the Johns Hopkins University in 1971 and was a faculty member at the University of Wisconsin-Madison from 1977 to 1994, serving as Chair of the ECE Department from 1990 to 1994. He held faculty and research positions at the University of Santa Catarina in Brazil, Kuwait University, the Max Planck Institute in Germany, the University of California-Berkeley, the European Molecular Biology Laboratory, Columbia University, and the University of Vienna.

His research contributions cover a broad spectrum of topics in optics and photonics including statistical optics, nonlinear optics, quantum optics, and image science. He is the author of *Photoelectron Statistics* (Springer-Verlag, 1978) and the co-author of *Fundamentals of Photonics* (Wiley, *First Edition* 1991, *Second Edition* 2007, *Third Edition* 2019) and *Introduction to Subsurface Imaging* (Cambridge University Press, 2011). He has published more than 600 papers in technical journals and conference proceedings. He holds nine patents.

Saleh served as the founding editor of the OSA journal *Advances in Optics and Photonics* (2008–2013), editor-in-chief of the *Journal of the Optical Society of America A* (1991–1997), and Chairman of the Board of Editors of OSA (1997–2001). He served as Vice President of the International Commission of Optics (ICO) (2000–2002), and as a member of the Board of Directors of the Laser Institute of America (2010–2011). He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), the Optical Society of America (OSA), the International Society for Optics and Photonics (SPIE), the American Physical Society (APS), and the Guggenheim Foundation. He received the 1999 OSA Beller Medal for outstanding contributions to optical science and engineering education, the 2004 SPIE BACUS award for his contribution to photomask technology, the 2006 Kuwait Prize, the 2008 OSA Distinguished Service Award, and the 2013 OSA Mees Medal. He is a member of Phi Beta Kappa, Sigma Xi, and Tau Beta Pi.

Fundamentals of Photonics, Third Edition. Bahaa E. A. Saleh and Malvin Carl Teich.
©2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Malvin Carl Teich received the S.B. degree in physics from the Massachusetts Institute of Technology in 1961, the M.S. degree in electrical engineering from Stanford University in 1962, and the Ph.D. degree from Cornell University in 1966. His first professional affiliation, in 1966, was with MIT Lincoln Laboratory. He joined the faculty at Columbia University in 1967, where he served as a member of the Electrical Engineering Department (as Chairman from 1978 to 1980), the Applied Physics and Applied Mathematics Department, the Columbia Radiation Laboratory in the Department of Physics, and the Fowler Memorial Laboratory at the Columbia College of Physicians & Surgeons. During his tenure at Columbia, he carried out research in the areas of photon statistics and point processes; quantum heterodyne detection; the generation of nonclassical light; noise in avalanche photodiodes and fiber-optic amplifiers; and information transmission in biological sensory systems. In 1996 he became Professor Emeritus at Columbia.

From 1995 to 2011, he served as a faculty member at Boston University in the Departments of Electrical & Computer Engineering, Biomedical Engineering, and Physics. He was the Director of the Quantum Photonics Laboratory and a Member of the Photonics Center, the Hearing Research Center, and the Graduate Program for Neuroscience. In 2011, he was appointed Professor Emeritus at Boston University.

Since 2011, Dr. Teich has been pursuing his research interests as Professor Emeritus at Columbia University and Boston University, and as a member of the Boston University Photonics Center. He is also a consultant to government and private industry and has served as an expert in numerous patent conflict cases. He is most widely known for his work in photonics and for his studies of fractal stochastic processes and information transmission in biological systems. His current efforts in photonics are directed toward the characterization of noise in photon streams. His work in fractals focuses on elucidating the information-carrying properties of sensory-system action-potential patterns and the nature of heart-rate variability in patients with coronary disorders. His efforts in neuroscience are directed toward auditory and visual perception, neural information transmission, and sensory detection. During periods of sabbatical leave, he served as a visiting faculty member at the University of Colorado at Boulder, the University of California at San Diego, and the University of Central Florida at Orlando.

Dr. Teich is a Life Fellow of the Institute of Electrical and Electronics Engineers (IEEE), and a Fellow of the Optical Society of America (OSA), the International Society for Optics and Photonics (SPIE), the American Physical Society (APS), the American Association for the Advancement of Science (AAAS), and the Acoustical Society of America (ASA). He is a member of Sigma Xi and Tau Beta Pi. In 1969 he received the IEEE Browder J. Thompson Memorial Prize for his paper "Infrared Heterodyne Detection." He was awarded a Guggenheim Fellowship in 1973. In 1992 he was honored with the Memorial Gold Medal of Palacký University in the Czech Republic, and in 1997 he received the IEEE Morris E. Leeds Award. In 2009, he was honored with the Distinguished Scholar Award of Boston University. He has authored or coauthored some 350 refereed journal articles/book chapters and some 550 conference presentations/lectures; he holds six patents. He is the co-author of *Fundamentals of Photonics* (Wiley, *First Edition* 1991, *Second Edition* 2007, *Third Edition* 2019) and of *Fractal-Based Point Processes* (Wiley, 2005, with S. B. Lowen).

INDEX

k surface
  anisotropic medium, 234–237
ABCD
  inverse ray-transfer matrix, 31
  law, 97–99
  ray-transfer matrix, 28, 1302
  transmission matrix, 380
  wave-transfer matrix, 259, 1303
g-parameters, 448
q-parameter, 81, 97

Absorption
  and scattering, 198–199
  atoms, 584, 589–593
  coefficient, 182, 198, 777
  coefficient in resonant medium, 189
  cross section, 199
  dielectric media, 181–184
  efficiency, 199
  indirect-bandgap semiconductor, 771
  left-handed media, 309
  occupancy probability, 773
  semiconductors, 769
  strong, 184
  transition rate, 592, 775
  weak, 183
Acceptance angle, 19, 25, 395
Acousto-optic devices, 958–967
  acoustic spectrum analyzer, 963
  figure of merit, 973
  filter, 966
  frequency shifter, 967
  intensity modulator, 959, 973
  intensity reflectance, 973
  interconnection capacity, 966
  optical isolator, 967
  reflector, 973
  scanner, 961
  spatial light modulator, 965
  switch, photonic, 959, 964, 1194
Acousto-optics, 943–974
  anisotropic media, 967–971
  Born approximation, 953, 973
  Bragg angle, 948
  Bragg condition, 948
  Bragg condition tolerance, 949
  Bragg diffraction, 945–958, 972, 973
  Bragg diffraction by beam, 956, 961
  Bragg diffraction by thin beam, 957
  Bragg diffraction, downshifted, 952
  Bragg diffraction, upshifted, 949
  coupled-wave equations, 953, 954
  Debye–Sears scattering, 957, 973
  Doppler shift, 950
  elasto-optic coefficient, 946
  figure of merit, 946
  index-ellipsoid modification, 970
  introduction to, 944–945
  Raman–Nath scattering, 973
  reflectance, amplitude, 948
  reflectance, intensity, 950
  scattering theory, 953, 973
  strain-optic coefficient, 946
Airy
  beam, 106–107
  beam generation, 107, 158
  disk, 131
  formulas, 261
  function, 107
  pattern, 131
Amplified spontaneous emission (ASE), 651–653, 699, 821, 830
Amplifier
  Brillouin fiber, 644
  chirped pulse, 1095
  electronic, 620, 621
  laser, 619–656, 659
  phase, 70
  phase-sensitive, 1062
  Raman fiber, 642–644
  reflection, 1065
  semiconductor, 817–831
  transmission, 1065
Anharmonic oscillator, 1071–1073
Anisotropic media
  k surface, 234–237

acousto-optics of, 967–971
arbitrary propagation, 232–233
biaxial crystal, 229
conductive, 882
dispersion relation, 234–237
double refraction, 238–239, 254
electro-optics of, 989–996
ellipsoid of revolution, 318
energy transport, 234–237
extraordinary refractive index, 229
four-wave mixing, 1074
gyration vector, 241
hyperbolic, 339–340, 352
hyperboloid of revolution, 318
impermeability tensor, 230
index ellipsoid, 230, 232–233
liquid crystals, 244–246, 996–1005
magneto-optics, 242–243
magnetogyration coefficient, 243
mixed-sign permittivities, 317, 352
negative uniaxial crystal, 229
negative-index, 338–339
nonlinear and dispersive, 1073
nonlinear optics of, 1066–1069
normal modes, 230, 232–233, 241
normal surface, 234–237
optic axis, 229
optical activity, 240–242
optics of, 227–239
ordinary refractive index, 229
permittivity tensor, 228
photoelastic effect, 968
positive uniaxial crystal, 229
principal axes, 229
principal refractive indices, 229
principal-axis propagation, 230–231
quadric representation, 229, 230, 317
rays, 234–237
refraction of rays, 239, 254
refractive indices of, 228–230
three-wave mixing, 1066–1069, 1073
uniaxial crystal, 229, 233, 236, 317
Verdet constant, 243
wavefronts, 234–237
Antenna
optical, 332–334
plasmonic, 332–334
Antibunching, photon, 780
Antireflection coating, 264, 301
Aperture
function, 129, 138
ring, 67, 118, 159
Approximation
Born, 192, 953, 973, 1020, 1076
Fraunhofer, 124, 125, 130
Fresnel, 49, 121, 125, 130
paraboloidal, 49
paraxial, 9, 50
quasi-static, 196, 197
quasi-stationary, 944
SVE, 51, 563, 954, 1050, 1064, 1086, 1111, 1127, 1133
Taylor-series, 10, 203, 1127
Volterra-series, 1070
WKB, 407
Array detectors, *see* Photodetectors
Atoms
absorption, 584, 589–593
atom amplifier, 602
atom chip, 602
atom interferometry, 601
atom optics, 601
atomic number, 564, 568
Bose–Einstein condensate, 601
bosonic, 568, 602
Doppler cooling, 599
electron configuration, 565, 566, 569
energy levels, 562
evaporative cooling, 600
fermionic, 568
Hartree method, 565
hydrogen, 564
interaction with broadband light, 590
interaction with light, 583–602
ionization energy, 568
isotopes, 568
laser cooling, 599
laser trapping, 600
lineshape function, 586
magnetic trap, 602
magneto-optical trap, 601
matter waves, 601
mononuclidic, 568
multielectron, 565–568
optical lattices, 601
optical molasses, 599
optical tweezers, 600
Pauli exclusion principle, 565
periodic table, 566, 567
photon recoil, 524, 555, 600
polarization gradient cooling, 600
relative atomic mass, 568
shells and subshells, 565
Sisyphus cooling, 600
spontaneous emission, 583, 587–588
spontaneous lifetime, 587, 588
spontaneous lifetime, effective, 589
stimulated emission, 585, 589–593

term symbol, 566, 569
transition strength, 586
Attenuation
and scattering, 198, 416, 432
coefficient, 182, 415–417
optical fiber, 415–417
Attosecond optics, 1141–1145
Auger recombination, 752, 821
Autocorrelation, 1288
Avalanche photodiodes, 895–907
avalanche buildup time, 899
breakdown voltage, 901
comparison with photodiodes, 927
conventional (CAPD), 895
dead space, 903
excess noise factor, 917–921
gain, 896, 898, 940
gain noise, 917–921
Geiger-mode, 904
group-IV, 902
initial-energy effects, 903
ionization coefficients, 895
ionization ratio, 896
materials, 901
position-dependent gradients, 904
position-dependent ionization, 904
punchthrough voltage, 901
reach-through device, 899
response time, 899
responsivity, 898
SACM device, 899, 902
SAM device, 899, 901, 902
SCIDCM devices, 897, 917
SCISCM devices, 896
sick space, 903
single-photon (SPAD), 904
Axicon, 14, 57, 106, 137

Band-structure engineering, 851
Bandgap
direct, 737
energy, 576, 733, 738
indirect, 737
wavelength, 738, 768
Bardeen, John, 731
Basov, Nikolai G., 619
Beam, optical, 79–109
accelerating, 107
Airy, 106–107, 158
Bessel, 57, 105, 136, 137
Bessel–Gaussian, 106, 181
Bessel/Gaussian comparison, 106
divergence–size tradeoff, 85, 106
donut, 103, 109
elliptic Gaussian, 102, 109
elliptic Hermite–Gaussian, 102
Gaussian, 51, 80–99
Hermite–Gaussian, 99–102, 105
holographic generation, 155
Ince–Gaussian, 105
introduction to, 80
Laguerre–Gaussian, 102–105, 155
nondiffracting, 57, 105, 136, 137
quality, 90
self-healing, 107
toroidal, 103, 109
vector, 179, 180
vortex, 104
Beamsplitter, 13, 63, 301, 549
dielectric-slab, 268
phase shift, 64, 269
polarizing, 239, 247
random partitioning, 539, 555
scattering matrix, 263
single-photon transmittance, 522
two-photon transmittance, 550
wave-transfer matrix, 263
Bell Laboratories, 1163
Bessel
-Gaussian beam, 106
beam, 57, 105, 136, 137
beam vs. Gaussian beam, 106
function, 105, 131, 398, 467, 1295
Bessel, Friedrich Wilhelm, 79
Bioluminescence, 608
Biprism, 14, 57
Fresnel, 14, 57
Bistability, optical
amplifying NL element, 1218, 1219
coupled microring lasers, 1219
devices, 1216–1220
dispersive NL element, 1216, 1220
dissipative NL element, 1217
embedded system, 1215
Fabry–Perot, nonlinear, 1217, 1218
hybrid, 1220
hysteresis, 1211
intrinsic system, 1215
MZI, nonlinear, 1216, 1218
photonic logic, 1223
principles of, 1213–1215
quantum-confined Stark effect, 1220
saturable absorber, 1217
self-electro-optic-effect device, 1220
spatial light modulator, 1220
systems, 1211
Blackbody radiation, 602–607
1D, 618, 923

2D, 617
spectrum, 604
correlated color temperature, 812
emission comparison, 617
Rayleigh–Jeans formula, 463, 606
Stefan–Boltzmann law, 606, 618
thermal equilibrium, 602
thermal light, 602
thermography, 606
Wien's law, 617
Bloch
modes, 257, 277–282, 295, 734, 1305
phase, 281
wavenumber, 278
Bloch, Felix, 255
Bloembergen, Nicolaas, 1015
Bohr
atom, 564
period, 564, 705
radius, 564, 784
Bohr, Niels, 561
Boltzmann distribution, 536, 581
Bonding
covalent, 573, 576, 743
ionic, 573, 576
metallic, 576
van der Waals, 573, 576, 743
Born approximation
for acousto-optics, 953, 973
for nonlinear optics, 1020
for scattering, 192–193
Born postulate, 563
Born, Max, 473
Bose–Einstein
condensate, 601, 602
distribution, 536
Boson, 524, 537, 568
Boundary
conditions, 164, 221
planar, 11
spherical, 14
Bragg
angle, 67, 270, 948
condition, 270, 948
condition, tolerance, 949
diffraction, 945–958, 972, 973
diffraction by acoustic beam, 956, 961
diffraction of optical beam, 955
diffraction, quantum view, 952
diffraction, Raman–Nath, 957
frequency, 270
reflection, 67
Bragg grating, 269–276, 698
chirp filter, 1099, 1161
distributed Bragg reflector, 269, 434, 466, 471, 798
fiber, 269, 687, 692
total reflection, 273
waveguide, 385
Bragg, William Henry, 943
Bragg, William Lawrence, 943
Brattain, Walter H., 731
Bremsstrahlung, 562
Brewster
angle, 224, 225, 247, 253
window, 225
Brillouin
fiber amplifier, 644
scattering, 613
zone, 278, 279, 283, 735, 738
zone, irreducible, 295
Bulk optics, 305, 354

Catadioptric system, 15
Cathodoluminescence, 607, 608, 618
Caustic
curve, 9, 17
surface, 397
Characteristic equation
optical fiber, 400
Chemiluminescence, 608
Cherenkov radiation, 562, 875
Chirp
coefficient, 1090
coefficient, angular-dispersion, 1097
coefficient, Bragg-grating, 1099
function, 1289
parameter, 1082
pulse amplifier, 1095
Chirp filter, 1088–1099
angular-dispersion, 1096
arbitrary phase filter, 1090
Bragg-grating, 1099, 1161
cascaded, 1090
chirp coefficient, 1090
diffraction-grating, 1098
envelope transfer function, 1088
for compression, 1094
for expansion, 1094
ideal filter, 1089
implementations of, 1095–1099
impulse response function, 1090
optical-fiber, 1103–1111, 1161
phase filter, 1089
prism, 1097, 1105, 1161
vs. quadratic-phase modulator, 1100
Circular

dichroism, 254
polarization, 213, 253, 524
Circulator, optical, 251, 1168, 1173
CMOS technology, 375, 907, 909
scientific CMOS (sCMOS), 909
Coherence
area, 485, 495
coherent light, 62, 152, 474, 544
cross-spectral density, 486
cross-spectral purity, 486
degree of coherence, 483
degree of temporal coherence, 478
distance, 495, 505
enhancement via filtering, 481, 489
extended radiator, 485
gain via propagation, 502
incoherent light, 485
introduction to, 474–475
length, 479
length, wave-mixing, 1032
longitudinal, 487
mutual coherence function, 483
mutual intensity, 485, 498, 502, 549
partial polarization, 506–510
power spectral density, 479
quantum, 483
quasi-monochromatic light, 485
random wavepacket sequence, 482
role in interference, 489–496
role in interferometry, 490, 493
spatial, 483–487
speckle, 406
spectral density, 479
spectral width, 481
temporal, 476–483
temporal coherence function, 477
time, 478
transmission, 497
van Cittert–Zernike theorem, 503
Wiener–Khinchin theorem, 480
Wolf equations, 484
Coherent
anti-Stokes Raman scattering, 614
detection, 74, 1266
lidar, 75
optical amplifier, 620
Colladon, Jean-Daniel, 353
Collimator, 9, 15
Color-rendering index, 812, 816
Comb, optical frequency, 721, 1141
applications, 722
Complex
amplitude, 46, 80
analytic signal, 71
degree of coherence, 483
degree of temporal coherence, 478
envelope, 47, 73, 80, 175, 211, 1080
representation, 45
wavefunction, 45, 72
Conductive medium, 320–326
power reflectance, 321
Conductivity, 320, 882
tensor, 882
Constitutive relation, 166
Converter
evanescent-to-propagating, 317
frequency, 548, 610, 611, 1027, 1054
incoherent-to-coherent, 989, 1005
indistinguishable-to-entangled, 550
mode, 789, 1265
space-to-time, 1119
wavelength, 1205
Convolution, 1288, 1294, 1296, 1300
Corkum, Paul B., 1078
Correlation, 1288
Coupled-wave theory
acousto-optics, 953, 954
four-wave mixing, 1059
three-wave mixing, 1047, 1127
waveguide, 378
Coupler
prism, 329
Critical angle, 12, 223, 225, 310, 327
complementary, 363
Cross section
absorption, 199
effective absorption, 585, 634
effective emission, 585, 634
Füchtbauer–Ladenburg equation, 588
lineshape function, 586
scattering, 195, 196
transition, 584, 588, 634, 644
transition strength, 586
Cross-phase modulation, 1041, 1197
Crystal
biaxial, 229
negative uniaxial, 229
positive uniaxial, 229
structure, 294
symmetry, 991
uniaxial, 229
Cutoff frequency
waveguide, 359, 366

Debye–Sears scattering, 973
Degeneracy parameter, 582
Delay, group, 200
Detectors, *see* Photodetectors

Dichroism, 247
circular, 254
Dielectric
boundary, reflection from, 53
boundary, refraction at, 53
constant, 167
Diffraction, 129–136
aperture function, 129
Bragg, 945–958
circular aperture, 131
dispersion analogy, 1112
focused optical beam, 132
Fraunhofer, 130–132, 158
Fresnel, 132–137, 158
Gaussian aperture, 134
grating, 59, 78, 158
nondiffracting beams, 105–107
nondiffracting waves, 136
periodic aperture, 135
rectangular aperture, 130
single-photon, 523
slit, 133
Talbot effect, 135
two pinholes, 158
Diffusion equation, 1111, 1112
Diode lasers, *see* Laser diodes
Dipole
electric, 177
magnetic, 178
moment, 177, 178
wave, 177, 194
Dirac equation, 565
Directional coupler, 381, 984
integrated-photonic, 986
nonlinear, 1186
soliton router, 1187
Dispersion, 184–191
-compensating fiber, 423
-flattened fiber, 423
-shifted fiber, 422
angular, 1096
anomalous, 202
chromatic, 422
coefficient, 201, 202, 419
compensation, 685, 1110, 1111, 1254
diffraction analogy, 1112
diffractive, 1096
group velocity, 201, 207, 370, 425, 550, 1104
in multi-resonance medium, 204
interferometric, 1096
management, 1255
material, 419, 1096
material and modal, 420
measures, 185
modal, 370, 389, 418, 1096
multipath, 1096
nonlinear, 425, 1096
normal, 202
optical-fiber, 418–426
polarization mode, 423, 1096
pulse propagation with, 199–203
spatial, 1096
waveguide, 421, 1096
Dispersion relation
anisotropic medium, 234–237
photonic crystal, 282, 285, 296
waveguide, 359, 370
Dispersive medium
anisotropic and nonlinear, 1073
four-wave mixing, 1074
nonlinear optics of, 1069–1074
three-wave mixing, 1073, 1076, 1162
Distributed Bragg reflector, 269, 434, 466, 471, 798
fiber, 687, 692, 720
laser, 843
Doped dielectric media, 568–572
actinide metals, 572
lanthanide metals, 571, 572
transition metals, 569
Doppler
effect, 75
radar, optical, 75
shift, 950
Double refraction, 238–239, 254
Double-slit experiment, 65, 493, 525
Michelson stellar interferometer, 505
role of source size, 495
role of spectral width, 495
Downconversion, 1077
Downconversion, parametric, 1027
Drude model, 322–326, 388
group velocity, 352
simplified, 324–326
Drude, Paul Karl Ludwig, 303
Drude–Lorentz model, *see* Drude model

Efficiency
absorption, 199
differential power-conversion, 670
external, 800
external differential, 836, 838
extraction, 670, 838
internal, 754, 796, 838
optical-to-optical, 671
optical-to-optical slope, 671
overall, 670

photon detection, 876
power-conversion, 670, 837, 838
scattering, 196
slope, 670, 837
wall-plug, 670, 838
Eigenvalue problem, 220, 279, 563, 1301
Eikonal equation, 26, 52, 343, 407
Einstein
$\mathbb{A}$ and $\mathbb{B}$ coefficients, 591
postulates, 591
Einstein, Albert, 515, 561, 871
Electric
dipole, 177
dipole moment, 177
field, 162
field, internal, 196
flux density, 163
Electro-optic devices
directional coupler, 381, 984
dynamic wave retarder, 980
intensity modulator, 981, 995, 1014
interferometric photonic switch, 1192
liquid-crystal SLM, 1002–1005
phase modulator, 973, 979, 995, 1013, 1014
self-electro-optic-effect device, 1220
spatial light modulator, 987–989
strain sensor, 1013
switch, 985
Electro-optics, 975–1014
anisotropic media, 989–996
double refraction, 1014
index-ellipsoid modification, 989
introduction to, 976–977
Kerr effect, 978, 994, 1037, 1077
Pockels effect, 978, 991–994, 1024, 1077
principles of, 977–989
Electroabsorption, 1010–1012
Franz–Keldysh effect, 1010
modulator, 1010, 1011
switch, 1010
Electrochromism, 1009
Electroluminescence, 608, 789
injection, 609, 790
Electromagnetic optics, 160–208
constitutive relation, 166
introduction to, 161–162
material equation, 173
relation to wave optics, 180
Electromagnetic wave
amplitude-modulated, 207
dipole wave, 177, 194
energy, 165
energy density, 176
Gaussian beam, 179
in a medium, 163
in absorptive medium, 181, 309–310
in anisotropic medium, 170
in conductive medium, 320–326
in dielectric medium, 166–180
in dispersive medium, 170, 184, 199
in double-negative medium, 308
in double-positive medium, 308
in free space, 162
in inhomogeneous medium, 168, 174
in negative-index material, 314
in nonlinear medium, 171
in resonant medium, 186
in single-negative medium, 308
intensity, 165, 172, 176
momentum, 165
monochromatic, 172–180
plane wave, 175, 307
power, 165, 172
Poynting vector, 165
scattered, 194, 196
spherical wave, 177
TEM wave, 175, 307
traveling standing, 207
vector beam, 180
Electron configuration, 566
actinide metals, 569
He, 566
lanthanide metals, 569
Ne, 566, 701
Ne^+, 701
Ne-like, 699
Ni-like, 699
rare-earth elements, 569
transition metals, 569
Electroweak theory, 515
Elements
actinide metals, 569, 572
lanthanide metals, 569, 571
periodic table, 566, 567
rare-earth, 571, 682, 687
semiconductor, 686
transition metals, 569, 745
Energy
activation, 885
anisotropic transport, 234–237
density, electromagnetic, 176
electromagnetic, 165, 172
optical, 44
photon, 518
Energy levels
alexandrite, 570

azimuthal quantum number, 564
bandgap energy, 576, 578, 733, 738
Bohr atom, 564
Boltzmann distribution, 581
C^{5+}, 564
CO_2 molecule, 574
conduction band, 576
Cr^{3+}:crysoberyl, 570
Cr^{3+}:sapphire, 570
crystal-field theory, 570
degeneracy, 572, 582
diatomic molecules, rotating, 573
diatomic molecules, vibrating, 573
dye molecule, 575
Fermi–Dirac distribution, 582
fine structure, 565
forbidden band, 576
H, 564
He, 566
hyperfine structure, 565
ionization energy, 568
ions, 568–572
lanthanide-ion manifolds, 572
ligand field theory, 570
magnetic quantum number, 564
manifold, 566, 572, 634, 683
molecular, 573
multielectron atoms, 565–568
Nd^{3+}:glass, 571, 595
Nd^{3+}:YAG, 571, 595
Ne, 566
occupation, 581
principal quantum number, 564
quantum dots, 580
relativistic effects, 565
rotational quantum number, 573
ruby, 570
spin–orbit coupling, 565
Stark splitting, 568, 601, 1010
triatomic molecule, vibrating, 574
valence band, 576
vibrational quantum number, 573
vibrational–rotational, 574
Zeeman splitting, 568, 599, 601
Etalon
Fabry–Perot, 69, 265–269, 679, 728
Evanescent wave
amplified, 316
at DPS-DPS boundary, 224
at DPS-SNG boundary, 310
decay, 120, 146
nanolaser, 862
near-field imaging, 146
waveguide, 368
Excess noise factor
CAPD, 917, 920, 940
dark-noise, 921
dead-space APD, 918
definition, 911, 916
modified, 920
photoconductive detector, 940
PMT, 916, 920, 940
position-dependent APD, 919
relation to gain variance, 916
SACM APD, 918
SAM APD, 918
SCIDCM CAPD, 940
SCISCM CAPD, 940
staircase APD, 919–921
superlattice APD, 921
Excitons
bulk semiconductors, 767
electroabsorption, 1010
organic semiconductors, 784, 810
quantum dots, 580, 850
quantum-confined, 779
Extinction
coefficient, field, 198, 311, 312, 368
coefficient, intensity, 198
coefficient, waveguide, 368
ratio, 983

Fabry, Charles, 433
Fabry–Perot
etalon, 69, 265–269, 471, 679, 728
finesse, 69, 266, 441, 443
free spectral range, 267, 438, 446
internal intensity, 68, 441
loss coefficient, 443
loss factor, 443, 444
losses, 440, 442–446
modal spectral width, 442
off-axis modes, 446
photon lifetime, 444, 471
quality factor, 445
resonance frequencies, 437, 446
resonator, 69, 265, 436–446, 661
spectral width, 440
transmittance, 266–268
Faraday
effect, 242–243
rotator, 250, 1173
Fast light, 204
Fermat's principle
maximum time, 38
minimum time, 6
Fermat, Pierre de, 3
Fermi

-Dirac distribution, 582, 747
energy, 582
function, 582, 746
inversion factor, 777
level, 747, 748, 785
tail, 748
velocity, 744
Fermion, 524, 565, 568
Dirac-, 744
Fiber optics, 391–432
infrared, 429
introduction to, 392–393
Fiber, optical, 18, 19
V parameter, 399
absorption bands, 416
acceptance angle, 19, 25, 395
attenuation, 415
Bragg grating, 269, 687, 692, 720
characteristic equation, 400
chromatic dispersion, 422
cladding, 392
communications, 1224–1286
core, 392
coupler, 414, 432
differential group delay, 423, 425
dispersion, 418–426
dispersion relation, 400
dispersion-compensating, 423, 1110
dispersion-flattened, 423
dispersion-shifted, 422
extrinsic absorption bands, 417
fiber optics, 391–432
fiber parameter, 399
grade-profile parameter, 396, 412
graded-index, 392, 396, 409, 432
group velocity, 404, 412, 432
guided rays, 393–397
guided waves, 397–415
holey, 256, 426–428
hybrid, 429
infrared, 429
introduction to fiber optics, 392–393
material and modal dispersion, 420
material dispersion, 419, 432
materials, 429–430
meridional ray, 394
metamaterial, 428
modal dispersion, 392, 418, 432
mode cutoff, 402
modes, 400, 409, 432
multicore, 413, 1263
multicore couplers, 413
multimode, 392
nonlinear dispersion, 425
number of modes, 403, 410
numerical aperture, 19, 395
optimal index profile, 412
photonic lantern, 414
photonic-crystal, 426–428
polarization mode dispersion, 423
polarization-maintaining, 406
power transmission ratio, 416
propagation constant, 404, 411, 432
quasi-plane wave solutions, 407, 432
second-harmonic generation, 1023
silica-glass, 392
single-mode, 392, 405
skewed ray, 394
specialty, 429
SPM, 1130
step-index, 392–396, 398–405, 408
waveguide, 392
waveguide dispersion, 421
Finesse, 69, 266, 441
First-order optics, *see* Paraxial optics
Fluorescence, 609, 811
fluorophore, 610
three-photon, 610
two-photon, 610
up-conversion, 610, 611
Focal length
arbitrary paraxial system, 31
cylindrical lens, 17
Fresnel zone plate, 67, 118
optical Kerr lens, 1039
paraboloidal mirror, 9
pulse focusing system, 1110
spherical lens, 16, 17
spherical mirror, 11
time lens, 1113
Focal point
SELFOC lens, 24
arbitrary paraxial system, 31
cylindrical lens, 346
paraboloidal mirror, 8
Four-wave mixing, 1042, 1059, 1074
degenerate, 1044, 1045
in supercontinuum generation, 1140
switch, photonic, 1199
Fourier optics, 110–159
amplitude modulation, 116
far-field Fourier transform, 124–126
Fourier transform via lens, 158, 549
free-space impulse response, 122
free-space propagation, 113–123
free-space transfer function, 119, 316
frequency modulation, 116
Fresnel approximation, 120–121

Fresnel zone plate, 118
Huygens–Fresnel principle, 123
imaging, 117
impulse response function, 112
introduction to, 111–112
optical Fourier transform, 124–128
periodic media, 286–289
plane wave, 113
pulsed waves, 1117
scanning, 117
spatial frequency, 111
spatial frequency and angle, 113
spatial harmonic function, 113
spatial spectral analysis, 114
transfer function, 112
Fourier transform
autocorrelation, 1288
circularly symmetric function, 1294
convolution, 1288, 1294
correlation, 1288
far-field, 124–126
one-dimensional, 1287–1293
optical, 124–128, 1118
pairs, 1289
Parseval's theorem, 914, 1288
properties, 1288, 1294
separable functions, 1294
short-time, 1083
spectroscopy, 491
table, 1289
two-dimensional, 1293–1295
using lens, 126–128, 158
window function, 1083
Fourier, Jean-Baptiste Joseph, 110
Franken, Peter, 1015
Franz–Keldysh effect, 1010
Fraunhofer
approximation, 124
diffraction, 130–132, 158, 1090
Fraunhofer, Josef von, 110
Frequency
-resolved optical gating, 1156
beat, 74, 967, 1153, 1267
conversion, 1027, 1054
instantaneous, 1081
modulation, spatial, 116
of light, 42
shifter, acousto-optic, 967
Fresnel
approximation, 49, 120–121, 1118
biprism, 14, 57
diffraction, 132–137, 158, 1090
equations, 223, 253, 258
integrals, 134
lens, 18
number, 50, 121, 125, 458
reflection, 639, 797
zone plate, 67, 118
Fresnel, Augustin-Jean, 210

Gabor, Dennis, 110
Gain
amplifier, parametric, 1056
amplifier, reflection, 1065
CAPD, 896
conversion, 1271
laser amplifier, 623
laser diode, 834
photoconductive detector, 884
Raman, 642, 1041
saturated, 647, 649
SCIDCM CAPD, 898, 940
SCISCM staircase APD, 940
secondary-emission, 916
semiconductor optical amplifier, 818
Gain coefficient
amplifier, parametric, 1057
peak, 821, 832, 869
quantum-well SOA, 828
Raman, 1041
saturated, 646, 649, 660
small-signal, 622, 660
SOA, 820
Gate, photonic logic, 1165, 1211
Gauss, Carl Friedrich, 79
Gaussian
chirped pulse, 1084–1085
pulse, 1084, 1161
pulse, Gaussian-beam analogy, 1113
Gaussian beam, 51, 80–99
$\mathbb{M}^2$ factor, 90
$ABCD$ law, 97–99
q-parameter, 81, 97
Bessel-beam comparison, 106
characterization, 88
collimation, 95
complex amplitude, 80–82, 123
complex envelope, 81
confocal parameter, 85
depth of focus, 85
divergence angle, 85, 109
elliptic, 102, 109
expansion, 95
focusing, 93, 109, 1122
Gaussian-pulse analogy, 1113
Gaussian-pulsed, 1120
Gouy effect, 86
intensity, 82

parameters, 89, 90, 108
paraxial approximation, 89
phase, 86
power, 83
properties, 82–91, 451
pulsed, 1087
quality, 90
radius of curvature, 86
Rayleigh range, 81
reflection from spherical mirror, 96
relaying, 94
shaping, 93
single-photon, 522
spherical-mirror resonator, 451, 453
spot size, 84, 109
standing wave, 471
through arbitrary system, 97
through components, 91–99
through free space, 98
through graded-index slab, 109
through thin lens, 91–93
through transparent plate, 99
vector, 179
waist radius, 84
wavefronts, 86
width, 84
Gaussian optics, *see* Paraxial optics
General Electric Corporation, 787
Geometrical optics, *see* Ray optics
Goos–Hänchen effect
shift, 253
waveguide, 371
Gordon, James P., 1078
Gouy effect, 86
Graded-index
GRIN material, 20, 343
SELFOC slab, 23
fiber, 24, 40, 396–397, 409
lens, 24
optics, 20–25, 343
ray equation, 20–21
ray equation, paraxial, 21
slab, 22
Graphene photonics, 719, 744
Grating
coupler, 377
diffraction, 59, 78
Grazing incidence, 223, 311, 434, 698
Group
delay, 200
index, 201
velocity, 200, 352
velocity dispersion, 201, 207, 370, 425, 550, 1104
velocity, fiber, 404, 412
velocity, photonic crystal, 284
velocity, waveguide, 360, 370
Group-IV photonics, 737, 743
2D materials, 745, 780
allotropes, 743
array detector, 909
avalanche photodiode, 902
GeSn-on-Si laser, 860
graphene photonics, 719, 744
microcavity laser, 859
photodiode, 892
Schottky-barrier photodiode, 894
SiC Schottky diode, 809
silicon photonics, 780, 808
transition-metal dichalcogenides, 745
Guided-wave optics, 353–390
Gyration vector, 241

Harmonic oscillator, 170, 323, 1072, 1298
analogy with optical mode, 543
energy, 542
quantum theory, 542, 543, 573
Heaviside, Oliver, 164
Helmholtz equation, 46, 81, 173, 192
coupled, 954, 1049, 1060
coupled paraxial, 1050
generalized, 257, 343, 1304
nonlinear, 1134
optical fiber, 397
paraxial, 51, 78, 81, 102
two-dimensional, 105
Hermite polynomials, 100, 542
Hermite–Gaussian beam, 99–102, 105
axial phase, 109
complex amplitude, 101
elliptic, 102
excess phase, 101
Hermite polynomials, 100
Hermite–Gaussian functions, 101
intensity, 102
power confinement, 109
superposition, 103, 109
Hermite–Gaussian functions, 542
Hermitian operator, 1302
Hero's principle, 6
Hertz, Heinrich, 871
Heterodyne
optical, 74, 545, 967, 1153, 1267
Heterostructures
organic semiconductors, 810
photoconductor, 886
photodiode, 891
SOAs, 825

High-harmonic generation, 1141–1145
in Ar, 1144
optical frequency comb, 723
recollisional model, 1141, 1145
Hilbert transform, 625, 1298
Hole burning
spatial, 673
spectral, 651, 674
Hologram, 148
computer-generated, 155
grating vector, 153
holographic optical element, 155
interconnection, 1222
interconnection capacity, 1170
Laguerre–Gaussian beam, 104, 155
metasurface, 342
oblique plane wave, 149, 159
point source, 149, 155, 159
rainbow, 154
volume, 153
Holography, 147–155
ambiguity term, 150
apparatus, 152
computer-generated, 155, 1171
dynamic, 1009
Fourier-transform, 150
hologram, 148
holographic code, 148
object wave, 148
off-axis, 150
optical correlation, 159
real-time, 1045
reconstructed wave, 149
reference wave, 148
spatial filters, 151
surface-relief, 1172
volume, 153
Huygens, Christiaan, 41
Huygens–Fresnel principle, 123, 193
Hyperbolic
media: Type-I & Type-II, 352
medium, 317–320
metamaterial, 339–340, 352

IBM Corporation, 787
Imaging, 137–147, 499
2-f, 127, 549
4-f, 139
4-f impulse response function, 141
4-f system transfer function, 139
diffraction limit, 146
equation, 11, 17, 32, 58, 138, 315
equation, incoherent light, 499
focal-plane array (FPA), 907
functional, 610
hyperbolic medium, 318–319
image intensifier, 875
incoherent light, 499–502
incoherent vs. coherent, 500–502
multiphoton microscopy, 611
near-field, 146, 334
negative-index slab, 315–317
optical correlation, 159
paraxial system, 31
perfect, 315, 318
phase object, 158
point-spread function, 499
scanning near-field microscopy, 147
single-lens, 58, 137, 141, 500
single-lens impulse response function, 141, 500
single-lens transfer function, 144
single-photon, 522
spatial filtering, 141, 159
spherical mirror, 11
structural, 610
subwavelength, 146, 315, 334
thick lens, 32
thin lens, 31
three-photon microscopy, 611
two-photon, 549, 550
two-photon microscopy, 610
two-point resolution, 159
X-ray, 705
Impedance, 176
complex, 183
imaginary, 308, 310
Impermeability, electric
effect of electric field, 990
in magneto-optic material, 243
tensor, 230, 989
Impulse response function, 1296, 1299
4-f imaging, 141
free space, 122
incoherent light, 499
single-lens imaging, 137, 141, 500
Ince–Gaussian beam, 105
Incoherent light, *see* Coherence
Index ellipsoid, 230, 232–233, 989
acoust-optic modification, 968–970
electro-optic modification, 989–990
Index of refraction, *see* Refractive index
Infrared
frequencies, 42, 853
LWIR band, 42, 853
molecular-fingerprint region, 853
MWIR band, 42, 853
optical fiber, 429

sensor card, 612
wavelengths, 42, 853
Injection lasers, *see* Laser diodes
Instantaneous frequency, 1081
Insulators
band structure, 577
Integrated
optics, 354, 375, 808
photonics, 354, 375, 808
Intensity
autocorrelation function, 1150
average, 475
Bessel beam, 106
Bessel-like beam, 107
electromagnetic, 165, 172, 176
elliptic Gaussian beam, 102
Gaussian beam, 82, 106
Hermite–Gaussian beam, 102
instantaneous, 476
irradiance, 44, 812
Laguerre–Gaussian beam, 103
measurement for pulse, 1146
optical, 44, 46
partially coherent light, 475
polychromatic light, 72
random, 476
scattered, 193, 194
spectral density, 480
Interconnect, optical, 1166–1178
circulator, 1168, 1173
computer-com, 1174–1177
diffractive, 1168
free-space, 1168–1172
guided-wave, 1172
holographic, 1177
inter-board, 1174
inter-chip, 1174
interconnection matrix, 1166
intrachip, 1176
introduction to, 1164–1165
isolator, 1167
nonreciprocal, 1167, 1173
nonreciprocal, multiport, 1167
optochip, 1174
rationale, 1177
refractive, 1168
Interference, 61–71, 489–496
Bragg reflection, 67
double-slit experiment, 65, 493, 525
equation, 61, 489
finite number of waves, 66, 1183
Fourier-transform spectroscopy, 491
Fresnel zone plate, 67
infinite number of waves, 68
light from extended source, 495
multiple waves, 65–71, 75
oblique-plane-wave, 64
OCT, 492
partially coherent light, 489–496
plane-wave and spherical-wave, 64
QOCT, 550
single-photon, 525, 555
spherical-wave, 65, 493
two-photon, 550
two-wave, 61, 74, 489
visibility, 78, 490
Interferometer, 62–71
double-slit, 65, 493, 495, 505, 512, 525
Fabry–Perot, 69, 70
Fabry–Perot, nonlinear, 1217
finesse, 69, 266, 441
gravitational-wave, 70–71, 545
gyroscope, 63
Hong–Ou–Mandel, 550
interferogram, 491
LIGO, 70, 545
Mach–Zehnder, 62, 1181
Michelson, 62, 70, 75, 78, 491
Michelson stellar, 505
multipath, 66, 1183
MZI, NL, 1185, 1216, 1218, 1223
nonlinear, 1154, 1156, 1162
role of spatial coherence, 493
role of temporal coherence, 490
Sagnac, 62
Sagnac, nonlinear, 1185, 1201
self-referenced spectral, 1154
single-photon, 526
spectral, 1153
stellar, 505
temporal, 74, 1153
Invisibility cloak, 347–348
metamaterials, 348
Ionization
coefficients, 895
coefficients, history-dependent, 903
coefficients, position-dependent, 903
energy of a donor electron, 742
energy of an atom, 568
energy of Ar, 1144
energy of H, 564, 568, 742
ratio, 896
Ions
actinide metals, 572
electron configuration, 569
lanthanide metals, 568, 571, 572
noble-gas lasers, 568

term symbol, 569
transition metals, 569
Irradiance, 44, 812
Isolator, optical, 250, 967, 1167

John, Sajeev, 255
Jones matrix, 217–221
cascaded devices, 219, 253
coordinate transformation, 219
diagonal, 222
field version, 509
half-wave retarder, 219, 220
linear polarizers, cascaded, 254
normal modes, 220
polarization rotator, 219, 549
polarizer, 217, 253
quarter-wave retarder, 218
wave retarder, 218, 253
wave-retarder cascade, 219, 253
Jones vector, 215–221
coordinate transformation, 219
field version, 507
normal modes, 220
orthogonal expansion, 215, 216
two-photon, 547, 548

Kaminow, Ivan Paul, 1224
Kao, Sir Charles Kuen, 1224
Keck, Donald B., 391
Kerr
coefficient, 978, 1037, 1077
effect, 978, 994, 1037
lens, 1039
medium, 1036, 1129, 1149
medium characteristic length, 1130
switch, 1148
Kerr, John, 975
Kramers–Kronig relations, 186, 1298

Laguerre
-Gaussian beam, 102–105, 155
generalized polynomial, 103
polynomial, 103
Laguerre, Edmond Nicolas, 79
Laser, 657–730
Q-switched, 708, 712, 716, 730
Ag^{19+}, 700, 706
alexandrite, 570, 645, 706
Ar^{+}-ion, 568, 645, 678, 706, 720
ArF exciplex, 645, 706
Brillouin fiber, 692
broadening, homogeneous, 672
broadening, inhomogeneous, 673
C^{5+}, 645, 699, 706
cascaded Raman fiber, 692
cascaded silicon Raman, 692
cavity dumping, 708, 730
ceramic hosts, 681
chemical, 696
CO, 854
CO_2, 645, 706, 720, 854
coherent-state generation, 544
cooling, 599
Cr^{2+}:ZnS, 645, 686, 706, 854
Cr^{2+}:ZnSe, 686, 854
Cr^{3+}:colquiriite, 682
Cr^{3+}:crysoberyl, 570, 645, 706
Cr^{3+}:sapphire, 570, 645, 657, 664, 681, 706, 716, 730
Cr^{4+}:forsterite, 645, 686, 706, 720
crystalline hosts, 681
Cu Kα, 645, 701
dopant ions, 681
DPSS, 682, 690, 789
dye, 696, 697
efficiency, 669
Er^{3+}:silica fiber, 645, 689, 706, 720
excimer, 695
exciplex, 695
extreme-ultraviolet, 697–702
Fe^{2+}:ZnS, 686
Fe^{2+}:ZnSe, 686
fiber, 687–691
four-level pumping, 630, 633
free-electron, 702–705
frequency pulling, 664
gain clamping, 666
gain switching, 707, 711, 730
gas, 695–696
glass hosts, 681
H_2O, 706
HAPLS, 640, 849, 1095
HCN, 706
He–Ne, 645, 706, 720
in-band pumping, 634, 686, 689, 697
incoherent-feedback, 693
InGaAsP, 645
inner-shell photopumped, 700
intracavity tilted etalon, 679
ion, 695
ionized-atom plasma, 699
Kr^{+}-ion, 568, 706
KrF exciplex, 706
lasing without inversion, 663
linewidth, 680
loss coefficient, 661
metal-nanocavity, 863
methanol, 706

microcavity, 854–862
microdisk, 859
microring, 859
microring, coupled, 1219
mode locking, active, 719
mode locking, passive, 719
mode-locked, 685, 709, 716–723
molecular, 695
MOPA, 688
multiple-mirror resonator, 679
multiquantum-dot, 850
multiquantum-well, 845
multiquantum-wire, 849
nanocavity, 862–864
nanoring, 863
nanosphere, 864
Nd^{3+}:CaF_2, 572
Nd^{3+}:glass, 571, 595, 645, 706, 720
Nd^{3+}:YAG, 571, 645, 683, 706, 720
Nd^{3+}:YVO_4, 645, 682, 706
Ne Kα, 645, 701, 706
non-plasmonic nanocavity, 863
number of modes, 672
optical vortex, 860
oscillation conditions, 662
oscillation frequencies, 664, 665
output characteristics, 666–680
overall efficiency, 706
petawatt, 640, 849, 1095
phase noise, 680
phonon-terminated, 686
photon lifetime, 662
photonic-bandgap fiber, 688
photonic-crystal, 860
photonic-crystal array, 861
plaser, 693
plasmonic nanocavity, 863
polarization, 677, 678
powder, 693
power-conversion efficiency, 706
pulsed, 707–721
pumping, 629, 635
quantum cascade, 851–854
quantum-confined, 844–854
quantum-dot, 850
quantum-well, 845
quantum-wire, 849
quasi-three-level pumping, 633
quasi-two-level pumping, 634
radar, 75
Raman fiber, 691–692
random, 693–695
rate equations, 709
rhodamine-6G dye, 645, 706, 720
ribbon fiber, 688
ruby, 570, 645, 657, 664, 681, 706, 716, 730
SASE, 703
Schawlow–Townes linewidth, 680
Se^{24+}, 699
seed, 688, 700, 701
SGDFB, 853
silicon Raman, 692, 809
slab-waveguide fiber, 688
solid-state, 681–686
spatial distribution, 675
spectral distribution, 671
spontaneous lifetime, 644
strained-layer QW, 846
theory of oscillation, 659–665
thin-disk, 683, 684, 706
three-level pumping, 632, 633
threshold, 662, 663, 698
thresholdless, 861
Ti^{3+}:sapphire, 645, 685, 706, 720
Tm^{3+}:silica-fiber, 690, 706
transient effects, 709–721
transition parameters, 644
trapping, 600
two-level pumping, 655
U^{3+}:CaF_2, 572
unipolar, 851
unstable-resonator, 677
vibronic, 570, 685, 686
W^{46+}, 699
wall-plug efficiency, 706
wavelengths, 644, 706
X-ray, 697–702
X-ray free-electron, 704–706
Yb^{3+}:silica fiber, 639, 688, 706, 720
Yb^{3+}:YAG, 645, 729
Yb^{3+}:YAG thin-disk, 684, 706
ZnO, 693, 694
Laser amplifier, 619–656, 659
ASE, 651–653, 699, 821, 830
bandwidth, 624
broadband, 655
coherent, 620
Doppler-broadened medium, 650
Er^{3+}:silica fiber, 641–642
four-level pumping, 630, 633, 729
gain, 623
gain coefficient, 622, 660
gain coefficient, saturated, 646, 649
gain, saturated, 647, 649
homogeneously broadened, 645–649
in-band pumping, 634, 640, 642
in-line, 636

incoherent optical, 620
inhomogeneously broadened, 649
line, 636
MOFA, 636
MOPA, 636, 688
National Ignition Facility, 639, 688
Nd^{3+}:glass, 638–640, 1095
noise, 651–653
nonlinearity, 645–651
optical fiber, 640–642
phase-shift coefficient, 625
photon statistics, 653, 656
population inversion, 582, 622–629
postamplifier, 636
power amplifier, 636
preamplifier, 636
pumping, 626–635
quasi-three-level pumping, 633
quasi-two-level pumping, 634
rare-earth-doped fiber, 640
rate equations, 626–629
rates and decay times, 626
ruby, 636–637
saturation, 645–651
saturation time constant, 629
spontaneous lifetime, 644
steady-state, 626
theory, 622–625
three-level pumping, 632, 633
transition characteristics, 644
two-level pumping, 655
wavelengths, 644
Laser diodes (LDs), 831–844
bipolar, 851
broad-area, 834, 848, 849
buried-heterostructure, 848
communications component, 1232
compare with LEDs, 839, 841
compare with SLEDs, 839, 841
confinement factor, 833
differential responsivity, 837
distributed Bragg reflector, 842
distributed-feedback, 842, 848
double-heterostructure, 825, 845
efficiency, 836
external differential efficiency, 836
external-cavity, 843
extraction efficiency, 838
far-field radiation pattern, 842
gain condition, 834
gain-guided, 834
Ge, 772
GeSn-on-Si, 860
group-IV, 859
III–antimonide, 854
III–V quantum-dot-on-Si, 859
in-band pumping, 817
index-guided, 834
interband, 851
interband cascade, 854
IV–VI, 854
lead-salt, 854
light–current curve, 837, 839
linewidth, 843
linewidth-enhancement factor, 843
mode-locked, 844
multimode MQW, 848
PbSnSe, 854
PbSnTe, 854
phase noise, 843
power output, 836
power-conversion efficiency, 838
ridge-waveguide, 847
Schawlow–Townes linewidth, 843
single-mode, 842
single-mode MQW, 847
slope efficiency, 837
spatial characteristics, 841
spectral characteristics, 839
threshold, 834
VCSEL, 856–858
VECSEL, 721, 859
wall-plug efficiency, 838
wavelength-tunable, 843
Layered media, 258–276
off-axis wave, 264
Lens, 16–18
aspheric, 17
biconcave, 17
biconvex, 17
collimating, 95
compound, 17
converging, 17
cylindrical, 17, 24, 40, 117, 118
diverging, 17
dome, 799, 814, 815
double-convex, 58
electro-optic, 976
expanding, 95
focal length, 16
focal point, 33
focusing, 58, 93, 109
Fourier transform, 126, 128, 158, 513, 549
Fresnel, 18
Fresnel zone plate, 67, 118
graded-index, 24, 60
hyperlens, 319

imaging, 58, 137, 139, 141, 158, 500
Kerr, 719, 1039
LED, 799
meniscus, 17
perfect, 315–317
plano-concave, 17
plano-convex, 17, 58
principal point, 33
relaying, 94
sequence, 36, 37
shaping, 93
spherical, 16
superlens, 316
thick, 33
thin, 57, 91
time, 1113, 1162
vertex point, 33
Lidar, 789
coherent, 75
PIC, 1238
Light
classical, 4, 539
guide, 18
interaction with atoms, 583–602
interaction with semiconductors, 766
line, 285
nonclassical, 4, 539, 544, 546
Light-emitting diodes (LEDs), 789–817
additive color mixing, 814–815
AMOLED, 1004
arrays, 815
bioinspired, 798
Ce^{3+}:YAG phosphor, 814
characteristics, 794–803
chip-on-board (COB), 815
color rendering index (CRI), 812
communications component, 1233
compare with incandescent, 812, 816
compare with LDs, 839, 841
compare with SLEDs, 839, 841
complementary colors, 813
correlated color temperature, 812
device structures, 803–817
die geometries, 798
discrete, 813
edge-emitting, 804
electronic circuitry, 803, 816
external efficiency, 800
extraction efficiency, 796, 869
illumination applications, 788, 811
indication applications, 788
infrared applications, 806
internal efficiency, 795, 800
light–current curve, 801, 839
lighting, 811–817
materials, 803–817
optics for, 15, 799
organic, 810, 816
output photon flux, 799
overall efficiency, 800
phosphor-conversion, 814
photonic-crystal, 798
plasmonic, 796
power-conversion efficiency, 800
quantum-dot, 807, 814
resonant-cavity, 800
response time, 802
responsivity, 801
retrofit lamps, 816
roughened-surface, 798
solid-state lighting, 811–817
spatial pattern, 799
spectral distribution, 802, 868
surface-emitting, 804
surface-mounted device, 814
trapping of light, 19, 39
ultraviolet applications, 808
visible applications, 807
wall-plug efficiency, 800
white, 813–815
WOLED, 810
Line broadening, 593–597
collision, 595
Doppler, 597
homogeneous, 572, 595
inhomogeneous, 572, 595
lifetime, 593
Linear system
causal, 1298
Hilbert transform, 1298
impulse response function, 1299
isoplanatic, 1299
Kramers–Kronig relations, 1298
modes, 1301–1305
one-dimensional, 1296–1299
point-spread function, 1299
shift-invariant, 1296, 1299
supermodes, 383
time-invariant, 1296
transfer function, 1297, 1300
two-dimensional, 1299–1300
Lineshape function, 586
Linewidth
-enhancement factor, 843
laser, 680
laser diode, 843
Schawlow–Townes, 680, 843
transition, 586, 644

Liquid crystal
cholesteric, 244
electro-optics of, 996–1005
ferroelectric, 1001
modulator, 997–1001
nematic, 244, 996
optics of, 244–246
parameters, 999
smectic, 244, 1001
switch, photonic, 1194
twisted nematic, 244–246, 999
wave retarder, 997–1001
Liquid-crystal display (LCD), 1002
active matrix, 1003
passive-matrix, 1002
segmented, 1002
Lithography
electron-beam, 348, 468, 696
EUV, 696, 702
focused ion-beam, 348, 696
holographic, 298
micro-, 298
multiphoton, 299, 611
X-ray, 696
LLNL, 639, 688, 698, 699, 849, 1095
Localized surface plasmon, 330–332
LED, 796
nanolaser, 862
resonance, 330
Logic
photonic logic gates, 1211
Lorentz
oscillator, 186, 1018, 1072
relativistic factor, 703
Lorentzian, 189, 483, 512, 593–595, 624, 625, 647, 730, 1080, 1292
Luminescence, 607–612
betaluminescence, 607
bioluminescence, 608
cathodoluminescence, 607, 618
chemiluminescence, 608
electroluminescence, 608
fluorescence, 609, 811
multiphoton fluorescence, 610
phosphorescence, 609, 811
photoluminescence, 609–612, 813
radioluminescence, 609
sonoluminescence, 607
up-conversion fluorescence, 611
Luminous
efficacy, 812, 817
flux, 812, 817

Magnetic
dipole moment, 178
field, 162
flux density, 163
Magnetization density, 164
Magneto-optics, 242–243
Faraday effect, 242, 1013
magnetogyration coefficient, 243
switch, photonic, 1195
Verdet constant, 243
Magnification
Gaussian beam, 92
shift-variant, 144
spherical boundary, 15
spherical lens, 17
spherical mirror, 11
time lens, 1162
Maiman, Theodore H., 657
Manley–Rowe relations, 1029, 1051, 1063
Maser, 659, 680
astrophysical, 694
Master-oscillator fiber-amplifier, 688
examples, 689
Master-oscillator power-amplifier
examples, 636, 639, 688, 689
Material
double-negative (DNG), 306
double-positive (DPS), 306
equation, 173, 228, 241, 243
hyperbolic, 317–320, 352
left-handed, 306, 309
negative-index (NIM), 309, 314–317
single-negative (SNG), 306
Matrix optics, 27–37
$ABCD$ matrix, 28, 1302
arbitrary paraxial system, 31
Bragg grating, 271
layered media, 258–276, 1303
periodic media, 280–286
periodic systems, 33–37, 447–450
ray-transfer matrix, 27, 1302
scattering matrix, 259–265
scattering vs. wave-transfer, 260
simple components, 29
thick lens, 32
transmission matrix, 380, 383
wave-transfer matrix, 258–265, 1303
Maurer, Robert D., 391
Maxwell's equations
boundary conditions, 164, 221
complex permeability, 309–310
complex permittivity, 309–310
dielectric constant, 167
electric field, 162
electric flux density, 163

impermeability tensor, 230
in a medium, 163, 172
in conductive medium, 320–326
in free space, 162
magnetic field, 162
magnetic flux density, 163
magnetization density, 164
permeability, 163, 167, 306, 343
permeability tensor, 343
permittivity, 163, 167, 306
permittivity tensor, 170, 228, 343
polarization density, 164, 1017
relative permittivity, 167
speed of light, 163, 167
vector potential, 177
wave equation, 163, 167, 169, 171
Maxwell, James Clerk, 160
Media, *see* Materials
Metals
band structure, 577
bound-electron absorption, 324
conductive media, 320–326
conductivity, 577
Drude model, 322–326
group velocity, 352
loss, 470
optics of, 304, 320–326
photoemission, 873
plasma frequency, 698
plasmonics, 320–334
reflectance, 324
work function, 873
Metamaterials
holey metallic film, 342
hyperbolic, 339–340, 352
invisibility cloak, 348
metasurfaces, 340–343
negative-index, 338–339
negative-permeability, 337–338
negative-permittivity, 336–337
optical fiber, 428
optics of, 304, 334–343
photonic-crystal, 256, 334, 428
point-dipole approximation, 335
Metameric white light, 813
Metasurfaces, 340–343
complementary, 340
holey metallic film, 342
hologram, 342
phase modulator, 341
reflection, 342
refraction, 342
Snell's law, modified, 342
Michelson stellar interferometer, 505
Micro-optics, 305, 1168
Microcavity, *see* Microresonator
Microcavity lasers, 854–862
Microresonator
microcavity, 463
microdisk, 465
micropillar, 465
microsphere, 466–468
microtoroid, 465
modal density, 465
modal volume, 464
photonic-crystal, 468
quality factor, 464
rectangular, 464–465
Microscopy
multiphoton, 610, 611
near-field, 147
three-photon, 611
two-photon, 610
Mie scattering, 197
Miniband, 579, 764, 779, 852
QCL, 852
Mirror
collimator, 9
elliptical, 9
paraboloidal, 8, 10
planar, 8, 53, 78
spherical, 9, 30, 78, 96
variable-reflectance spherical, 97
MIT Lincoln Laboratory, 787
Mixing, optical, 74, 967, 1153, 1267
Mode locking, 76, 709, 716–723
applications, 721
examples, 685–690, 696, 720
external-cavity LDs, 844
fiber lasers, 844
harmonic, 721
Kerr-lens, 719, 1039
methods, 719
optical frequency comb, 721
parameters, 718
properties, 716, 718, 730
QCLs, 721, 854
QD lasers, 851
saturable absorber, 719
SESAM, 719
VECSELs, 721, 859
Modes
discrete linear system, 1302
eigenfunction, 1301
eigenvalue, 1301
eigenvector, 1301
homogeneous medium, 1305
integral operator, 1303

linear system, 1301–1305
normal, 1304
ordinary differential equation, 1304
partial differential equation, 1304
periodic medium, 1305
resonator, 1304
supermodes, 383
zero-point energy, 518, 573, 586
Modulation
amplitude shift keying, 1259
binary phase shift keying, 1259, 1273
constellation, 1259
differential phase shift keying, 1259
digital, 1258
field, 1257, 1266
frequency shift keying, 1259
intensity, 1258
multilevel coding, 1259
on–off keying, 931, 1259, 1272
phase shift keying, 1259
pulse code, 1258
quadrature amplitude, 1259
quaternary phase shift keying, 1259
spectral efficiency, 1259
Modulator
acousto-optic, 959
electroabsorption, 1010, 1011
intensity, 998
intensity, acousto-optic, 973
intensity, electro-optic, 981–983, 995
intensity, magneto-optic, 1013
intensity, push–pull, 1014
interferometric, 981
liquid-crystal, 997–1001
Mach–Zehnder, 981, 1014
multiquantum-well, 1010
optically addressed SLM, 1004
parallel-aligned SLM, 1005
phase, 998
phase, acousto-optic, 973
phase, cascaded, 1014
phase, electro-optic, 973, 979, 995
phase, opto-optic, 1076
quadratic phase, 1100
spatial light, 987–989
spatial light, acousto-optic, 965
spatial light, liquid-crystal, 1002
Molecules, 572–575
covalent bonding, 572
dye, 575
ionic bonding, 572
rotating diatomic, 573
van der Waals bonding, 572
vibrating diatomic, 573
vibrating triatomic, 574
Momentum
electromagnetic, 165
localized photon, 523
localized wave, 523
photon, 523–524, 554, 555
radiation pressure, 524, 555
Momentum, angular
photon orbital, 524
photon spin, 524
Mourou, Gérard, 1078
Multiphoton
absorption, 1017
detection, 939, 1156, 1162
fluorescence, 610
lithography, 298, 299, 611
microscopy, 610, 611
photoluminescence, 610
Multiple access
code-division, 1276
frequency-division, 1276
time-division, 1276
Multiplexing
code-division, 1261
CWDM, 1263
DWDM, 1263
electronic, 1261
frequency-division, 1260
optical, 1261
space-division, 1263–1266
time-division, 1208, 1260
wavelength-division, 1179, 1262
Multiquantum
-dot lasers, 850
-well lasers, 845
-wire lasers, 849
well, 764, 1010

Nano-optics, *see* Nanophotonics
Nanocavity lasers, 862–864
Nanophotonics, 193–199, 305, 326–343, 386–388, 428, 469–470
nanolasers, 862
subwavelength imaging, 147, 315
Nanoresonator, 469–470
metallic nanodisk, 469
metallic nanosphere, 470
Nanosphere
dielectric, 196–197
metallic, 330–332
scattering from, 196–197, 330–332
Near-field imaging, 146, 334
Negative-index
materials, 314

metamaterials, 338–339
Network, fiber-optic, 1274–1281
broadcast-and-select, 1277
bus, 1275
interface, 1276
local-area (LAN), 1274
mesh, 1275
multi-hop broadcast-and-select, 1278
ring, 1275
star, 1275
topologies, 1275
wavelength-routed, 1279
WDM, 1277
Newton, Sir Isaac, 3
Nobel laureates, 110, 160, 619, 657, 731, 867, 943, 1015, 1224
Nobel lectures, 157, 431, 551, 615, 616, 726, 727, 782, 783, 867, 868, 938, 972, 1075
Noise
1/*f*, 923
ASE, 651, 830
background, 911
circuit, 910, 922–930
dark-current, 911
gain, 910, 915
generation–recombination (GR), 940
laser phase, 843
optical amplifier, 651
photoconductor, 940
photocurrent, 912
photodetector, 909–923
photoelectron, 539, 910, 912
photon, 533–539, 910, 911
pink, 923
semiconductor optical amplifier, 821
shot, 518, 703, 912, 913
superluminescent diode, 830
thermal, 922
Nondiffracting beams, 105–107
Nondiffracting waves, 136
Nonlinear optical coefficients, 1018, 1019, 1025, 1037, 1066, 1068
Nonlinear optics, 171, 1015–1077
anharmonic oscillator, 1071–1073
anisotropic dispersive medium, 1073
anisotropic medium, 1066–1069, 1073
Born approximation, 1020, 1076
chi-two medium, 1021
coherence length, wave-mixing, 1032
coupled waves, 1047
coupled-waves, 1059, 1127
cross-phase modulation, 1041
DFG, 854, 1027
differential-equation description, 1071
dispersive medium, 1069–1074
downconversion, 1027, 1077
electro-optic effect, 1024
extreme, 1141
five-wave mixing, 1077
four-wave mixing, 1042, 1044, 1059, 1074, 1077, 1140
frequency conversion, 1027, 1054
HHG, 723, 1141
holography, real-time, 1045
idler, 1027
integral-transform description, 1070
introduction to, 1016–1017
Kerr effect, optical, 1037, 1042
Kerr medium, 1036
Manley–Rowe, 1029, 1051, 1076
Miller's rule, 1073
nonlinear coefficients, 1018, 1019, 1025, 1037, 1066, 1068
nonparametric, 1017, 1028
parametric, 1017
parametric interactions, 1026–1029
periodic poling, 1035, 1036
phase conjugation, 1044, 1063
phase matching, 1026, 1029–1033
phase-mismatching tolerance, 1076
photonic-crystal soliton, 1141
plane-wave conjugation, 1045
polarization density, 1017–1019
poling, 1036
pump, 1027
quasi-phase matching, 1034, 1076
Raman gain, 1041
rectification, optical, 1023
rectification, pulsed optical, 1128
refraction, nonlinear, 1038
scattering theory, 1020, 1076
Schrödinger equation, nonlinear, 1040, 1135, 1138
second-order, 1021–1036, 1047–1059
self-focusing, 1039
SFG, 1027
SFG and SHG combined, 1077
SHG, 1021, 1027, 1049, 1051
SHG efficiency, 1022, 1052
SHG phase mismatch, 1053
signal, 1027
solitary wave, 1131
soliton self-frequency shift, 1140
soliton, spatial, 1039
soliton, spatiotemporal, 1139
soliton, temporal, 425, 1130–1139
SPDC, 548, 1027

spherical-wave conjugation, 1045
SPM, 425, 1038, 1129, 1130, 1140
supercontinuum generation, 1139
THG, 1037, 1063
third-order, 1036–1047, 1059–1066
three-wave mixing, 545, 1025, 1026, 1043, 1050, 1061, 1066, 1073, 1076, 1077, 1127, 1162
THz pulse generation, 1128
tuning curves, 1029
two-wave mixing, 1009, 1026, 1042
ultrafast, 1126–1145
up-conversion, 1054, 1055
Volterra-series expansion, 1070
walk-off effect, 1126
wave equation, 1019, 1047, 1133
wave restoration, 1046
Nonlinear-optic devices
amplifier, parametric, 1027, 1056
amplifier, phase-sensitive, 1062
amplifier, reflection, 1065
amplifier, transmission, 1065
DFG, 1027
downconverter, 1027, 1077
FPI, 1217, 1218
intensity autocorrelator, 1150, 1162
lens, optical Kerr, 1039
loop mirror, 1186
modulator, opto-optic phase, 1076
MZI, 1216, 1218
oscillator, doubly resonant, 1058
oscillator, parametric, 1027, 1057
oscillator, phase-conjugation, 1066
oscillator, singly resonant, 1058
router, directional-coupler, 1186
router, nonlinear MZI, 1185
router, nonlinear Sagnac, 1185
router, soliton, 1187
SFG, 1027
streak camera, 1149
switch, FWM, 1199
switch, optical Kerr, 1148, 1198
switch, photonic, 1196–1211
switch, Sagnac, 1201
switch, SFG, 1197
switch, SHG, 1147
switch, XPM, 1197–1199
up-converter, 1027
Nonparametric processes, 1028
Normal modes
anisotropic medium, 230, 232–233
optically active medium, 241
polarization system, 220, 221
Numerical aperture, 19, 25, 39, 181, 366, 395, 432

Ohm's law, 320, 882
Optical
frequency comb, 721, 1141
Kerr effect, 1037, 1042
lattices, 601
molasses, 599
OADM, 1179, 1180, 1279
pathlength, 5, 6, 53, 78
phase conjugation, 1044, 1063
sectioning, 492, 550
tweezers, 600
Optical activity, 240–242
gyration vector, 241
normal modes, 241
rotatory power, 242
Optical coherence, *see* Coherence
Optical coherence tomography, 492
frequency-domain, 493
quantum, 550
time-domain, 492
Optical components, 8–20, 53–61
active-matrix LCD, 1003
active-matrix OLED, 1004
antireflection coatings, 264, 301
axicons, 14, 57, 106, 137
backlight, 1002
beam combiners, 13
beam directors, 14
beamsplitters, 13, 63, 247, 263, 268, 301, 522, 539, 549
catadioptric, 15
circulator, 1173
collimators, 9, 15, 799
diffraction gratings, 59, 78
electro-optic prism, 983
fiber couplers, 413
fibers, 392, 422–423, 426, 429
filter, acousto-optic, 966
frequency shifter, acousto-optic, 967
graded-index, 22–24, 40, 60, 396
integrated, 354, 808
isolator, 250, 967, 1167
LED optics, 15, 799
lenses, 16–18, 57, 58, 319
mirrors, 9, 10, 53
modulators, acousto-optic, 959
passive-matrix LCD, 1002
photonic lantern, 414
plates, 39, 55
polarizers, 217, 253, 520
prisms, 13, 14, 56, 57

resonators, 434, 447, 459, 463, 516
scanners, acousto-optic, 961
segmented LCD, 1002
space switches, acousto-optic, 964
spatial light modulator, 965, 1002
spectrum analyzer, AO, 963
spiral phase plate, 104
superprism, 1180
thin films, 264, 301
waveguide couplers, 376–383
waveguides, 355, 363, 372, 386
Optical fiber amplifier, 640–642
communications component, 1234
compare with SOA, 830
Optical fiber communications, 1224–1286
analog, 1253
analog coherent, 1271
attenuation, 1246, 1247, 1252
attenuation compensation, 1253
balanced homodyne receiver, 1268
balanced mixer, 1268
bit error rate, 1245
coherent, 1266–1274
coherent receiver advantages, 1269
components, 1226–1238
direct vs. heterodyne, 1274
direct vs. homodyne, 1274
dispersion, 1246, 1249, 1252
dispersion compensation, 1254, 1255
dispersion management, 1255
evolution, 1240–1243
eye diagram, 1245
fibers, 1226–1231
heterodyne receiver, 1267
heterodyne vs. direct, 1274
homodyne BPSK, 1273
homodyne OOK, 1272
homodyne QPSK, 1273
homodyne vs. direct, 1274
homodyne vs. heterodyne, 1274
introduction to, 1225–1226
local oscillator, 1267
modulation, 1257–1260
multiplexing, 1260–1263
networks, 1274–1281
optical amplifier, 1234
performance, 1243–1246
photodetector, 1235–1237
photonic integrated circuit, 1238
power budget, 1247
receiver sensitivity, 1246
soliton, 1256
SONET standard, 1239, 1276
sources, 1232–1234
systems, 1238–1257
time budget, 1249
Optical indicatrix, *see* Index ellipsoid
Optical materials
2D, 745, 780
fused silica, 19, 186, 191, 204, 205, 207, 394, 416, 417, 419, 641, 961, 1110
GaAs, 207
glass, BK7, 1098, 1105, 1107, 1116
glass, phosphate, 639, 643
glass, phosphosilicate, 643, 692
host glass, 640
$LiNbO_3$, 254, 375, 973, 980, 992, 995, 1006, 1036
$LiTaO_3$, 992, 995
periodically poled, 1035, 1036
quartz, 240, 242, 254, 1015
soft glasses, 429
TeO_2, 958
TMDs, 745, 780
Optical receiver
analog receiver sensitivity, 929–930
bipolar-transistor amplifier, 925
bit error rate, 911, 932
circuit-noise parameter, 923–925
digital receiver sensitivity, 931–934
FET amplifier, 925
on–off keying (OOK), 931
resistance-limited, 924
sensitivity, 911
signal-to-noise ratio, 925–929
SNR dependence on APD gain, 928
SNR dependence on bandwidth, 928
SNR dependence on photon flux, 926
Optical system, periodic, 33–37, 447–450
GRIN plate, 40
harmonic trajectory, 35
lens sequence, 36
lens-pair sequence, 37
periodic trajectory, 35
ray position, 33
resonator, 37, 40
Organic semiconductors, 742, 876
Oscillator
harmonic, 170, 323, 542, 543, 573, 1072, 1298
Lorentz, 186, 1018, 1072
optical parametric, 1027, 1057
phase-conjugation, 1066

Paraboloidal
approximation, 49
mirror, 8

surface, 86
wave, 49, 51, 81
Parametric
amplifier, 1027, 1056
downconverter, 1027
oscillator, 1027, 1057
oscillator, doubly resonant, 1058
oscillator, singly resonant, 1058
processes, 1017
switches, 1197
Paraxial
approximation, 9, 27, 50
Helmholtz equation, 51, 78
imaging system, 31
optics, 5, 9, 27–37
ray equation, 21, 22, 40
rays, 9
system, focal length, 31
system, focal point, 31
system, principal point, 31
system, vertex point, 31
wave, 50, 1086
wave equation, generalized, 1124
Parseval's theorem, 914, 1288
Partially coherent light, *see* Coherence
Pauli exclusion principle, 565, 582, 734
Pendry, Sir John, 303
Penetration depth, 308, 312
Periodic media
2D periodic structure, 292
3D periodic structure, 294
Fourier optics, 286–289
matrix optics, 280–281
Periodic table
elements, 566, 567
semiconductors, 737
Permeability, magnetic, 163, 167, 343
complex, 306, 309–310
negative, 306
tensor, 343
Permittivity, electric, 163, 167, 306
complex, 183, 198, 306, 309–310
effect of electric field, 990
effective, 198, 320
frequency-dependent, 184
in magneto-optic material, 243
negative, 306
relative, 167
tensor, 170, 228, 343
Perot, Alfred, 433
Phase
-mismatching tolerance, 1076
-sensitive amplifier, 1062
-shift coefficient, 660
amplifier, 70
matching, 288, 1026, 1029–1145
noise, 680, 843
spiral, 103, 155
velocity, 48, 200, 284, 352
Phosphorescence, 609, 811
Photoconductors, 875, 883–886
detection circuit, 939
doped extrinsic, 886
extrinsic, 885
gain, 884
intrinsic, 883
noise, 940
response time, 885
spectral response, 885
Photodetectors, 871–942
array detectors, 907, 1149
array readout circuitry, 908
avalanche photodiodes, 895–907
bolometer, 872
CCD readout circuitry, 908
charge-coupled device (CCD), 908
circuit noise, 922–930
CMOS readout circuitry, 909
communications component, 1235
digital photon-counting device, 906
electron-multiplying CCD, 909
external photoeffect, 873
extrinsic photoconductive, 886
focal-plane array (FPA), 907
FROG, 1156
gain, 879
gain noise, 915–921
general properties, 876–883
Golay cell, 872
intensified CCD (ICCD), 909
intensity, pulse, 1146–1151
intensity-autocorrelation, 1150
internal photoeffect, 873
introduction to, 872
microbolometer, 872, 907
microchannel plate, 875
minimum-detectable signal, 911
negative-electron-affinity, 874
noise, 909–923
noise-equivalent power, 911
optical-pulse, 1146–1158
optical-pulse phase, 1152–1156
organic, 876
performance measures, 910
photoconductors, 875, 883–886
photodiodes, 887–894
photoelectric, 872
photoelectric emission, 873–875

photoemission equation, 873
photomultiplier, 874, 904
photon detection efficiency, 876
photon-number-resolving, 904, 905
phototube, 874
plasmonic, 878
pyroelectric, 872
QDIP, 886
quantum efficiency, 876
QWIP, 886
Ramo's theorem, 880
RC time constant, 883
reach-through APD, 899
resonant-cavity, 878
response time, 880
responsivity, 878
SACM APD, 899, 902
SAM APD, 899, 901, 902
secondary emission, 874
signal-to-noise ratio, 910, 913, 916
silicon photomultiplier (SiPM), 905
single-photon, 904
specific detectivity, 911
spectral intensity, pulse, 1151–1152
spectrogram, pulse, 1156–1158
streak camera, 1149
superconducting, 907
thermal, 872
thermocouple, 872
thermopile, 872
transit-time spread, 880, 939
transition-edge sensor (TES), 907
two-photon, 939, 1156, 1162
vacuum photodiode, 874
Photodiodes, 887–894
p–i–n junction, 889
p–n junction, 887
avalanche, 895–907
comparison with APDs, 927
depletion layer, 887
edge-illuminated, 890, 892
evanescently coupled, 890
group-IV, 892, 894
heterostructure, 891
metal–semiconductor, 893
modes of operation, 888
open-circuit operation, 888
photon-trapping microstructures, 891
photovoltaic operation, 888
response time, 888
reverse-biased operation, 889
Schottky-barrier, 893, 894
short-circuit operation, 888
solar cell, multi-junction, 892
traveling-wave configuration, 890
Photoelastic
effect, 968
tensor, 969
Photoluminescence, 608–612, 813
applications, 610
multiphoton, 610
quantum dots, 580
Photometry
illuminance, 812
luminous efficacy, 812
luminous flux, 812
photopic luminosity function, 812
Photon, 516–529
antibunching, 780
at beamsplitter, 522
boson, 524, 537
energy, 518
Gaussian-wavepacket, 527
helicity, 524
in Fabry–Perot resonator, 555
in Gaussian beam, 522
in Mach–Zehnder, 526
in Young interferometer, 525
interference, 525
lifetime, 662
momentum, 523–524, 554, 555
monochromatic, 527
optics, 514–541
orbital angular momentum, 104, 524
polarization, 519
polychromatic, 527, 555
position, 521
position and time, 527
spin angular momentum, 524
time, 526
transmission through polarizer, 520
wavepacket, 527
Photon detectors, *see* Photodetectors
Photon stream, 529–541
number of photons, 531, 532
partitioned, 539–541
photon flux, 530, 532
photon-flux density, 530, 532
photon-number statistics, 533–541
randomness, 532
spectral density, 531
Photon-number statistics, 533–541
Bernoulli, 539
Bernoulli trial, 912
binomial, 540, 541, 546, 557
Bose–Einstein, 536, 537, 539, 556
coherent light, 533
counting time, 533

doubly stochastic, 538, 556
exponential density function, 539
geometric, 537
Mandel's formula, 538
mean, 533, 535
mean under absorption, 556
negative-binomial, 556
noncentral-chi-square density, 656
noncentral-negative-binomial, 653
partitioned distribution, 540, 912
partitioned SNR, 541, 912
photon number, 533
photon-number distribution, 533
Poisson, 534, 544, 555
random partitioning, 539, 556
random selection, 539, 912
signal-to-noise ratio, 535, 911
sub-Poisson, 546
thermal light, 536
uniform, 557
variance, 535
Photonic
bandgap, 282, 285, 296
integrated circuits, 354, 809, 1176, 1238
lantern, 414
Photonic crystal, 277–302
2D, 292–294
3D, 294–299
band structure, 282, 285, 296
bandgap, 282, 285, 296
Bloch modes, 277, 1305
dispersion relation, 285
fabrication, 297, 299
fibers, 426–428
group velocity, 284
holes and poles, 298
holes on diamond lattice, 297
inverse-opal, 297, 298
laser, 860
laser array, 861
lattice defects, 298
metal–dielectric array, 388
metamaterial, 256, 334
omnidirectional reflection, 290
one-dimensional, 277–291
optics, 255–302
optics, introduction to, 256–258
phase velocity, 284
point defects, 298
projected dispersion diagram, 285
silicon, 297
soliton generation, 1141
switch, photonic, 1200
thresholdless laser, 861
waveguide, 385–386
woodpile, 297
Yablonovite, 297
Photorefractivity, 1005–1009
applications, 1009
real-time holography, 1009
simplified theory, 1006–1009
Planck
Planck's constant, 518
spectrum, 605
Planck, Max, 515, 516, 529
Plane wave, 47, 111, 113
acoustic, 946
conjugate, 1045
electromagnetic TEM, 175, 307
partially coherent, 487
pulsed, 1086
quasi-plane wave, 407
reflection, 221
refraction, 221, 238
wavefronts, 48
wavefunction, 48
Plasma
frequency, 322
wavelength, 322
Plasmonics, 320–334, 387
LEDs, 796
nanolasers, 862
photodetectors, 878
resonator, 469–470
switch, photonic, 1201
waveguide, 386–388
Pockels
coefficient, 978, 1025, 1077
effect, 978, 991–994, 1024
readout modulator (PROM), 988
Pockels, Friedrich, 975
Poincaré sphere, 213, 508
Point-spread function, 1299
Poisson, Siméon Denis, 871
Polarization, 211–221
p, 222
s, 222
-maintaining fiber, 406
autocorrelation, 506
circular, 213, 253, 509, 524
coherency matrix, 507
complex envelope, 211, 214
complex polarization ratio, 213
coordinate transformation, 219
cross-correlation, 506
degree, 509
electric-field vector, 211

ellipse, 211–212
Jones matrix, 217–221, 509
Jones vector, 215–221, 507
linear, 212, 508, 524
matrix representation, 215–221, 1302
mode dispersion, 423
normal modes, 220, 221
optics, 209–254
optics, introduction to, 210–211
orthogonal, 222, 252
orthogonal expansion, 216
orthogonal Jones vectors, 215
parallel, 222
partial, 506–510
photon, 519
Poincaré sphere, 213, 508
rotator, 219, 245, 249, 252, 253, 549
Stokes parameters, 213, 507
transverse-electric, 222, 223
transverse-magnetic, 222, 224
two-photon, 547
unpolarized light, 508
Polarization devices
rotator, 252
Polarization density, 164, 1017
Polarization devices, 247–251
anti-glare screen, 253
cascade, 219, 253, 254
circulator, 251
coordinate transformation, 219
dichroic, 247
Faraday rotator, 250, 1173
fast and slow axes, 218, 248, 253
half-wave retarder, 219, 220, 253
intensity control, 249
isolator, 250
linear-polarizer cascade, 254
nonlinear Kerr switch, 1148, 1198
nonreciprocal, 250–251
normal modes, 220
polarizer, 217, 247–248, 253
polarizing beamsplitter, 248
quarter-wave retarder, 218
rotator, 219, 249, 549
router, 1184
wave retarder, 218, 248, 253
wave-retarder cascade, 219, 253
Polychromatic light, 71–76
intensity, 72
Ponderomotive force, 703, 1142
Postulates
Born postulate, 563
Einstein postulates, 591
ray optics, 5–8
wave optics, 43–44
Power
electromagnetic, 165, 172
Gaussian beam, 83
optical, 44
scattered, 194, 195
spectral density, 479
Poynting
theorem, 168
vector, 165, 173, 312
Principal
axes, 229
refractive indices, 229
Principal point
SELFOC lens, 24
arbitrary paraxial system, 31
Prism, 13, 14, 56
chirp filter, 1097, 1105, 1161
coupler, 329, 352, 377
dispersion compensation, 685
electro-optic, 983
laser-line selector, 677
spatial, 116
superprism, 1180
Prokhorov, Aleksandr M., 619
Propagation
along principal axis, 230–231
anisotropic medium, 232–233
free space, 113–123
gain of spatial coherence, 502
homogeneous medium, 6
partially coherent light, 497–498
van Cittert–Zernike theorem, 503
Pulse, optical
attosecond, 1141
characteristics, 1079–1083
chirp filtering of, 1088–1099
chirp parameter, 1082
chirped, 1082
chirped-Gaussian, 1084–1085
complex envelope, 73, 1080
compression, 1099, 1109, 1130, 1144
detection, 1146–1158
dispersion length, 1106
down-chirped, 1082
Fourier-transform-limited, 1084
frequency-to-space mapping, 1101
FROG, 1156
Gaussian, 1080, 1084–1085, 1161
Gaussian beam, 1087, 1120
Gaussian, bandwidth-limited, 1084
in dispersive media, 199–203
instantaneous frequency, 1081
intensity autocorrelation, 1150, 1162

intensity detection of, 1146–1151
light bullet, 1139
linear filtering, 1088
nonseparability, 1117, 1125
phase detection of, 1152–1156
plane wave, 73, 1086
propagation in fiber, 1102–1115
rectification, 1128
sech, 1040, 1080, 1133, 1136, 1161
shaping, 1101–1102
slowly varying, 1086
soliton, 1133
soliton self-frequency shift, 1140
soliton, spatiotemporal, 1139
soliton, temporal, 1130–1139
spatial characteristics, 1086–1088
spectral intensity, 1080
spectral intensity detection, 1151
spectral phase, 1080
spectral shift, 1117
spectral width, 1080
spectrogram, 1083, 1156–1158
spherical wave, 78, 1086
SPM, 1129, 1130, 1140
spread, 201
streak-camera detection, 1149
sub-femtosecond, 1141
supercontinuum light, 1139
temporal broadening, 1116, 1117
temporal width, 1080
three-wave mixing, 1127, 1162
THz pulse generation, 1128
time-to-space mapping, 1102
time-varying spectrum, 1083
transform-limited Gaussian, 1084
up-chirped, 1082
wavepacket, 73
Pupil function, 138, 141, 500
generalized, 143
Purcell factor, 598, 796, 800, 855, 861

Quadric representation, 229, 230
Quantum
circuits, 550
cutting, 610
defect, 634, 690
electrodynamics, 4, 515
entanglement, 547
mechanics, 562
optics, 4, 515, 541–550
Quantum cascade laser (QCL), 851–854
heterogeneous, 852
single-frequency, 853
strained-layer, 853
tunable, 853
Quantum dot
infrared photodetector (QDIP), 886
Quantum dots, 579–580, 607, 766
applications, 580
artificial atoms, 580
core–shell, 580
excitons, 580, 850
fabrication, 579
lasers, 850–851
LEDs, 807, 814
mode-locked lasers, 851
photoluminescence, 580
self-assembly, 580, 807, 850
silicon photonics, 780, 808
single-photon emitter, 546, 780
SOAs, 829
synthesis, 579
Quantum number
azimuthal, 564
magnetic, 564
orbital angular momentum, 566
overall angular momentum, 566
principal, 564
rotational, 573
spin, 564, 566
vibrational, 573
Quantum state, 541–550
amplitude-squeezed, 545
binomial, 546, 557
biphoton, 548
coherent, 544
entangled, 547
Fock, 546
intensity-squeezed, 546
number, 546
phase-squeezed, 545
photon-number-squeezed, 546
quadrature-squeezed, 544
sub-Poisson, 546
thermal, 536
twin-beam, 548
two-photon, 546
Quantum well, 578, 761–764
infrared photodetector (QWIP), 886
lasers, 845
SOAs, 826–830
Quantum wire, 579, 765
lasers, 849
Quantum-confined
excitons, 779
lasers, 844–854
structures, 760–766
Quasi-phase matching, 1034, 1076

Radar
Doppler, 75
laser, 789
Radiometry
irradiance, 812
radiant flux, 812
Raman
-Nath scattering, 957, 973
cascaded fiber laser, 692
cascaded silicon laser, 692
distributed fiber amplifier, 643
fiber amplifier, 642
fiber laser, 691
gain, 1041
lumped fiber amplifier, 643
scattering, 613
silicon laser, 692, 809
stimulated scattering, 613, 642, 691, 1140
Stokes shift, 643, 691
Ramo's theorem, 880, 939
Random light, *see* Statistical optics
Rate equations
amplifier radiation absent, 626
amplifier radiation present, 627
broadband radiation, 617
photon-number, 709
population-difference, 710
thermal light, 604
Ray equation, 20–21
paraxial, 21
Ray optics, 3–40
introduction to, 4–5
postulates, 5–8
relation to electromagnetic optics, 4
relation to quantum optics, 4
relation to wave optics, 4, 26, 42, 52
Ray-transfer matrix, 27–33
4×4, 40
SELFOC plate, 40
air followed by lens, 31
arbitrary paraxial system, 31
cascade of components, 30
determinant, 31
free-space, 29
inverse, 31
lens system, 40
parallel transparent plates, 30
planar boundary, 29
planar mirror, 29
skewed rays, 40
special forms, 28
spherical boundary, 29
spherical mirror, 30
spherical-mirror resonator, 447
thick lens, 32
thin lens, 29, 31
Rayleigh
-Jeans formula, 463, 606
inverse fourth-power law, 194, 417
range, 81
scattering, 193–197, 416, 612
Rayleigh, Lord (John William Strutt), 160
Reciprocal
lattice, 292
Rectification, optical, 1023
pulsed, 1128
Reflectance
at boundary, 222–225
between GaAs and air, 226
between glass and air, 226, 253
from a Bragg grating, 950
from a plate, 227
of a conductive medium, 321
power, 226
quarter-wave film, 264, 301
Reflection, 53, 221–227
at a dielectric boundary, 7
at a DPS-SNG boundary, 310
at a metasurface, 342
at a mirror, 7
at an absorbing boundary, 253
Bragg grating, 273
Brewster angle, 224, 225, 247
circularly polarized light, 253
critical angle, 12, 223, 225, 310, 327
external, 223, 224
frustrated total internal, 329, 378
internal, 223, 225
omnidirectional, 290, 302
phase shift, 224
total internal, 12, 13, 15, 18, 223, 253, 353, 392, 698
transverse-electric polarization, 223
transverse-magnetic polarization, 224
Refraction, 53, 221–227
all-angle, 320
at a dielectric boundary, 7, 15
at a hyperbolic medium, 319
at a metasurface, 342
at a spherical boundary, 14
at normal incidence, 346
Brewster angle, 253
conical, 254
double, 238–239, 254
external, 11
internal, 11
negative, 314–317

nonlinear, 1038
rays in anisotropic media, 239, 254
transverse-electric polarization, 223
transverse-magnetic polarization, 224
without reflection, 345
Refractive index, 5, 168, 183
air, 208
anisotropic media, 228–230
extraordinary, 229, 233, 236
frequency-dependent, 184
group, 201
negative, 308
optical materials, 184, 191
ordinary, 229, 233, 236
resonant medium, 189
Sellmeier equation, 191
Resolution
acousto-optic filter, 967
acousto-optic scanner, 961
electro-optic scanner, 983
liquid-crystal display, 1003
multiphoton lithography, 611
multiphoton microscopy, 611
optical-pulse detection, 1148
routing device, 1182
wavelength demultiplexer, 1184
Resonant medium, 186
anharmonic oscillator, 1071–1073
Lorentz oscillator, 186, 1018, 1072
Resonator
g-parameters, 448
axial modes, 456
bow-tie, 439–440
circular (2D), 460–461, 469
cold, 662
concentric, 449
conditionally stable, 448
confinement condition, 448–450, 453
confocal, 449
diffraction loss, 457, 458, 1304
energy per mode, 538
Fabry–Perot, 69, 265, 436, 661
fiber-ring, 434
finesse, 69, 266, 441
finite apertures, 457, 1304
free spectral range, 267, 438, 446
frequencies, 437, 439, 446, 460
guided-wave, 434, 471
integrated-optic-ring, 434, 692
losses, 440
microdisk, 464
micropillar, 464
microresonator, 463–468
microring, 1180
microsphere, 464
microtoroid, 464
modal density (1D), 440
modal density (2D), 460
modal density (3D), 462–463
modal volume, 435
modes, 1304
modes, standing wave, 437, 471
modes, traveling wave, 438
multiple microring, 1181
nanodisk, 469
nanoresonator, 469–470
nanosphere, 330–332, 470
number of modes, 472
optics, 433–472, 661
periodic optical system, 37, 40, 448
photonic modes, 469
photonic-crystal, 464, 468
planar-mirror, 436–446
plano–concave, 453
plasmonic, 330–334, 469
quality factor, 435, 445, 708
rectangular (2D), 459–460
rectangular (3D), 461–465, 517, 1305
ring, 434, 439–440
size vs. resonance wavelength, 435
spectral width, 440
spherical-mirror, 447–458
stability diagram, 449, 471
stable, 448
symmetric, 449
transmittance, 729
transverse modes, 456
traveling wave, 439
unstable, 448, 471
whispering-gallery modes, 461
Resonator, spherical-mirror, 447–458
axial modes, 456
free spectral range, 455, 456
Gaussian modes, 451
Hermite–Gaussian modes, 455
modes, 450–458
ray-transfer matrix, 37, 447, 448
resonance frequencies, 455, 456
symmetric, 449, 453, 471, 472
transverse modes, 456
Responsivity
avalanche photodiode, 898
differential, 837
LD, 839
LED, 801, 839
photodetector, 878, 1146
SLED, 839
Retarder, 218, 248

half-wave, 219, 220, 253
quarter-wave, 218, 254
quartz, 254
Ring
aperture, 67, 118, 159
benzene, 742
network, 1275
resonator, 434, 439–440
Rotator
Faraday, 250, 1173
nonreciprocal polarization, 251, 1173
polarization, 245, 249, 252
Router, passive optical, 1178–1187
add–drop multiplexer, 1179, 1279
arrayed waveguides, 1182, 1223
broadcast-and-select, 1179
demultiplexer, 1179
intensity-based, 1185
introduction to, 1164–1165
Mach–Zehnder interferometer, 1181
multipath interferometer, 1183
multiplexer, 1179
phase-based, 1185
polarization-based, 1184
waveguide grating, 1182
wavelength-based, 1179–1184
wavelength-division multiplexer, 1179
Russell, Philip St John, 1224

Saturable absorber, 649
bistable device, 1217
Scalar wave optics, *see* Wave optics
Scanner
acousto-optic, 961
electro-optic, 983
holographic, 117
Scattering, 612–614
and absorption, 198–199
and attenuation, 198, 416, 432
anti-Stokes, 613
Bragg diffraction, 953, 973
Brillouin, 613
CARS, 614
coefficient, 198
Debye–Sears, 957, 973
dielectric nanosphere, 196–197
efficiency, 196
elastic, 192–197
Huygens–Fresnel principle, 193
Maxwell-Garnett mixing rule, 199
metallic nanosphere, 330–332
Mie, 197, 613
quasi-static approximation, 196, 197
Raman, 613
Raman–Nath, 957, 973
Rayleigh, 193–197, 416, 432, 612
small scatterers, 193
stimulated Brillouin, 614, 644, 692
stimulated Raman, 613, 642, 691, 1140
Stokes, 613
strong, 196
volume fraction, 198
weak, 192–194
Scattering matrix, 259–265, 549
beamsplitter, 263
dielectric boundary, 262, 264
dielectric slab, 263, 265
homogeneous medium, 260
lossless medium, 261
lossless symmetric system, 262
relation to wave-transfer matrix, 260
Schawlow, Arthur, 657, 1015
Schrödinger equation
nonlinear, 1040, 1135, 1138
time-dependent, 562, 1304
time-independent, 542, 563, 1304
Schultz, Peter C., 391
Second-harmonic generation, 730, 1021, 1049, 1051
efficiency, 1022, 1052
phase mismatch, 1053
switch, 1147
Self-focusing, 1039
Self-phase modulation, 425, 1038
in supercontinuum generation, 1140
pulse, optical, 1129, 1130
Sellmeier equation, 191, 781
Semiconductor optical amplifiers, 817
bandwidth, 819, 869
compare with OFAs, 830
double-heterostructure, 825
gain, 818
gain coefficient, 820
heterostructures, 825
peak gain coefficient, 821, 869
pumping, 822
quantum-dot, 829
quantum-well, 826–830
SLEDs, 830
switch, photonic, 1193, 1196
waveguide, 829
Semiconductors, 577–580
k-selection rule, 770
p–i–n junction, 759, 889
p–n junction, 756, 887
p–n junction, biased, 757
absorption, 768, 777

AlGaAs, 806
AlGaN, 807
AlInGaN, 808
AlInGaP, 806
allotropes, 743
alloy broadening, 794
Auger recombination, 752
bandgap energy, 576, 733, 738
bandgap wavelength, 738, 768
bowing parameter, 785
Brillouin zone, 735
bulk, 578, 766–778
carrier concentrations, 748, 751
carrier generation, 752
carrier injection, 753
carrier recombination, 752
carriers, 734
degenerate, 750
density of states, 745
density of states, joint, 770
depletion layer, 756, 887
direct-bandgap, 737
dopants, 741
doped extrinsic, 886
drift velocity, 880
effective mass, 736
electroluminescence, 789–794
electron affinity, 873
elemental, 737
energy bands, 577, 733, 1305
energy–momentum relations, 735
excitons, 767, 784, 810, 811, 1010
extrinsic, 741
Fermi function, 746
Fermi inversion factor, 777
fundamentals, 731–766
GaAs, 805
GaAsP, 805
gain coefficient, 776
GaN, 807
heterojunction, 759
II–VI materials, 740
III–nitride materials, 738, 807–808
III–V materials, 738–739, 803–808
impact ionization, 876, 895
indirect-bandgap, 737
InGaAs, 805
InGaAsP, 805
InGaAsSb, 806
InGaN, 807
interaction with light, 766–781
internal efficiency, 754
intrinsic, 741
IV–VI materials, 741
Kronig–Penney model, 733
law of mass action, 750
minibands, 579, 764, 779, 852
mobility, 880
multiquantum-well, 764
nanocrystals, 579
nonradiative recombination, 752
occupancy probabilities, 746, 773
optics, 766–786
organic, 742
periodic table, 737
photoconductors, 875, 883–886
photoemission, 873
quantum dots, 579, 766
quantum wells, 578, 761–764
quantum wires, 579, 765
quantum-confined, 760, 779
quasi-equilibrium, 751
recombination coefficient, 752
recombination lifetime, 753
refractive index, 781
SESAM, 719
Shockley equation, 758
Si photonics, 375
SiC, 737
silicon photonics, 692, 780, 808, 1238
superlattice, 579, 764, 779, 852
transition probabilities, 774
Vegard's law, 739, 785
Semimetals
band structure, 577, 740, 741
graphene, 744
massless Dirac fermions, 744
Shockley, William B., 731
Shot noise, 518, 703, 912, 913
Silicon photonics, 375, 692, 808, 1176
direct-mounting integration, 809
flip-chip integration, 809
heteroepitaxy, 809, 859
heterogeneous integration, 809
hybrid approach, 809
microring laser, 859
OEIC, 1238
PIC, 809, 1238
PLC, 1238
quantum dots, 780
Single-mode
fiber, 405
waveguide, 359, 366
Skin depth, *see* Penetration depth
Slow light, 204
Snell's law, 54, 184, 221
at a boundary, 7

at a metamaterial boundary, 346
at a metasurface, 342
modified, 238, 342
negative refractive index, 307, 314
proof, 8
Solids, 575–580
covalent, 575
doped dielectric media, 568
ionic, 575
metallic, 576
molecular, 576
van der Waals, 573, 576, 745
Soliton
N-soliton wave, 1136
collision, 1199
condition, 1132
dark, 1138
directional-coupler router, 1187
envelope, 1133
fundamental, 1135, 1162
generation, 1137
higher-order, 1136
interaction, 1137
laser, 1138
optical fiber communications, 1256
period, 1136
photonic-crystal, 1141
sech pulse, 1133, 1136
self-frequency shift, 1140
solitary wave, 1131
spatial, 1039
spatial–temporal analogy, 1138
spatiotemporal, 1139
switch, photonic, 1199
temporal, 1130–1139
vector, 1200
SONET, 1239, 1276
Sonoluminescence, 607, 608
Spatial
coherence, 483
dispersiveness, 241
filter, 141, 151
frequency, 111, 113
harmonic function, 113
hole burning, 673
Lambertian pattern, 799
laser emission pattern, 675
LD emission pattern, 841
LED emission pattern, 799
solitons, 1039
spectral analysis, 114
Spatial light modulator
acousto-optic, 965
bistable, 1220
digital micromirror device, 1190
electro-optic, 987–989
liquid-crystal, 1002–1005
optically addressed, 987, 1004
parallel-aligned (PAL-SLM), 1005
Pockels readout optical (PROM), 988
Speckle, 406
Spectral
density, 479
hole burning, 651
packet, 596
width, 481
Spectrogram, 1083, 1156–1158
Wigner distribution function, 1157
Spectrum analyzer
acoustic, 963
interferometric, 1151
optical, 1151
Speed of light, 5, 42, 43, 163, 167, 184, 200, 204, 516, 1016
Spherical wave, 48, 177
complex amplitude, 77
conjugate, 1045
intensity, 77
paraboloidal approximation, 49
partially coherent, 488
pulsed, 1086
reflection, 78
wavefronts, 48
wavefunction, 48
Spin
-allowed transitions, 609, 811
-forbidden transitions, 609, 811
-orbit coupling, 565, 566, 811
-spin coupling, 565
angular momentum, 524
electron, 564, 746
multiplicity, 566
photon, 516, 524, 746
singlet state, 566, 575, 609, 811
triplet state, 566, 575, 609, 811
Spontaneous emission
atoms, 583, 587–588
enhanced, 464, 597, 796, 800
inhibited, 464, 617
occupancy probability, 773
peak rate, 786
Purcell factor, 598, 796, 855, 861
semiconductors, 769
spectral intensity, 775
transition rate, 592, 775
Stark effect, 568, 601
light shift, 600
quantum-confined, 1010

Statistical optics, 473–513
gain of spatial coherence, 502
imaging with incoherent light, 499
interference, 489
interferometry, 490, 493, 550
introduction to, 474–475
longitudinal coherence, 487
optical intensity, 475
partial polarization, 506
quantum interferometry, 550
spatial coherence, 483
spectrum, 476
temporal coherence, 476
transmission of random light, 497
Step-index fiber, 393–396, 398–405, 408
Stimulated Brillouin scattering, 614, 644
Stimulated emission
atoms, 585, 589–593
occupancy probability, 773
semiconductors, 769
semiconductors, indirect-gap, 772
transition rate, 592, 775
Stimulated Raman scattering, 613, 642, 691, 1140
Stokes
parameters, 214, 216, 507
vector, 214
Stokes, George Gabriel, 210
Strain
sensor, 1013
tensor, 968
Sum-frequency generation, 1027, 1197
Supercontinuum generation, 1139
Superlattice, 579, 764, 779, 852
quantum cascade laser, 852
Superluminescent diodes (SLEDs), 830
compare with LDs, 839, 841
compare with LEDs, 839, 841
light–current curve, 839
Surface plasmon
resonance spectroscopy, 329
resonance, 330
Surface plasmon polariton, 326
at DPS-SNG boundary, 311, 351
at metal–dielectric boundary, 326
nanolaser, 862
surface-charge wave, 311
Susceptibility, electric, 166
complex, 181
frequency-dependent, 184
resonant medium, 188
tensor, 170
Switch
acousto-optic, 959
electro-optic, 985
electroabsorption, 1010
ferroelectric liquid crystal, 1001
waveguide, 381
Switch, photonic, 1187–1211
acousto-optic, 964, 1194
all-optical, 1196–1211
architectures, 1187
Banyan switch, 1210
buffer, 1209
characteristics, 1188, 1196
circuit switching, 1209
configurations, 1189
electro-optic, 1192
energy required, 1202
Franz–Keldysh effect, 1193
FWM, 1199
heat dissipation, 1203
implementations, 1188
introduction to, 1164–1165
liquid-crystal, 1194
magneto-optic, 1195
mechano-optic, 1190
MEMS, 1190
nonlinear Kerr, 1148, 1198
nonlinear optical retardation, 1198
nonlinear Sagnac interferometer, 1201
opto-optic, 1196–1211
optoelectronic, 1188
optomechanical, 1190
packet, 1209
parametric, 1197
photonic-crystal, 1200
plasmonic, 1201
programmable delay, 1209
QCSE, 1193
quantum limit, 1202
quantum-confined Stark effect, 1010
ROADM, 1204, 1280
second-harmonic generation, 1147
semiconductor, 1193
SFG, 1197
SOA, 1193, 1196
soliton, vector, 1200
soliton-collision, 1199
space, 964, 1187–1203
space–wavelength, 1204
switching time, 1190, 1201
thermo-optic, 1195
time–space–time, 1207
time-division demultiplexer, 1207
time-division multiplexer, 1207
time-domain, 1206–1209
time-slot interchange, 1208, 1223

TOAD, 1201
wavelength converter, 1205
wavelength selector, 1203
wavelength-channel interchange, 1204
wavelength-selective, 1203–1206
XGM, 1196
XPM, 1197–1199

Tail
band, 750, 821, 870, 1010
Fermi, 748
Urbach, 778, 1010
Talbot effect, 135
TEM wave, 175, 307, 355, 506
Temperature
BEC formation, 602
blackbody spectrum, 605
correlated color, 812, 816
Doppler cooling limit, 600
earth, 606
single-photon recoil limit, 600
sub-recoil cooling, 600
sun, 606
thermographic images, 606
Tensor
conductivity, 882
constraints, 969, 1066, 1074
elasto-optic, 969
electric permittivity, 170, 228, 343
electric susceptibility, 170
first-rank, 228
fourth-rank, 969, 990, 1066
geometrical representation, 228
impermeability, 230, 989
index ellipsoid, 230
linear electro-optic, 990
magnetic permeability, 343
photoelasticity, 969
quadratic electro-optic, 990
quadric representation, 229, 230
second-order nonlinearity, 1066
second-rank, 170, 228–230, 343, 882
strain, 968
strain-optic, 969
third-order nonlinearity, 1066
third-rank, 990, 1066
zeroth-rank, 228
Terahertz
frequencies, 42, 853, 854, 1128
Term symbol
actinide metals, 569
atoms, 566
He, 566
lanthanide metals, 569
occupied subshells, 566
rare-earth elements, 569
transition metals, 569
Thermal light, 602–607
blackbody radiation, 602
rate equation, 604
Rayleigh–Jeans formula, 606
spectrum, 604
Stefan–Boltzmann law, 606, 618
thermography, 606
Wien's law, 617
Thermo-optic effect, 1195, 1198
Third-harmonic generation, 1037, 1063
Three-wave mixing, 545, 1025, 1043, 1050, 1061, 1066, 1073, 1076
pulsed, 1127, 1162
Time
-varying spectrum, 1083
lens, 1113, 1162
Townes, Charles H., 619
Transfer function, 1297, 1300
4-f imaging system, 139
free space, 119–122, 316
single-lens imaging, 144–145
Transformation optics, 343–348
cylindrical focusing, 346
refraction at normal incidence, 346
refraction without reflection, 345
transformation principle, 344–347
Transition
absorption, 584, 589
excitonic, 767, 779
free-carrier, 767
impurity-to-band, 767
interband, 766, 779
intersubband, 779
intraband, 767
miniband, 779
phonon, 767
spontaneous, 583, 587
stimulated, 585, 589
Transmittance
biprism, 57
complex-amplitude, 54, 222–225
diffraction grating, 59
Fabry–Perot, 266–268
Fresnel biprism, 57
graded-index plate, 60
optical components, 54, 497
plate of varying thickness, 55
power, 226
prism, 56
thin lens, 57
transparent plate, 55

Tyndall, John, 353

Ultrafast optics, 1078–1162
  introduction to, 1079
  linear, 1115–1125
  nonlinear, 1126–1145
  pulse characteristics, 1079–1088
  pulse compression, 1088–1102
  pulse detection, 1146–1158
  pulse propagation in fibers, 1102
  pulse shaping, 1088–1102
Ultraviolet
  EUV band, 697
  frequencies, 42
  UVA band, 42, 808
  UVB band, 42
  UVC band, 42, 808
  VUV band, 42
  wavelengths, 42
Uncertainty relation
  duration–bandwidth, 1290
  field quadratures, 543
  Heisenberg, 528, 543, 1291
  position–momentum, 543
  time–energy, 527, 555
Undulator, 702
Uniaxial crystal, 229, 233, 236, 317–320
Units, radiometric and photometric, 812
  illuminance, 812
  irradiance, 812
  luminous efficacy, 812
  luminous flux, 812
  radiant flux, 812
Up-conversion, 1027, 1054, 1055
  fluorescence, 611

Vacuum state, 518, 544
van Cittert–Zernike theorem, 503–505
VCSELs, 856–858
VECSELs, 721, 859
Vector
  beam, 179, 180
  potential, 177
Velocity
  group, 200, 204, 284, 352, 360, 370, 404, 412
  information, 204
  phase, 48, 200, 204, 284, 352
Verdet constant, 243
Veselago, Victor Georgievich, 303
Visibility, 78, 490
Visible
  frequencies, 42
  wavelengths, 42

Vortex, optical, 104
  laser, 860
  phase singularity, 104
  topological charge, 104, 155

Wave
  -particle duality, 521
  acoustic plane, 946
  beating, 74, 967, 1153, 1267
  complex amplitude, 46
  complex analytic signal, 71
  complex envelope, 47, 80, 175
  complex representation, 45
  complex wavefunction, 45, 72
  conjugate, 78, 1045
  cylindrical, 78
  evanescent, 120, 146, 224, 316, 368
  in a GRIN slab, 78
  localized, 523
  monochromatic, 44–52
  nondiffracting, 136
  paraboloidal, 49, 51, 81
  paraxial, 50, 1086
  partially coherent, 487, 488
  plane, 47, 111, 113
  polychromatic, 71–76
  pulsed, 73, 78, 1086
  quasi-monochromatic, 72
  quasi-plane, 407
  restoration, 1046
  retarder, 218, 248
  retarder, dynamic, 980
  retarder, liquid-crystal, 997–1001
  retarder, voltage-controlled, 998
  spherical, 48, 77, 78
  standing, 78
  stationary, 136
  wavefronts, 47
  wavefunction, 43, 45, 71
  wavelength, 47
Wave equation, 43, 45, 72
  diffusion equation, 1111
  generalized paraxial, 1124
  in free space, 163
  in homogeneous medium, 171
  in inhomogeneous medium, 169, 953
  in medium, 167
  in nonlinear dispersive medium, 1133
  in nonlinear medium, 1019, 1047
  SVE, 1111
  SVE, nonlinear, 1133
Wave optics, 41–78
  introduction to, 42–43
  postulates, 43–44

vs. electromagnetic optics, 180
vs. ray optics, 52
Wave-transfer matrix, 258–265, 1303
antireflection film, 264
beamsplitter, 263
cascade of elements, 259
dielectric boundary, 262, 264
dielectric slab, 263, 265
homogeneous medium, 260
lossless medium, 261
lossless system, 262
relation to scattering matrix, 260
Wavefronts
anisotropic medium, 234–237
helical, 103, 155, 524
plane-wave, 48
spherical-wave, 48
Wavefunction
complex, 45, 716, 1080, 1158
electron, 562, 582, 734, 1291, 1304
entangled-photon, 547
harmonic-oscillator, 542
plane-wave, 48
quantum mode, 543
single-photon, 521
spherical-wave, 48
two-photon, 546
Waveguide, optical, 354
arrays, 383
asymmetric planar, 372
bounce angles, 357, 365, 390
Bragg-grating, 385
channel, 374
cladding, 363
confinement factor, 369, 390
core, 363
coupled-mode theory, 378–384
couplers, 378–383
coupling, 376–384, 390, 985
cutoff, 359, 366, 390
cylindrical, 392
dispersion relation, 359, 370
evanescent wave, 368
extinction coefficient, 368
fiber, 392
field distributions, 358, 367, 389, 390
GaAs/AlGaAs, 375
glass, 375
Goos–Hänchen effect, 371
group velocity, 360, 370
InGaAsP, 375
input coupling, 376
$LiNbO_3$, 375
materials, 375
metal–insulator–metal, 387
metal–slab, 388, 390
modes, 355–357, 363–366
multimode fields, 362
number of modes, 359, 366, 367, 390
numerical aperture, 366
optical-power flow, 359, 363
periodic, 384
photonic-crystal, 385–386
planar dielectric, 363–372, 389, 390
planar-mirror, 355–363, 1305
plasmonic, 386–388
power-transfer ratio, 381
propagation constants, 357, 365
rectangular dielectric, 373, 390
rectangular mirror, 372
side coupling, 377
silica-on-silicon, 375
silicon-on-insulator, 375
single-mode, 359, 366, 390
switch, 381
transfer distance, 378, 381, 985
two-dimensional, 372–375
Wavelength, 47
γ-ray, 697
-division multiplexer, 1179
activation, 885
bandgap, 738, 768
converter, 1205
de Broglie, 601, 761
infrared, 42, 853
laser, 644, 706
laser amplifier, 644
plasmon, 312
ultraviolet, 42, 697
visible, 42
X-ray, 697
zero-dispersion, 1140
Wavenumber, 48, 173
complex, 182
Wavepacket, 73
mode, 527
random sequence, 482
single-photon, 527, 594
velocity, 360
Wavevector, 47
Width of a function
$1/e$-, 1292
3-dB, 1292
FWHM, 1292
Gaussian, 1290–1292
measures of, 1290–1293
power-equivalent, 1291
power-RMS, 1290

root-mean-square (RMS), 1290
Wiener–Khinchin theorem, 480, 512
Wigner distribution function, 1157
WOLED, 810, 816, 1002
Wolf, Emil, 473

X-ray
energies, 697
hard-X-ray (HXR) band, 697
imaging, 705
optical components, 698
soft-X-ray (SXR) band, 697
wavelengths, 697

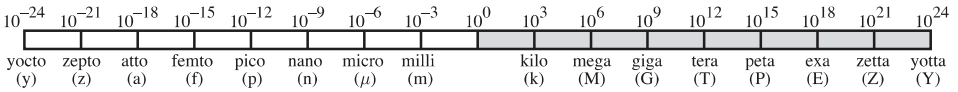
Yablonovitch, Eli, 255
Yablonovite, 297
Young, Thomas, 41

Zeeman effect, 568, 599, 601
Zero-point energy, 518, 536, 573, 586
Zone plate, Fresnel, 67, 118

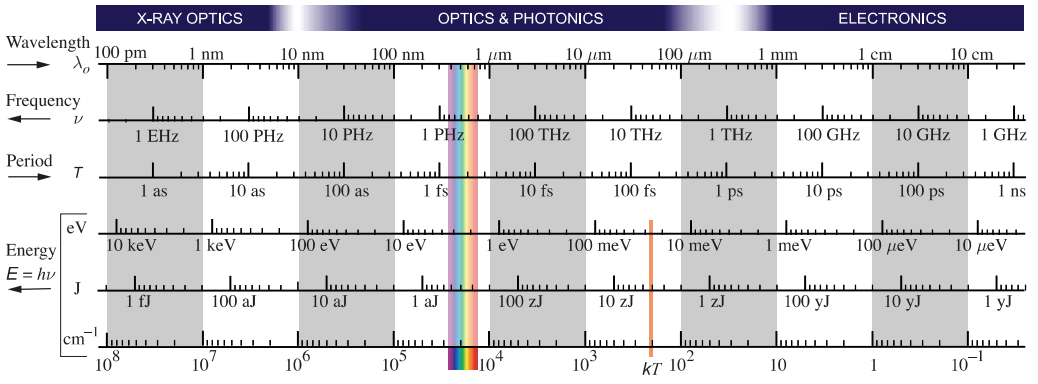
USEFUL CONSTANTS

Speed of light in free space	c_o	2.9979×10^{8}	m/s	Planck's constant	h	6.6261×10^{-34} J · s
Permittivity of free space	ϵ_o	8.8542×10^{-12}	F/m	Electron charge	e	1.6022×10^{-19} C
Permeability of free space	μ_o	1.2566×10^{-6}	H/m	Electron mass	m_0	9.1094×10^{-31} kg
Impedance of free space	η_o	376.73	Ω	Boltzmann's constant	k	1.3807×10^{-23} J/°K

PREFIXES FOR UNITS

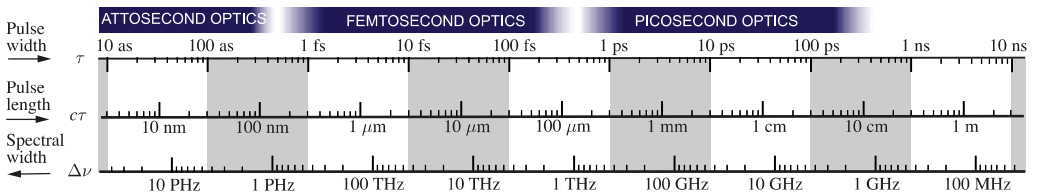


THE PHOTON

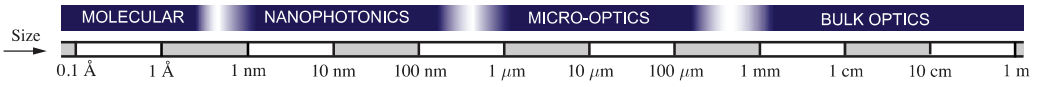


A photon of free-space wavelength $\lambda_O = 1$ μm has frequency $\nu = 300$ THz, period $T = 3.33$ fs, and energy $E = 1.24$ eV = 199 zJ = 10^4 cm^{-1}. At room temperature ($T = 300^o$ K), the thermal energy $kT = 26$ meV = 4.14 zJ = 209 cm^{-1}.

OPTICAL PULSES



PHOTONIC STRUCTURES



Fundamentals of Photonics

A complete, thoroughly updated, full-color third edition

Fundamentals of Photonics, *Third Edition* is a self-contained and up-to-date introductory-level textbook that thoroughly surveys this rapidly expanding area of engineering and applied physics. Featuring a blend of theory and applications, coverage includes detailed accounts of the primary theories of light, including **ray optics**, **wave optics**, **electromagnetic optics**, and **photon optics**, as well as the **interaction of light with matter**. Presented at increasing levels of complexity, preliminary sections build toward more advanced topics, such as **Fourier optics and holography**, **photonic-crystal optics**, **fiber and guided-wave optics**, **LEDs and lasers**, **acousto-optic and electro-optic devices**, **nonlinear optical devices**, **ultrafast optics**, **optical interconnects and switches**, and **optical fiber communications**. The third edition features an entirely new chapter on the **optics of metals and plasmonic devices**. Each chapter contains highlighted equations, exercises, problems, summaries, and selected reading lists. Examples of real systems are included to emphasize the concepts governing applications of current interest. Each of the chapters of the second edition has been thoroughly updated.

BAHAA E. A. SALEH, PhD, has been Distinguished Professor and Dean of CREOL, The College of Optics and Photonics at the University of Central Florida, since 2009. He is also Professor Emeritus at Boston University. Saleh is the author of *Photoelectron Statistics* and *Introduction to Subsurface Imaging*, and is the Founding Editor of *Advances in Optics and Photonics*. He is a Fellow of APS, IEEE, OSA, and SPIE, and is the recipient of the OSA Beller Medal, the OSA Mees Medal, the SPIE BACUS Award, and the Kuwait Prize.

MALVIN CARL TEICH, PhD, is Professor Emeritus at Boston University and Columbia University, and a member of the Boston University Photonics Center. He is the coauthor of *Fractal-Based Point Processes* and is a Fellow of the IEEE, OSA, SPIE, APS, AAAS, and ASA. He is the recipient of the IEEE Browder J. Thompson Memorial Prize Award, the IEEE Morris E. Leeds Award, the Memorial Gold Medal of Palacký University, and the Distinguished Scholar Award of Boston University.

WILEY

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.