The Measurement of TIME

Time, Frequency and the Atomic Clock

Claude Audoin &
Bernard Guinot

CAMBRIDGE

The Measurement of Time

Time, Frequency and the Atomic Clock

This book provides a comprehensive introduction to the physics of time and time measurement, from an historical perspective to the modern day. It discusses the stability and accuracy of atomic frequency standards, including different types of oscillators and atomic clocks, and covers the most recent developments and uses of these devices. The precision of atomic clocks and the atomic time scale are considered in the context of fundamental physical research, with relation to general relativity and searches for possible time variation in the fundamental constants. The authors also discuss International Atomic Time and its relationship to Coordinated Universal Time and the time scales used in astronomy, as well as applications such as the Global Positioning System (GPS).

The book will be ideal as an introduction for graduate students or researchers entering the fields of time and frequency metrology and precise astronomical observation. It will also be useful as a reference for scientists working in these and other applied areas, such as geophysics, atomic physics, astronomy and telecommunications. General readers with a background in science should also find this a fascinating book.

CLAUDE AUDOIN received his doctorate in physics from the Université Paris Sud in 1967. He was Director of the Laboratoire de l'horloge atomique, Orsay, from 1972 to 1985 and managed research programmes related to hydrogen masers, optically pumped caesium beams, trapped ions, frequency stabilised lasers and frequency metrology. He has published numerous papers on the subject of atomic frequency standards and their metrological properties, and is co-author of a previous book covering the quantum physics of atomic frequency standards. Dr Audoin is also on the editorial board of *Metrologia* and is a member of the Bureau des longitudes.

BERNARD GUINOT was an officer in the merchant navy before receiving his masters degree in mathematics in 1952. He then joined the Paris Observatory and in 1958 received the Doctorat d'Etat. His research was devoted to fundamental astronomy, space geodesy and the measurement of time, and he is the author of numerous papers on these subjects. From 1964 to 1988, Dr Guinot was Director of the Bureau international de l'heure, where he was involved in the transition to modern techniques for measuring the rotation of the Earth and the development of International Atomic Time, which is now under the responsibility of the Bureau international des poids et mesures. He is a member of the Bureau des longitudes and the Academia Europaea.

The Measurement of Time

Time, Frequency and the Atomic Clock

Claude Audoin
CNRS, Orsay, France

Bernard Guinot
Paris Observatory, France

Translated by Stephen Lyle

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521800808

Originally published in French as *Les Fondements de la Mesure du Temps* by Masson, Paris

First published in English 2001

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data

Audoin, Claude, 1933–
The measurement of time: time, frequency and the atomic clock / Claude Audoin, Bernard Guinot.
p. cm
Includes bibliographical references and index.
ISBN 0 521 80080 3 – ISBN 0 521 00397 0 (pbk.)
1. Time measurements. Atomic clocks. I. Guinot, Bernard. II. Title.
QB213.A88 2001
528'.7–dc21 2001025507

ISBN 978-0-521-80080-8 Hardback
ISBN 978-0-521-00397-1 Paperback

Contents

Acknowledgements *page* x

1 Introduction to the book 1

2 The principles of time measurement 6
2.1 Introduction 6
2.2 Time and reproducibility: the notion of duration 7
2.2.1 Chaos 7
2.2.2 Transposing the measurement of time to the measurement of another quantity 7
2.2.3 The local nature of phenomena 8
2.2.4 Differential ageing? 9
2.2.5 Reproducibility and measurement standards 9
2.3 Time, evolution and time scales 10
2.4 The two modes of time measurement 11
2.4.1 Use of reproducibility 12
2.4.2 Use of a dynamical model 13

3 Time measurement and theoretical models 16
3.1 The Newtonian model and absolute time 17
3.2 Special relativity 19
3.2.1 Physics in inertial systems 19
3.2.2 Lorentz transformation and invariance of the interval 21
3.2.3 Time in special relativity 22
3.2.4 Field of application of special relativity 23
3.3 General relativity 24
3.3.1 Overview 24
3.3.2 Post-Newtonian approximation 27
3.4 Summary and conclusions 36

4 The evolution of time measurement 38
4.1 Date, calendar and hour 38
4.2 Time measurement based on alternation of day and night 40
4.2.1 Mean solar time 40
4.2.2 Universal time scales and time zones 41
4.2.3 Towards a unique realisation of Universal Time 43
4.2.4 Definition of the second before 1960: the second of mean solar time 46
4.3 Time based on Solar System dynamics: Ephemeris Time 46
4.3.1 First doubts on the uniformity of the Earth's rotation 46
4.3.2 Acceptance of irregularities in the Earth's rotation 47
4.3.3 Definition of the second from 1960 to 1967: the ephemeris second 50
4.3.4 Ephemeris Time: a scale reserved for astronomers 51
4.4 Atomic time measurement 51
4.4.1 Atomic frequency standards and the atomic definition of the second (1967) 51
4.4.2 From Universal Time to International Atomic Time 56
4.5 Coordinated Universal Time 61
4.6 Final comments 64

5 Clock time 65
5.1 Introduction 65
5.2 Frequency and time instability 67
5.2.1 Definitions 67
5.2.2 Notation 67
5.2.3 Measurements in the time domain 69
5.2.4 Measurements in the spectral domain 71
5.2.5 Modelling frequency and time fluctuations 72
5.2.6 Characterising frequency instability 73
5.2.7 Characterising time instability 81
5.2.8 Determining contributions from each oscillator 82
5.3 Mechanical oscillators 83
5.4 Piezoelectric oscillators 84
5.4.1 Resonators 84
5.4.2 Oscillators 86
5.4.3 Performance levels 87
5.5 Atomic time and frequency standards. The notion of accuracy 87
5.6 Time and frequency comparisons 89
5.6.1 General observations 89

5.6.2 Relationship between frequency and proper time increment 90
5.6.3 Time comparisons 90
5.6.4 Frequency comparisons 104

6 Atomic frequency standards 109
6.1 Spectroscopic basis for atomic frequency standards 109
6.1.1 Invariance and universality of atomic properties 109
6.1.2 Spectral properties of alkali atoms 110
6.1.3 Selection rules 116
6.1.4 Resonance width 118
6.2 Operations involved in an atomic clock 119
6.2.1 Observation in very dilute media 119
6.2.2 Preparation and detection of atomic states 119
6.2.3 Eliminating the Doppler effect 123
6.2.4 Interaction between atoms and electromagnetic field 125
6.2.5 Associated electronic systems 126
6.3 Realising an approximation to the isolated atom at rest 127
6.4 Caesium beam clocks 129
6.4.1 Magnetically deflected caesium beam clock 130
6.4.2 Optically pumped caesium beam clock 162
6.4.3 Laser-cooled caesium clock 170
6.4.4 Laser-cooled rubidium clock 184
6.5 Hydrogen masers 184
6.5.1 Basic principles of the hydrogen maser 185
6.5.2 Oscillation condition 187
6.5.3 Active hydrogen maser 191
6.5.4 Passive hydrogen maser 201
6.5.5 Metrological properties of the hydrogen maser 206
6.5.6 Cryogenic hydrogen masers 210
6.6 Rubidium cell clocks 211
6.6.1 Description 211
6.6.2 Operating principles 213
6.6.3 Electronics 215
6.6.4 Metrological properties of rubidium cell clocks 215
6.6.5 Current studies and prospects 217
6.7 Stored ion clocks 217
6.7.1 Confining ions in a radiofrequency trap 218
6.7.2 Confinement of a cloud of ions 222
6.7.3 Description of a frequency standard using a cloud of mercury ions 224

6.7.4 Metrological properties of frequency standards using a cloud of mercury ions 228
6.7.5 Ytterbium ion frequency standard 229
6.7.6 Laser-cooled mercury ion frequency standard 229
6.8 Other atomic frequency standards 230
6.8.1 Magnesium beam frequency standard 231
6.8.2 Optical frequency standards 231
6.8.3 Optical frequency standards based on individual ions at rest 231
6.8.4 Optical frequency standards for time measurements 233
6.9 Conclusion 234

7 Atomic time measurement 236
7.1 Definitions 236
7.1.1 Definition of the second 236
7.1.2 International Atomic Time and its relationship with geocentric and barycentric coordinate times 237
7.2 Establishing International Atomic Time 240
7.2.1 Time scale algorithms 240
7.2.2 Algorithm for International Atomic Time 241
7.3 Properties of TAI and UTC 252
7.3.1 Reliability 252
7.3.2 Accuracy of readings 252
7.3.3 Frequency stability 254
7.3.4 Frequency accuracy 254
7.4 World organisation of time measurement 255
7.4.1 Coordinated Universal Time 255
7.4.2 Local representations of UTC and independent local atomic times 258
7.4.3 Other forms of atomic time 259
7.5 Dissemination of time and frequency 260
7.5.1 Frequency 260
7.5.2 Time 261
7.6 Summary and prospects 262

8 Astronomical times 264
8.1 Universal Time 264
8.1.1 Conceptual definition 264
8.1.2 Practical definition of UT1 266
8.1.3 Other forms of Universal Time 269
8.1.4 Measurement techniques 269
8.1.5 Irregularities in the Earth's rotation 273

8.1.6 Use of Universal Time 276
8.2 Ephemeris Time 278
8.2.1 Definition 278
8.2.2 Determining Ephemeris Time 279
8.2.3 Difficulties and current solutions 280
8.3 Pulsar Time 281

9 Ultraprecise time and frequency applications 286
9.1 Fundamental research 287
9.1.1 Metrology 287
9.1.2 Measuring physical constants 289
9.1.3 Impact on atomic physics 290
9.1.4 Spacetime structure and gravity 291
9.2 Positioning, geodesy and navigation 295
9.2.1 Principles of time and frequency methods 295
9.2.2 The Global Positioning System 297
9.2.3 Other positioning systems 301
9.3 Very Long Baseline Interferometry (VLBI) 302
9.4 The TOPEX/POSEIDON mission 303
9.5 Telecommunications 306

Appendices
Appendix 1 Acronyms for time laboratories 308
Appendix 2 Abbreviations 311
Appendix 3 Definitions of base units in the SI system 313
Appendix 4 International services 315

References 317
Index 331

Acknowledgements

We would like to thank our colleagues and the organisations that have helped us to prepare this book in various ways, through discussions, provision of unpublished papers, or figures and photographs, supply of useful information and critical rereading of parts of the text: Felicitas Arias, Jacques Azoubib, Roland Barillet, Andreas Bauch, Bernard Cagnac, Pierre Cérez, André Clairon, Noël Dimarcq, Robert E. Drullinger, Bernard Dubouis, Martine Feissel, Dominique Ferrand, Georges Fréon, François Gonzalez, Michel Granveaud, Peter Kartaschoff, Philippe Laurent, Michel Lefebvre, Wlodzimierz Lewandowski, Gérard Petit, Pierre Petit, John D. Prestage, Pascal Rochat, Geneviève Théobald, Claudine Thomas, the Associazione Elettrotecnica ed Elettronica Italiana, the Institut de mécanique céleste et de calcul des éphémerides at the Paris Observatory, and the International Bureau of Weights and Measures (BIPM, Bureau international des poids et mesures). Special thanks also to Liliane Bergeal who sorted out a good few administrative problems and to Pascale Michel for her skilful typing. We are most grateful to Stephen Lyle who took care of the translation from French into English. This book has been written at the suggestion and with the encouragement of the Bureau des longitudes, Paris.

1

Introduction to the book

The use of atomic properties for time measurement was born in 1955 when the first caesium beam frequency standard began regular operation in the United Kingdom. Of course, other types of atomic frequency standard had already been attempted. However, they had proved unable to provide a unit of time and uniform time scales superior to those based on celestial motions, in the way that the caesium standard could. Caesium standards soon proliferated, whilst their accuracy advanced by leaps and bounds. The good agreement between them inspired confidence and, in 1967, it was possible to make an atomic definition of the second to replace the previously used astronomical definition. Several years later, the existence of an adequate atomic time scale for the management of world affairs was officially recognised by the 14th General Conference on Weights and Measures (1971).

Despite these achievements, over a quarter of a century later, the atomic measurement of time has still not really become familiar. Could this be due to a deeply ingrained habit of measuring out our everyday existence by the movements of the celestial bodies? Or could it be a reaction to the lack of poetry in atomic clocks, or their inscrutable accuracy? This may be so, but there are also more objective reasons that can sometimes make it difficult to accept atomic time. One of these is the need to formulate the measurement of time within a relativistic framework, and another is the complex relationships between atomic time and the time of the dynamical theories used by astronomers. The latter problem was a subject of controversy over twenty years or so, and even the International Astronomical Union was not completely able to lay the ghost with a resolution adopted in 1991. Recently, some have thought that the discovery of rapid pulsars, the so-called millisecond pulsars, might bring the measurement of time back within the jurisdiction of astronomy.

Meanwhile, in the peace of their laboratories, physicists are barely susceptible to such interpellations. Following a period of relative stagnation which

lasted about twenty years, a sudden gain in the accuracy of frequency standards was achieved in 1994. The new techniques that made this improvement possible are also full of promise for further progress in the near future. On the other hand, the world of industry has not limited its involvement. Manufactured quartz clocks and atomic clocks of various types have shown continual improvement. The instrument favoured by fundamental metrological research has always been the caesium clock and thousands of these have already been produced. In 1994, a new manufactured version came out which scored a tenfold enhancement over the whole range of relevant qualities, compared with previous versions.

Through the industrial manufacture of timing equipment, important applications of time–frequency techniques have seen the light of day. An example is the Global Positioning System (GPS), in full operation by 1995. This interests a whole range of potential users, from geodesists seeking to represent the whole globe to the nearest centimetre, to ramblers, not forgetting the military applications for which the system was originally devised.

Considering these recent advances in the already well-tested atomic techniques of time measurement and the new problems posed by such a gain in quality, it seemed an opportune moment to present something of their foundations: the basic principles, construction of the best standards, ways of comparing frequencies and times with the ultimate accuracy, and ways of accessing the primary standards, that is, the best realisations of the second and the International Atomic Time at the highest level of accuracy. It also seemed useful to recall the role of astronomical time scales, which remain important for various reasons.

As we have already mentioned, the fundamental measurement of time can no longer be conceived of outside a relativistic framework. Up to the present time, Einstein's general relativity, the simplest of these theories, has always remained in agreement with experimental results. We shall therefore adopt it as our working hypothesis. However, we must express some reservations in this regard. To begin with, we had hoped to treat the measurement of time within the framework of classical physics and then indicate, in small print as it were, what would be required to take relativistic effects into account. But then it seemed to us that this method ran the risk of promoting a widely held view that general relativity is in some sense a mere addition to classical theory, that its contribution to science consists in making a set of relativistic corrections. In reality, general relativity is a complete model for the structure of space and time, as well as for gravitational effects. We have therefore directly established the useful developments within the context of this theory. The reader who is barely familiar with such questions should nevertheless persevere, for we shall

go no further than recalling the postulates of the theory and its mathematical consequences in the greatest simplicity.

The inevitable linearity of a written account is poorly suited to the many interactions between the varied aspects of time measurement. The brief description given here of the various chapters in the book is intended to orient the reader with regard to this problem.

Chapter 2 recalls the fundamental principles of time measurement for macroscopic phenomena, the only ones to be considered here. In truth, this is essentially a review of H. Poincaré's analysis, whose relevance extends to the relativistic context.

Chapter 3 shows how time measurement relates to general relativity. In an extended space, time is one element of a four-dimensional coordinate system and cannot be dissociated from the spatial coordinates. We shall define the coordinate systems used and establish some equations relating local physics to spacetime coordinates. We shall also discuss the conventions involved in synchronisation and comparison of distant clocks.

Chapter 4 has a historical objective. It shows how time measurement has evolved, in particular, during the twentieth century. We shall discuss the development of ideas stimulated by development of techniques, as well as the problems raised by transition from astronomical measurement of time to its atomic measurement.

Chapter 5 describes the tools involved in using atomic time standards, and in particular, methods for characterising their metrological qualities and comparing one with another. The characterisation of frequency stability has led to new developments in statistics. These refer to time series of measurements and the definition of variances which allow manufacturers and users of frequency standards to speak a common language. The characterisation of the accuracy of a frequency is also a problem specific to time measurement, although it could be extended to any physical quantity whose unit is directly defined by a natural phenomenon that is considered to be reproducible. Concerning the comparison of frequencies and time scales at a distance, we shall see that it is a crucial question in the sense that it often gives rise to greater uncertainties than those involved in the quantities to be compared.

Chapter 6 treats atomic frequency and time standards, also called atomic clocks. We shall begin by reviewing, on an elementary level, the main ideas of atomic physics and spectroscopy required to understand the workings of these standards and their principal features. There are several sorts of atomic clock, all making use of some transition between atomic levels of the same type, but in different atoms, such as caesium, hydrogen and rubidium. Even ions are used, like the mercury ion. It is the caesium clocks built and used in metrolog-

ical laboratories that best achieve the definition of the second. These are the primary frequency and time standards. Their main feature is accuracy. Industrial versions of the caesium clock allow a relatively inexpensive dissemination of the unit of time, and hence of atomic time itself. Other types of atomic clock will also be described. They have been built or developed because they are specially suited to certain fields of application. Some are bulky but exceptionally stable, whilst others are less stable but rather compact. In this chapter, we shall go into the principles by which atoms are manipulated using the radiation from lasers. This experimental technique, which appeared recently, makes it possible to slow down and trap atoms before launching them in a highly controlled manner at very low speeds. Since 1996, this technique has led to an accuracy that excels that of all other primary frequency and time standards. The potential in these methods for manipulating atoms by laser beams, combined with the extraordinary international competition that they arouse, are a firm guarantee of future improvement in the performance of atomic clocks.

Chapter 7 is mainly devoted to the construction of the unique atomic time scale taken by convention as the international standard, namely, the International Atomic Time (denoted in all languages by TAI, from the French name Temps atomique international). We shall see how worldwide cooperation has been able to establish and maintain such a scale, and we shall investigate its qualities and the way it is disseminated. But we shall also see how it has been necessary to compromise, accepting as the practical standard a time scale based on the TAI but ingeniously adapted to maintain connections with the Universal Time (UT) as specified by Earth's rotation. This hybrid is known as Coordinated Universal Time (officially denoted by the acronym UTC). We shall show how the so-called proper second required for laboratory work can be obtained from the TAI or UTC.

Chapter 8 gives definitions of the various astronomical times in common use. These time scales have been gradually elaborated by observation of celestial bodies throughout a long period of history. Their usual definitions carry the traces of such historical developments and are thereby somewhat obscured for those unfamiliar with astronomy. We have thus made every effort to bring out the basic concepts upon which the operational definitions given by astronomers are founded. We shall investigate the purpose of these time scales. In particular, the Universal Time, witness to Earth's rotation, is a very important physical quantity requiring continuous measurement. Indeed, the efforts devoted to measuring Universal Time remain on a par with those required to maintain atomic time. Concerning pulsars, although they cannot provide a good measure for time, their study is nevertheless hampered by the uncertainties involved in atomic time. This is one of the most demanding applications for atomic clocks.

Finally, Chapter 9 presents some applications selected for the high level of quality in the time and frequency standards they require. Naturally, these applications belong to the area of fundamental research. The use of time standards in more utilitarian programmes is concomitant with a certain degree of sacrifice as far as their ultimate qualities are concerned, in favour of reliability and reduced costs. However, the more practical instruments sometimes lag only slightly behind those destined for laboratory use when it comes to accuracy and stability. The Global Positioning System provides a good illustration for this. As we are all either direct or indirect users of this system, we shall go into some degree of detail. We shall also show, as an example, how the measurement of time is relevant to a space-based oceanographic mission, TOPEX/POSEIDON. Indeed, we shall see that it underlies many aspects of the mission, with a high requisite level of quality.

2

The principles of time measurement

2.1 Introduction

The range of physical time spans is vast, from the Planck time of 10^{-43} s, which is the shortest we can conceive of in physics, to those encountered in cosmology, of the order of tens of billions of years, or 10^{17} to 10^{18} s.

From this range, we shall be concerned with only that very small part which has given rise to accurate measurements. Roughly speaking, this extends from periods of the order of 10^{-15} s, associated with visible atomic transitions, to periods of several thousand years, let us say 10^{11} s, over which astronomical observations have been made. This is the domain of typical laboratory experiments and dynamical astronomy, which have each in turn required or supplied time standards: the domain of time metrology.

Until the birth of relativistic theories at the beginning of the twentieth century, time and space formed an external and immutable framework within which our various activities could be accomplished. We were there to perceive this framework through the presence and evolution of beings and objects. At the same time, we were convinced that it existed independently of its content, and even that it would have existed without any content at all. Today, it is thought that spacetime cannot be conceived of without mass and energy, and that there is no privileged time such as Newtonian absolute time. Such a conception forces us into redoubled humility before nature. Let us admit that we do not know what time is. We know only that we need a temporal coordinate and three spatial coordinates to represent physical phenomena and invent mathematical models for them. Why these four independent coordinates and not five, or more? It is quite simply because four are enough to ensure that no measurable inconsistency occurs in our physical models.

Freed from absolute time, it becomes easier to accept a pragmatic approach to defining, or rather, to measuring time. Indeed, time is defined in such a way

that mathematical models of nature remain simple. It is worth adding that the other physical quantities, such as length, mass, and so on, should be defined in the same way. It is thus by examining criteria of simplicity that we shall be able to tackle the basics of time measurement, define a unit of time and choose a time scale that can be taken as a world standard for dating events.

2.2 Time and reproducibility: the notion of duration

Nothing could be simpler than to take a well-known observational feature of our relationship with nature and turn it into a postulate. This feature is the reproducibility of any experience in time and in space. It was H. Poincaré who first stated this postulate in his discussion on the measurement of time [2.1] (translation): 'When we use the pendulum to measure time, what postulate do we implicitly assume: it is that the duration of two identical phenomena is the same; or, if we prefer, that the same causes require the same time to produce the same effects.'

However, this simplicity is illusory. Several claims demand our attention.

2.2.1 Chaos

Experience teaches us only [2.1]: 'that approximately like causes take roughly the same time to produce roughly the same effects.'

Poincaré's assertion here may be somewhat surprising. He was one of the pioneers in the study of deterministic chaos according to which roughly similar causes can produce quite different effects in the same time.

The existence of chaotic phenomena does not contradict the reproducibility postulate. It merely reminds us that caution is of the essence when applying it. In metrology, the aim is to provide reproducible standards, and it is founded on well-behaved phenomena for which an approximate application of the reproducibility postulate remains possible. It is indeed precisely the role of the metrologist to recognise and use such phenomena, whose dependence on poorly controllable perturbing causes remains weak and bounded.

2.2.2 Transposing the measurement of time to the measurement of another quantity

When Poincaré speaks of the same causes and the same effects, he assumes that we already have standards for assessing them. A measurement of time, that is, in this case, duration, can only ever be accessible to us through the measurement of another quantity.

Returning to his pendulum example, if we wish to obtain the unit duration, the second, by a single there-and-back oscillation, we must define with great accuracy some 'point' of reference through which the pendulum passes and obtain a signal at the moment it passes there. In practice, nothing is instantaneous. A photoelectric method could give a signal extending over 0.001 s. The second would then only be given to a relative accuracy of 10^{-3}. But the reproducibility postulate becomes useful when we can repeat the experiments without interruption. For then, although the read time remains unchanged, the error in the unit is, on average, divided by the number of cycles observed. In the pendulum example, it reduces to 10^{-8} after one day's observation. At the present time, the second is defined as the duration of 9 192 631 770 periods corresponding to a specified atomic transition. Simply counting periods would lead in one day, provided none were missed, to a relative uncertainty in the second of 10^{-15}.

In the two examples above, a measurement of time is converted into a voltage measurement, but the reproducibility postulate can reduce this measurement to a simple, almost qualitative act of recognition, in such a way that any uncertainty in the volt has no effect whatever on that in the second. This highlights the advantages of referring to phenomena of very short duration.

2.2.3 The local nature of phenomena

In order to control the causes of the phenomenon, in the hope that they will be the same whenever we wish to reproduce it, the experiment must be a *local* experiment, that is, rather close to the experimenter.

Are there any naturally reproducible large scale or distant phenomena that are sufficiently rapid to answer the needs of accuracy? At the present time, we do not know of any. The rotation of the Earth seemed for a long time to fulfill these requirements, until there came proof of its irregularities, during the second half of the twentieth century. The rapidly rotating pulsars discovered since 1982, with periods in the millisecond range, would have been excellent standards for duration, had the rotation not been slowing down!

If a distant and well reproducible phenomenon were discovered, we could base the definition of the unit of duration on observations of it. In fact, we would certainly do so, if there were some advantage to it. However, this would require recourse to theory, if only to the theory of propagation of electromagnetic waves. Used in the purest possible way, without appealing to any theory that might one day be contested, reproducibility leads to local measurement standards.

2.2.4 Differential ageing?

Poincaré's postulate mentions causal relations. We may wonder what happens when two phenomena occur in parallel, independently of one another. If a chemical reaction reaches conclusion whilst our atomic clock counts 10^{10} periods, can we be sure that in a million years from now, the same reaction will still take the same number of periods? The postulate allows us to assert that it will indeed be the same number. In fact, there is nothing to stop us establishing a causal connection between the two phenomena, for example, by triggering the reaction after n atomic periods and considering the global phenomenon 'period counting and chemical reaction'. The duration of this global phenomenon must be invariable. The postulate leaves no room for differential ageing of the various natural phenomena. Concerning the question as to whether local phenomena and their laws age globally, this would suppose the existence of some fictional time by which we could make our assessment. The existence of such a time is a pointless hypothesis.

Notwithstanding, it is as well to refrain from unjustified overconfidence as regards the reproducibility postulate. Although it has never yet been found to fail, it may one day be brought into question by experimental progress.

2.2.5 Reproducibility and measurement standards

From the above discussion, we can conclude that the definition of measurement units to be used for local physics can be based upon the reproducibility of local experiments. Examining the definitions of the seven basic units in the SI system (Système international d'unités), viz., the metre, the kilogram, the second, the ampere, the kelvin, the mole and the candela (definitions given in Appendix 3), reveals that this is currently the case for five of them. The exceptions are the kilogram and the kelvin. The definition of the kilogram as the mass of a unique object appeals to the supposed permanence of a property of this object, which is a form of reproducibility. With regard to the kelvin, its definition appeals to a more complex model, namely, the laws of thermodynamics based on additional postulates.

The unit of duration, the second, normally referred to as the unit of time, is therefore, in its present definition, a unit to be established locally, just like the other units. This has consequences when we have to treat measurements within the model provided by general relativity, and this is often necessary, precisely because of the great accuracy with which we can determine the second.

2.3 Time, evolution and time scales

Let us now turn to evolving phenomena which we have no reason to believe are reproducible. Whilst such phenomena abound, we shall be concerned with those belonging to the dynamics of celestial bodies, because of the high quality of the mathematical models developed to describe them, and perhaps also because, in the past, time was always associated with the motions of such heavenly objects.

In order to describe these motions, we need only a scale of coordinates in time, that is, a time scale, whose main qualities are that everyone should accept it and that it should be accessible to all. (This already raises the question of how to synchronise clocks used by distant observers, a discussion we shall postpone until Chapter 3. Let us just say for the moment that we can synchronise.) But if we wish to establish a model for these motions that allows us to predict them as in the astronomical ephemerides, the temporal argument in the theory must satisfy further criteria. What are these criteria? Let us turn once again to Poincaré's own account, rather than attempting to paraphrase:

> They [astronomers] define duration in the following way: time must be defined in such a way that Newton's law of gravitation and second law of motion both hold. The law of gravitation is an observational truth and, as such, is only approximate. This shows that we have once again only an approximate definition. If we assume that we now adopt another means of measuring time, the experiments upon which Newton's law of gravitation is founded would nevertheless have the same meaning. It is just the statement of the law that would change, because it would be translated into another language. It would obviously be much less simple. The definition implicitly adopted by astronomers can thus be resumed in the following way: time must be defined in such a way that the equations of mechanics become as simple as possible. In other words, there is no way of measuring time that is truer than all others. The one that is generally adopted is merely the most convenient.

We thus find ourselves with two definitions for the measurement of time, namely, the time which stems from the reproducibility of local phenomena, and the time that arises from Newtonian dynamics. Are these two definitions equivalent? Until the beginning of the twentieth century, there was every reason to believe that either of these definitions would lead to representations of the same absolute time, differing only by their level of quality. However, since then, Einstein's general relativity has appeared on the scene. According to this theory, only local time can be directly measured with a clock. In other words, it is the *proper time* of this clock or an observer in the immediate vicinity that is measured. The time which now renders the equations of motion simple (or rather, the least complicated), over an extended region of space including, for example, the Solar System, is just a coordi-

nate freely chosen for its virtues with respect to the problem at hand and to which no physical reality is attributed. It is the *coordinate time* in an arbitrarily chosen system of spacetime coordinates. We shall return to these questions in Chapter 3. Their increasing practical importance is related to progress in time standards. Let us just bear in mind for our present general purposes that theoretical relationships exist between the proper time of an observer and the various coordinate times that he or she might be called upon to define.

Consequently, time metrology consists in supplying practically applicable representations:

- for the unit of proper time used by any observer for local experiments and for observations (even when they refer to distant objects);
- for the various coordinate times useful to astronomers, geodesists and engineers involved in determining spacecraft orbits or, to put it briefly, all those requiring precise models in spaces which vastly exceed the confines of the laboratory.

As soon as we adopt a theoretical dynamical model that is satisfactory in the sense that no experiment has yet found fault with it, time metrology can be based either upon local measurements, or upon the experimental determination of a coordinate time. Although both methods are theoretically equivalent, they differ profoundly in their results, to such an extent that, in 1960, the question was raised as to whether it would not be better to base the unit of time on atomic properties (observed locally) and to leave the determination of time scales to astronomy. It is to this practical aspect that we shall now turn.

2.4 The two modes of time measurement

We no longer distinguish local experiments from those involving large regions of space, since we have accepted that there exists a suitably all-embracing theoretical framework. The following discussion can be understood, if desired, in the familiar context of classical physics and Newtonian dynamics. In the spirit of the present work, it refers to the basics of metrology: how to define the second, and how to define the time scale used as the ultimate time reference.

In this section, we shall investigate two modes of time metrology, depending on whether we start from a phenomenon that anyone might reproduce in their laboratory, or from a time scale given by the apparent motions of the celestial bodies.

2.4.1 Use of reproducibility

The phenomenon taken as the standard must be defined in such a way that its duration can be considered as ideally constant. This means that we must be able to specify strictly the same causes, or else that we are in a position to eliminate any cause that could be viewed as a perturbation. The second, defined as an agreed number of standard durations, then has an ideally constant duration.

Naturally, the seconds actually achieved differ from ideal seconds. These differences can be reduced up to a point by time averages. But since they are not independent for successive seconds, and they do not average to zero, an optimal averaging period exists beyond which nothing more can be gained. (We shall return to these statistical problems when discussing the stability of oscillators in Chapter 5.) It is also possible to average over the seconds realised in several laboratories.

Once these averages have been taken, the resulting second suffers uncertainties from two sources:

- experimental uncertainties, in each laboratory, that are usually impossible to deal with, since the experiment itself occurred in the past and the only possible corrections are those concerning processing errors recognised subsequently, which rarely happens;
- uncertainties due to the way averages have been carried out, including those due to algorithms and comparison between standards (e.g., problems of statistical weighting).

In any case, these uncertainties remain bounded. Over the years, they gradually decrease, thanks to technical progress with instrumentation and a better understanding of the basic phenomenon which means that it can be purified by eliminating the causes of perturbation.

Let us now consider how a time scale can be made. Such a scale is obtained by cumulating seconds, in what is sometimes called an *integrated time scale*. But as the seconds accumulate, so also do their errors. The potential discrepancy between the scale realised and the ideal scale, made by cumulating ideal seconds, is always on the increase. Indeed, it can soon take on embarrassing proportions. As an example, the second was achieved in 2000 with an uncertainty of the order of 1×10^{-15} s, which may seem quite insignificant. However, it follows that the integrated time scale may deviate by one microsecond in 30 years when compared with the ideal scale that would be obtained by cumulating rigorously exact seconds. For certain applications, such a discrepancy cannot be disregarded. For example, we shall see later that the arrival of pulses from a pulsar can be timed to within 0.3 μs. The discrepancy between

the practical time scale and ideal time may still limit the work that can be done on the rotation of these objects.

Let us also note two consequences of choosing an integrated time scale as the world standard.

- In order to ensure the durability of the scale and make the best use of the existing standards for duration, averages must be taken. In other words, to use the technical expression, we must set up algorithms for the time scales. Various algorithms provide mutually diverging time scales, and there is no real way of saying which is the best. As the standard time scale must be unique, one must be chosen by convention between the results obtained by different institutions.
- To err is human, and this kind of error is also integrated. It may arise from experimental perturbations or comparison of standards, and affect transmission and centralised processing of data. As practical considerations make it impossible to modify the standard scale retrospectively (all the dates attributed to past events would have to be adjusted!), such integrated errors remain forever impressed upon history.

It was therefore with good reason that the idea of adopting an integrated time scale for world time was not very popular around 1960, when the first experimental scales of atomic time began to be established, based on the reproducibility postulate applied to frequencies of atomic transitions. But the advantages of atomic time over astronomical times proved to be decisive. In 1971, atomic time became official time, sixteen years after the first caesium atomic time standard had become operational. Despite the impatience of the atomic timekeepers, these sixteen years were, all things considered, a short enough period indeed for such a major development.

2.4.2 Use of a dynamical model

The main apparent motion of the celestial bodies is their daily path across the sky due to the Earth's rotation. However, the Earth's rotational period, although varying only slightly, is not strictly reproducible. Qualitatively, the causes of these irregularities are known, but this does not mean that we have been able to establish a satisfactory dynamical model. We shall return to this in Chapter 8 when discussing Universal Time. For the moment, let us consider the dynamics of the Solar System.

What the theory achieves is an ephemeris, giving the spatial coordinates of the celestial bodies (planets and their satellites) in terms of a time argument,

the coordinate time of the theory. For our present purposes, the spatial coordinates must be geocentric, in a non-rotating system, so that the observer need only apply small corrections depending on his or her position and velocity, to account for the Earth's orbital motion and rotation.

The measurement of time is then very simple, at least in principle. The coordinates of any body in the Solar System are measured at some moment which we time with a clock. The ephemeris then gives the date corresponding to these coordinates, expressed by its temporal argument. We thus have the time from the theory in the form of a correction to the reading of the clock. The operation is repeated as often as desired (or possible), the clock serving only to interpolate between the astronomical observations.

The primary data is therefore the time scale. The unit of time can be taken as the running unit interval of the graduation of this scale. To provide for possible revisions of the ephemeris, we may also define the unit of time as a unique time interval between two specified astronomical events. In this case, it is simply one of the two constants of integration in the equations of motion (the other fixing the origin from which time is counted) and only a physical realisation is provided by the ephemeris.

But needless to say, the way is long from principle to practice. If we really hope to achieve a unique time scale, a great many conditions must be satisfied. Measurements must refer to a single astronomical object, e.g., the Sun, the Moon or some other. One particular ephemeris must serve as reference by convention. The spatial frame of reference (which is a realisation of the spatial coordinate system) must also be fixed by convention. The various measurements, referred to a single clock, must be analysed and combined by an organisation in charge of giving the official time.

Having taken all these precautions, we must raise the question of errors affecting the scale produced, that is, its deviation from the ideal time of the theory, apart from a linear function of time that corresponds to the arbitrary choice of a unit and origin for time.

- Observational errors concerning the date are bounded (in contrast to what happens with integrated times). They can be reduced by technical progress. We note that it is preferable to choose a celestial body with rapid apparent motion.
- The dynamical model, currently general relativity, may be questioned.
- Application of the model incurs errors such as those which arise when series expansions are truncated or numerical integrations carried out. But it should be noted that, since observations are permanently acquired, it is possible to

improve the theory and its application in such a way as to obtain a revised time scale.

- Finally, and of the utmost importance, the ephemeris is affected by errors in initial conditions for the integration, that is, errors in the values taken for positions and velocities.

The last in this list raises a question of logic. How can we obtain velocities if we do not first establish a way of measuring time? This question was what most preoccupied the eminent astronomers who set up *Ephemeris Time* during the first half of the twentieth century. Without going into the complexity of their work, we may say that the main reason for their success is due to the lucky circumstance that the period of Earth's rotation is not such a bad standard for duration (involving relative variations of 10^{-7} to 10^{-8}). Thanks to this good fortune, the equations of motion could be established and initial velocities rather accurately measured. Moreover, the orbital motions are almost periodic so that, as far as time measurements are concerned, the theory involves small terms whose imperfections have little consequence for the final result.

In any case, the dynamical measurement of time rests in part on past observations whose uncertainties are then carried over to the future. The advantage of providing a single scale, as soon as a reference ephemeris has been chosen, is somewhat tempered by the need to revise the ephemeris as past observations accumulate. We shall see, in Section 4.3 on Ephemeris Time, that other difficulties have prevented this measure of time from being adopted as the practical world time standard.

Notwithstanding, realisations of time based on dynamics and their long term comparison with others based on atomic properties remain a test for both dynamical and atomic theories. For example, it has been asked whether the gravitational 'constant' is really constant.

3

Time measurement and theoretical models

At the beginning of the 1960s, fifty years after the publishing of Einstein's general relativity, time measurement and indeed the whole of metrology still made no reference to it. Of course, some experiments were aimed at putting general relativity to the test, but in the words of C. Will [3.1]: 'the attitude toward the theory seemed to be that, whereas it was undoubtedly of importance as a fundamental theory of nature, its observational contacts were limited to the classical tests and cosmology. As a consequence, general relativity was cut off from the main stream of physics.' Only the predictions of special relativity were commonly taken into account.

However, at this time, interest in relativistic theories was rekindled and began to grow rapidly. The reason can be found in the development of theoretical astrophysics and cosmology, supported by observational progress. We shall see that this experimental progress was largely due to the proliferation of techniques involving time and frequency. At the same time, therefore, there was a real need to fit time measurement into a more satisfactory theoretical framework than the one provided by classical mechanics and special relativity.

However, whilst a great many research scientists and engineers must now take relativistic effects into account in their work, general relativity has not for this reason become more familiar as a tool. In truth, communication between metrology and general relativity has always been difficult. The real cause for this barrier undoubtedly lies in the fact that general relativity is based on differential geometry, whilst metrology deals with finite quantities. The application of general relativity to time measurement and, more generally, to metrology and macroscopic physics thus brings us to the fundamental principles of the theory. This is what we shall examine now, although rather briefly. We shall review the way the concepts of space and time have evolved from the Newtonian model to that proposed by general relativity. The reader can go further with these subjects by consulting, for example, [3.2]. Another useful source is

the report [3.3] which discusses the application of general relativity to metrology, and which inspired parts of the present chapter and Chapter 5. Those who find such considerations disheartening or tiresome are advised to move directly to the summary at the end of this chapter, where they will find the main things they need to know for the purposes of the present book.

The relativistic treatment of time measurement involves spatial coordinates. In this chapter, we shall define the necessary celestial and terrestrial coordinate systems. Physical realisations of these systems, in the form of numerically specified coordinates for selected objects, are known as *reference frames*. Methods for constructing such frames will be examined in Chapter 8. For the time being, we shall indicate their accuracy and official status.

3.1 The Newtonian model and absolute time

Let us begin with a strict definition of an *event*. This is a geometrical object, namely, a point in spacetime which exists in itself. The first contact between Armstrong's boot and the lunar surface is an example of an event, insofar as we may neglect the area of his sole and the time required to crush down the upper layers of dust. An event can be identified by a description, as in the present example, and this emphasises its intrinsic nature. However, it is often useful to attribute a numerical code to it, namely, its four spacetime *coordinates*, within a *coordinate system*. The latter is chosen specifically because it is the most suitable for the aim at hand. In the example cited above, the archives conserve the *date* (a time coordinate) in the *Coordinated Universal Time* scale, whilst the television viewer would have been more interested in the official local time. Selenocentric spatial coordinates would be the most suitable for describing the place where the footprints could be seen, but the engineers at NASA must have followed Apollo's trajectory in a geocentric system.

Coordinates are therefore an essential auxiliary tool, but they serve only to label events, rather than to define them.

In the Newtonian model of spacetime, it is assumed that there exist subsets of *simultaneous events*. Simultaneity is viewed as a physical property that cannot be influenced in any way. A particular cross section of spacetime can be identified by the date t_U in a time scale U, or by a date t_V in another scale V, but the relation $t_V = t_V(t_U)$ is a function of t_U alone. In each cross section of simultaneity, spatial coordinates can be freely chosen, within the framework of Euclidean geometry.

One consequence of immediate importance in the present context is that the time interval between two events A and B is viewed as an intrinsic property of that pair of events. As soon as the temporal coordinate axis has been chosen,

it is unambiguously expressed as the difference between the dates of A and B, and no further information need be supplied. This would seem to be an obvious truth, and yet extremely precise measurements, now common practice in physics, belie this beautiful property of the Newtonian universe.

Putting together the simultaneous sections of this Newtonian spacetime, we now require a dynamical model. It is based on the well known principles of Newtonian mechanics. We shall restrict ourselves to Newton's first law, as this will be enough to understand what such a model can tell us.

According to this principle, an isolated material point moves in a straight line with uniform speed. Needless to say, this is an idealisation from experiments or observations made in a rough and ready manner. The isolated point can be represented by a marble on a horizontal plane surface, time can be measured by a clock based on the notion of reproducibility (e.g., a pendulum), lengths can be measured with a rigid rod, and we observe that the motion is approximately uniform. But what does the first law really tell us? Does it define the isolated point or the uniform velocity? What measurement standards can be used to assess this uniformity? We must accept that such questions are posed in vain. The law merely tells us to adopt 'good' coordinates, such that the spatial coordinates x^i $(i = 1, 2, 3)$ are linear functions of the time coordinate t. The Newtonian model, like any model, is a construction of the human mind. It is logically developed from a minimal number of postulates, viz., Newton's laws of mechanics. It is motivated by experience, but once created, lives independently of it. It is the physicist's task to construct the coordinates, and the measurement standards for time, length, mass and angle in accordance with the theory. With these constructions, the physicist must then check that all measured phenomena correspond, within the limits of experimental uncertainty, to the theoretical predictions based upon the model. If any disagreement should occur, either the coordinate systems and measurement standards were poorly chosen, or else the theory has been incorrectly applied (by omission of perturbing effects, calculation errors, and so on), or even, as a last resort, the theory itself is unsatisfactory.

Neither the first law of motion nor the other laws of Newtonian dynamics are affected by the coordinate transformation

$$\begin{aligned} t &\longrightarrow t' = t \,, \\ x^i &\longrightarrow x'^i = x^i + v^i t \,, \end{aligned} \tag{3.1}$$

where the v^i are three arbitrary constants. (They are also unaffected by the trivial changes of unit and origin for the axes.) Whereas absolute time is an essential element, absolute space is clearly an unnecessary notion in this dy-

namical model. The set of spatial coordinate systems defined by the transformation (3.1) constitute the Galilean coordinate systems and the invariance of the dynamical laws in these systems is the Galilean principle of relativity.

We must stress the fact that the (spatio-temporal) coordinate system is *defined* by the equations of motion. It is the simplest form of these equations that leads to Galilean coordinate systems and absolute time. To define, as is sometimes done, the axes of the Galilean system by the fact that they remain fixed relative to the stars or quasars is not correct. This is not a definition, but an approximate observed feature. If we say that the spatial axes and absolute time are graduated in metres and seconds, in accordance with the definitions of these quantities as supplied by the SI system, we are in no way defining the graduation of these axes. This is merely an assumed property which is approximately corroborated by experience.

3.2 Special relativity

3.2.1 Physics in inertial systems

Unlike dynamics, both optics and electrodynamics seemed to be in a position to identify an absolute space through measurements of the speed of light. The crisis that was to lead to the abandonment of absolute space began with experiments by Fizeau (1849) and Foucault (1850), reaching its climax with Maxwell's electrodynamics (1869) and attempts to measure the absolute orbital velocity of the Earth by Michelson (1881), then Michelson and Morley (1887). Maxwell's equations are not invariant under the Galilean transformation (3.1) and imply that light has constant speed in absolute space. However, observations showed that all observers drag absolute space along in their stead.

Einstein's special relativity (1905) resolved this crisis. In his theory, spatial reference frames are still Galilean frames, or *inertial frames*, considered to be removed from all gravitational fields. These frames all have uniform rectilinear motions relative to one another. In the following, we shall not attempt to describe special relativity starting from a minimal number of postulates, since this would lead to an inappropriate level of subtlety for the present purposes. We review in as simple a manner as possible the principles and consequences of the theory, stressing those features most relevant to metrology. The price to be paid for these simplifications is a certain redundancy and circularity.

(a) Invariance of the speed of light

For each inertial frame, the speed of light in vacuum c is the maximal speed of any signal or particle. This speed is isotropic and is the same in all frames.

We come back to the problem of measurement standards which already worried us when we considered Newton's first law. As far as time is concerned, we assume that clocks based on reproducibility in the way discussed in Chapter 2 are qualified to check directly or indirectly that c is constant. This is the case for an atomic clock producing the SI second. As far as length is concerned, the standard was, historically, a metal bar. But in 1960, the metre was defined as a certain multiple of the wavelength λ_0 of an atomic transition of krypton. It was then implicitly assumed that the frequency ν_0 of this transition and the speed of light were constant. Any attempt to check the constancy of c using atomic clocks and metres defined in this way would be quite pointless. The 17th General Conference on Weights and Measures (1983) explicitly based the definition of the metre on the constancy of c (see Section 4.4).

(b) The geometry of space

In each inertial frame, space is Euclidean. The *distance* D between two fixed points in a particular frame is a well defined and measurable quantity. It is convenient to use orthonormal Cartesian coordinates x^i such that

$$D = \left[(\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2\right]^{1/2} , \tag{3.2}$$

where Δx^i are the coordinate differences between the two points. In such conditions, coordinates are themselves directly measurable quantities using measurement standards, for example, according to the procedure described in (d) below.

(c) Time coordinate and synchronisation

In order to define the time coordinate t for an inertial frame, we may appeal to a system of clocks fixed within the frame. By definition, what we read on a clock C is the *proper time* τ_C of this particular clock. Let us consider two fixed clocks C and D, distant from one another. A light signal is emitted by C at time $\tau_C(\text{emit})$ as recorded by C. It is then received and reflected by D at time $\tau_D(\text{reflect})$, as recorded by D, and returns to C at time $\tau_C(\text{return})$, as recorded by C. According to *Einstein's synchronisation convention*, C and D are synchronised if

$$\tau_D(\text{reflect}) = \frac{\tau_C(\text{emit}) + \tau_C(\text{return})}{2} .$$

By principles (a) and (b), the fact of being synchronised according to Einstein's convention is a transitive relation in an inertial frame. This means that, if C and D are synchronised, and if also D and E are synchronised, then so also are C and E. It is therefore possible to synchronise all clocks fixed in a given

inertial frame and to adopt the time τ of one of them (its *proper time*) as the time coordinate t.

(d) Spatial coordinates

Accepting the postulate that light has constant speed c, and also the present definition of the metre, the distance between two points A and B fixed relative to an inertial frame can be unambiguously measured from the time Δt required by a light signal to cross the vacuum from A to B. The spatial axes are graduated by appealing to the propagation of light along these axes (this propagation occurring in straight lines).

(e) Einstein's principle of equivalence

We may now state a fundamental principle of physics, which says that no non-gravitational phenomenon can be used to distinguish one inertial frame from another. In other words, mathematical models of (non-gravitational) physics have the same form in all inertial frames when we use identical measurement standards, fixed in these frames. This implies that the constants of physics are invariant when they are measured using these standards. This principle, stated in different ways, is often called *Einstein's principle of equivalence.*

3.2.2 Lorentz transformation and invariance of the interval

The principles stated above are incompatible with the Galilean transformation (3.1) between the coordinates of two inertial frames. They require the Lorentz transformation. If the system R′ is in translational motion with respect to the system R, at some constant speed v as measured in R, along the x^1 axis of R, this transformation is given by

$$\begin{aligned} t &\longrightarrow t' = \beta(t - vc^{-2}x^1) \,, \\ x^1 &\longrightarrow x'^1 = \beta(x^1 - vt) \,, \\ x^2 &\longrightarrow x'^2 = x^2 \,, \\ x^3 &\longrightarrow x'^3 = x^3 \,, \end{aligned} \tag{3.3}$$

where $\beta = [1 - (v/c)^2]^{-1/2}$ and the origins coincide at $t = 0 = t'$.

The Lorentz transformation guarantees invariance of c (this is straightforward enough to show) and Maxwell's equations. However, it should not be considered as a mathematical artefact designed to this end. It reflects a fundamental property of nature and its various consequences have never been contradicted even by the most accurate experiments.

Moreover, if Δt and Δx^i are the coordinate differences between two events A and B, the quantity Δs^2 defined by

$$\Delta s^2 = -c^2\Delta t^2 + \sum_i \left(\Delta x^i\right)^2 \tag{3.4}$$

is invariant under the change of coordinates (3.3). This quantity Δs is called the *interval* between the two events. Equation (3.4) gives the spacetime *metric* for special relativity, known as the *Minkowski metric*. It shows how the interval is a quantity associated with a pair of events that can be measured without referring to any particular system of coordinates. Of course, in R, for example, we can measure the temporal and spatial coordinates, deducing Δt and, by (3.2), the value of D. But these quantities have no universal reality, depending as they do on the choice of coordinates. They are said to be *coordinate quantities*.

In each inertial frame, the propagation of light is characterised by $\Delta s^2 = 0$, its speed being c.

3.2.3 Time in special relativity

The unit of time used in an inertial frame must be supplied by a standard that is fixed relative to the frame. This is summarised by saying that the second must be viewed as the *unit of proper time* (understanding that we mean the proper time of the frame in which it is used).

The synchronisation convention involves clocks fixed relative to some coordinate system. It depends on the frame chosen, in the sense that clocks synchronised within R will not be synchronised within R′. As is well known, this leads to the famous apparent time dilation observed in moving clocks. In order to remove the slightest ambiguity, it is useful to specify exactly what the quantities appearing in (3.3) are intended to mean. In these equations, t and t', like x^i and x'^i, are the coordinates of one and the same event (existing in itself, independently of the means used to record it). They are therefore *readings* taken along the coordinate axes, in such a way that (3.3) expresses the relations between readings relative to the two different coordinate systems.

As far as time is concerned, we have seen that t can be considered as equivalent to the readings τ of synchronised clocks fixed relative to R. We may likewise imagine a large number of clocks fixed relative to R′, physically identical to those in R, whose readings τ' are equal to t'.

Two successive readings τ_1 and τ_2 of a clock fixed in R correspond to two events for which the time coordinates t_1 and t_2 in R, and t'_1 and t'_2 in R′, satisfy

$$t_2 - t_1 = \tau_2 - \tau_1 \,, \qquad t'_2 - t'_1 = \beta(t_2 - t_1) = \tau'_2 - \tau'_1 \,. \tag{3.5}$$

Since $\beta > 1$, the observer fixed relative to R′ concludes that the proper time of the clock in R, which is in relative motion, increases less quickly than the proper time of his or her own clock.

In this argument, there has been no question of sending signals between R and R′. However, if the observer in R′ receives the frequency of the clock in R, moving at relative velocity $\mathbf{v}$ (with magnitude v), via an electromagnetic signal, then he or she must take the classical Doppler effect into account. The relation between the emitted frequency f, measured in R, and the received frequency f', measured in R′, is

$$f' = f\left[1 + \frac{\mathbf{v}\cdot\mathbf{n}}{c}\right](1 - v^2/c^2)^{-1/2}\,, \tag{3.6}$$

where $\mathbf{n}$ is the unit vector along the (straight) path of the signal, pointing in the direction of propagation.

In addition, let us consider two events A and B such that a clock can be transported at constant speed along a straight line from A to B. In the coordinate system of the frame R attached to this clock, since t is equal to τ,

$$\Delta s^2 = -c^2\Delta t^2 = -c^2\Delta\tau^2\,. \tag{3.7}$$

In an arbitrary inertial frame R′, the proper time of a clock transported in this way from one event to another can be used to measure the interval between such events. It is related to the primed coordinates by

$$\Delta\tau^2 = \Delta t'^2 - \frac{1}{c^2}\sum_i(\Delta x'^i)^2\,. \tag{3.8}$$

3.2.4 Field of application of special relativity

Special relativity may seem to be of very limited interest since it excludes gravity and hence, strictly speaking, any distribution of mass. However, it is of key importance in physics and metrology. One reason is that we can achieve an approximation to an inertial frame through a freely falling laboratory. Another is that the theory can even be applied in an accelerated laboratory. In both cases, the gravitational interaction between masses inside the laboratory, arising from equipment or its users, is so small that it can always be neglected, except of course in experiments specifically designed to study this interaction (e.g., a measurement of the gravitational constant).

In a freely falling laboratory, the acceleration almost completely compensates for gravitational effects. The only remaining effects are those due to the lack of uniformity of the gravitational field (because of the finite distance to the masses generating it). The residual potential is called the tidal potential, by

analogy with the potential that generates tidal effects in the terrestrial oceans and landmasses, whilst the Earth is freely falling in the gravitational field of all the celestial bodies (which is mainly due to the Sun and the Moon). It gives rise to effects of the order of 10^{-16} per metre, in relative values, close to the Earth's surface. It is not negligible for the best atomic frequency standards in space laboratories. In less stringent applications, such a laboratory is locally a good approximation to an inertial frame, provided it is not rotating.†

In an accelerated, rotating laboratory, such as a laboratory on Earth, subject to the acceleration due to gravity and carried along by the Earth's rotation, special relativity can still be applied. For this to be possible, the difference in the Earth's gravitational potential between various points in the laboratory must have negligible effects. This is commonly the case in a typical room. (However, for time measurements, the gravitational frequency shift is about 10^{-16} per metre of altitude, in relative values, and we are coming to the point where the very spatial extent of an atomic clock could no longer be treated as 'local'.) In addition, inertial effects such as weight, the Coriolis force, and so on, must be taken into account in a classical manner. As an example, we shall see that certain atomic clocks use a beam of atoms in a fixed cavity in the laboratory. In that case, we must take into account the curvature of the beam due to the weight of its constituent particles, and possibly other classical weight effects. But, in local experiments, it would be a mistake to take into account relativistic effects due to the presence of the Earth's gravitational field, as long as it may be considered uniform. It should be remembered that photon trajectories are only straight lines in an inertial frame (a property used by laser gyrometers).

3.3 General relativity

3.3.1 Overview

General relativity is a complex theory, especially with regard to its developments in astronomy and cosmology. This is in spite of the fact that it is founded on very simple principles, which will fortunately be adequate for the metrological problems treated in this book.

Whereas special relativity applies, in principle, to an empty and therefore unreal universe, general relativity is a model of the real world, where mass and energy find their proper place. It is a theory of gravitation. However,

† The problem of rotation is complex in relativity. However, it is not of critical importance in time measurement. In the present book, we adopt a kinematic definition for the condition of non-rotation, viz., it is the absence of apparent rotation relative to the most distant known objects, the quasars.

gravitational effects are not viewed as an action at a distance, in the way proposed by the 'frozen' framework of Newtonian spacetime. Rather, they appear as local geometrical properties of a Riemannian spacetime, curved by the presence of mass and energy.

In general relativity, there are no theoretically privileged coordinate systems (although there are systems that are more convenient than others, in the sense that they lead to equations that are easier to handle). Let x^μ, $\mu = 0, 1, 2, 3$, be the chosen coordinates, without worrying for the moment about their meaning in terms of space and time. The spacetime is equipped with a metric relating coordinate differences between any two infinitely close events, with coordinates x^μ and $x^\mu + \mathrm{d}x^\mu$, say, to a quantity $\mathrm{d}s$ called the *interval* as in special relativity, except that now the interval is infinitesimal. The relation is

$$\mathrm{d}s^2 = g_{\alpha\beta}(x^\mu)\,\mathrm{d}x^\alpha\,\mathrm{d}x^\beta\ , \tag{3.9}$$

where we apply the summation convention on repeated indices and $g_{\alpha\beta} = g_{\beta\alpha}$, so that there are ten independent components $g_{\alpha\beta}$. These $g_{\alpha\beta}$ are the components of the *metric tensor* and they are functions of the four coordinates x^μ. Hence, a coordinate change leaves the scalar $\mathrm{d}s^2$ invariant and it is considered to be a measurable quantity using the usual means, that is, rods and clocks.

In the Newtonian model and also in special relativity, the 'good' coordinates were physical quantities that could be measured anywhere and at any time without ambiguity. The coordinates in general relativity never have such physical properties globally, whatever we may do. This is a key difference for metrology. In order to emphasise this fact, many authors consider coordinates simply as numbers (i.e., without dimensions, in the metrological sense), rather like telephone numbers [3.4] or street numbers. But in practice, when actually making measurements, this viewpoint is difficult to implement. In fact, there is no entirely satisfactory solution to the problem of dimension and unit for coordinates in metrology [3.3].

Since the Minkowski metric has proven to be wholly successful in local experiments, there must exist, for each event x^α, a change of coordinates $x^\mu \to x'^\mu$ such that the metric takes on a form analogous to (3.4) with $x'^0 = ct'$, viz.,

$$\mathrm{d}s^2 = -c^2(\mathrm{d}t')^2 + (\mathrm{d}x'^1)^2 + (\mathrm{d}x'^2)^2 + (\mathrm{d}x'^3)^2\ . \tag{3.10}$$

The expression in terms of differentials reminds us that it is an approximation to apply special relativity to a space of finite extent. Mathematically, this change of coordinates is made possible by the symmetry of the $g_{\alpha\beta}$, together with a condition on the eigenvalues of the matrix they make up. In fact, there are infinitely many primed coordinate systems which solve the above problem,

since the resulting local coordinate systems may have any orientation and velocity.

The form of (3.10) brings out an extremely important feature for time measurement. The spacetime trajectory or *worldline* of a clock, with proper time denoted by τ, is described by the successive values of all four coordinates. In an infinitesimal displacement of the clock, a local coordinate system of the form given by (3.10) can be attached to it, in such a way that the dx'^i are zero, for $i = 1, 2, 3$. Then in a similar way to (3.7), we have

$$d\tau^2 = -c^{-2}\, ds^2 \,. \tag{3.11}$$

If finitely separated events A and B are such that a clock can be transported from A to B, the increase $\Delta\tau$ in proper time recorded by the clock will be given by

$$\Delta\tau = c^{-1} \int_C \left| g_{\alpha\beta}\, dx^\alpha\, dx^\beta \right|^{1/2} \,, \tag{3.12}$$

the integral being evaluated along the worldline C of this clock. The quantity $\Delta\tau$ is therefore measurable, although it does not express an intrinsic or objective property of the pair of events, since various choices for C may be available.

The $g_{\alpha\beta}$ are related to the distribution of matter and energy by Einstein's field equation, a tensorial equation which plays a similar role to Poisson's equation in Newtonian dynamics. Expressed in coordinate form, Einstein's equation provides six independent differential equations for the determination of the ten independent $g_{\alpha\beta}$. This underdetermination expresses the fact that we may freely choose the four coordinates x^μ, by arbitrary coordinate transformation from any specific solution.

Concerning particles with negligible mass that are not subject to any forces other than gravitational effects (referred to as *free particles* in the present context), there is a further postulate which states that they follow spacetime *geodesics*. These are worldlines G for which the quantity

$$l = \int_G \left| g_{\alpha\beta}\, dx^\alpha\, dx^\beta \right|^{1/2} \tag{3.13}$$

is extremal. General relativity is thus a gravitational theory well suited to describing the motions of planets, for example. The worldlines of photons are characterised by the fact that ds^2 is always zero.

General relativity leads to many predictions, especially in astrophysics and cosmology, and for the time being, none has been contradicted by observation. Since Einstein's work, other relativistic theories have been invented and

several of them have passed the same observational tests (see Chapter 9). However, none of these theories has turned out to be indispensable [3.1]. For our metrological purposes, modest as they are, we shall appeal to general relativity because it is the simplest model whose predictions agree with measurements, within the bounds of uncertainty in the latter.

3.3.2 *Post-Newtonian approximation*

(a) General form of the metric

We shall now make use of the freedom left to us by the theory to choose the most suitable coordinates. We shall also make assumptions compatible with the best time measurements yet achieved. Although we are only concerned with metrology and dynamics within the Solar System, it will not be convenient to cover this region of the Universe with a single coordinate system. We shall require at least two systems which are not rotating in space, one with spatial origin at the barycentre of the Solar System, the *barycentric system*, and the other with its origin at the centre of mass of the Earth, the *geocentric system*. We shall also need a geocentric system rotating with the Earth. We now describe the way the non-rotating systems were defined in 1991 by the International Astronomical Union (IAU) [3.5].

The IAU recommends that spacetime coordinates $x^0 = ct$, x^1, x^2, x^3 be chosen in such a way that in each coordinate system, centred on the barycentre of any system of masses, the square ds^2 of the interval should be expressed to the lowest level of approximation in the form

$$\begin{aligned} ds^2 &= -c^2\,d\tau^2 \\ &= -\left(1-\frac{2U}{c^2}\right)(dx^0)^2 + \left(1+\frac{2U}{c^2}\right)\left[(dx^1)^2+(dx^2)^2+(dx^3)^2\right], \end{aligned} \tag{3.14}$$

where c is the speed of light, τ the proper time and U the sum of the Newtonian gravitational potentials for the system of masses under consideration (taken as zero at infinity) and a Newtonian tidal potential generated by bodies outside the system, this potential being expressed in such a way that it vanishes at the barycentre. By convention, the potential is taken as positive.

The IAU also stipulates that the measurement units to be used are the SI second and the SI metre, for the proper time and proper length (i.e., to measure ds^2).

The metric (3.14) is an approximate solution to Einstein's equations. However, it has become inadequate for establishing precise planetary ephemerides.

At the General Assembly in 2000, the IAU therefore recommended adoption of *harmonic coordinates*, using them to express the metric to higher order terms in $1/c$. It also provided a relativistic expression for the relevant (scalar and vector) potentials. The IAU also developed the application of this extended metric to time. However, in the present book, which is mainly concerned with time measurement on and near the Earth, the metric (3.14) provides an adequate approximation to the new one. In the geocentric system, it generates relative frequency errors of at most 10^{-18}, out to 300 000 km from the Earth, and these are much smaller than current uncertainties in frequency standards. The metric (3.14) will thus be used in the following developments.

The ratio U/c^2 is always small compared with unity for our present preoccupations. On the Earth's surface, for example, it takes values around 1.5×10^{-8} in the barycentric system and 0.7×10^{-9} in the geocentric system. Space is quasi-Euclidean and the spatial coordinates are very close to classical Cartesian coordinates. The coordinate $t = x^0/c$ is close to the proper time of clocks moving at slow velocities compared with c in the coordinate systems under consideration. It is referred to as *barycentric* or *geocentric coordinate time*.

Coordinates defined in this way clearly have dimensions of length and time. The practice in metrology is to attribute a single unit, the SI unit, to all quantities having the same dimension. The distinction between these quantities must then by made through their definitions and not through the use of different units. We shall follow this practice, expressing spatial and temporal coordinates in metres and seconds. This has the disadvantage of causing some confusion between *proper quantities*, which are directly measurable using standards, and *coordinate quantities* (the coordinates themselves or quantities constructed from them) for which the graduation scale varies relative to the proper units, depending on the place and time. It is essential to bear in mind this difference between such quantities. Whenever necessary, the unit on the coordinate graduation scale will be referred to as the *scale unit*.

We shall frequently refer to the relation between the proper time of a clock and the coordinate time. It is based on the expression for the ratio between $\mathrm{d}\tau$ and $\mathrm{d}t$. We may set

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = 1 - h(t, \mathbf{x}) \,, \tag{3.15}$$

where $\mathbf{x}$ represents the triplet x^i $(i = 1, 2, 3)$, as a reminder that this relation depends on all four coordinates of the clock. On a given worldline, $\mathbf{x} = \mathbf{x}(t)$, and we may write, simplifying the notation,

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = 1 - h(t) \,. \tag{3.16}$$

The increment $\Delta(t - \tau)$ during a coordinate time interval $(t_0,\ t)$ is given by

$$\Delta(t - \tau) = \int_{t_0}^{t} h(t)\,\mathrm{d}t\ . \tag{3.17}$$

Using the metric (3.14),

$$h(t) = c^{-2}\left[U(t) + \frac{1}{2}v(t)^2\right] + O(c^{-4})\ , \tag{3.18}$$

where v is the coordinate speed of the clock, defined by

$$v = \frac{\left[(\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2\right]^{1/2}}{\mathrm{d}t}\ , \tag{3.19}$$

and $U(t)$ is the value of the gravitational potential at the location of the clock. The terms of order $O(c^{-4})$ in (3.18) are negligible for our present purposes and will no longer be mentioned. The potential U and coordinates x^i, $i = 1, 2, 3$, are functions of t along the worldline of the clock. In the integral, t can be replaced by τ, since this involves an error of order $O(c^{-4})$.

(b) Coordinate systems and their realisation

In (3.14), x^α represents coordinates in any coordinate system. To avoid confusion, a different notation will be used for each system. The notation $t = x^0/c$, x^i will be reserved for non-rotating geocentric coordinates, and $t_\mathrm{B} = u^0/c$, u^i will be used for barycentric coordinates.

The spatial axes of the barycentric coordinate system recommended by the IAU are centred upon the barycentre of the Solar System. They have fixed directions relative to the directions of the compact extragalactic sources known as *quasars*, with u^3 pointing roughly along the axis of rotation of the Earth on 1 January 2000 and u^1 roughly towards the equinox for this date.

The spatial axes of the non-rotating geocentric coordinate system are centred upon the centre of mass of the Earth (including its fluid envelopes). They have the same directions as those of the barycentric system. In the metric, expressed in the form (3.14), U is then the gravitational potential of the Earth, to which must be added the Sun–Moon tidal potential.

The coordinate transformations between the barycentric and geocentric coordinate systems are complicated when treated rigorously [3.6]. However, such a degree of rigour will not be necessary here.

We shall also need coordinate system X^i rotating with the Earth. It is attached globally to the lithosphere, taking into account a model for the motion of the tectonic plates (a few centimetres per year). The X^3 axis points in a direction close to the Earth's axis of rotation, and the X^1 axis defines the origin

for longitudes, in such a way that the longitude at Greenwich is roughly zero. This system derives from the non-rotating geocentric system by a spatial rotation that takes into account the motions of the Earth's axis of rotation both in space and in the Earth itself. The geocentric coordinate time t is the same in the rotating and non-rotating systems.

In the rotating geocentric system, the metric (3.14) takes the form

$$\mathrm{d}s^2 = -\left(1 - \frac{2U}{c^2}\right) c^2\,\mathrm{d}t^2 + \left(1 + \frac{2U}{c^2}\right)\left[\mathrm{d}r^2 + r^2\,\mathrm{d}\phi^2 + r^2\cos^2\phi\left(\omega^2\,\mathrm{d}t^2 + 2\omega\,\mathrm{d}t\,\mathrm{d}L + \mathrm{d}L^2\right)\right], \tag{3.20}$$

where

- ϕ is the geocentric latitude (the angle between the geocentric direction of the place and the plane of the equator),
- L is the longitude (positive towards the east),
- r is the geocentric coordinate distance

$$r = \left[(X^1)^2 + (X^2)^2 + (X^3)^2\right]^{1/2},$$

- ω is the angular speed of rotation of the Earth (the constant approximate value $\omega = 7.292\,115 \times 10^{-5}$ rad/s is good enough),
- U is the gravitational potential of the Earth plus the tidal potential, as in the non-rotating system.

We shall now consider how these coordinate systems may be set up in practical terms. Concerning time, the IAU has chosen an origin for t_{B} and t. Each of these scales is then uniquely determined, since the scale unit is determined by using SI units for the interval $\mathrm{d}s$. These scales are referred to as *Barycentric Coordinate Time* (TCB) and *Geocentric Coordinate Time* (TCG), distinguished from the corresponding purely theoretical concepts by the use of capital letters and the acronym. Physical realisations of TCB and TCG can be obtained from International Atomic Time in the way explained in Chapter 7.

As far as spatial coordinates are concerned, once the proper unit of length, the metre, has been chosen, the graduation of the coordinate axes is also determined.† It remains only to orient the axes.

† For their own specific reasons, astronomers also use the *astronomical unit* of length, approximately equal to 1.5×10^{11} m. Its relationship to the metre is determined experimentally. The astronomical unit (AU) must also be considered as a proper unit, serving as a basis for graduating spatial axes.

The barycentric coordinate system is realised physically by attributing invariable angular coordinates to the distant extragalactic sources known as quasars. By international convention, the *International Celestial Reference Frame* (ICRF) is used. It comprises a list of the right ascensions and declinations of about 600 quasars, published by the International Earth Rotation Service (IERS) [3.7]. Uncertainties in the coordinates range from $0.0002''$ to $0.0005''$. As already observed, the geocentric coordinate system is oriented in the same manner.

Concerning the spatial coordinates in the rotating geocentric system, the physical realisation adopted by convention is the *International Terrestrial Reference Frame* (ITRF). It comprises a list of coordinates at a fixed reference date, and the velocities of about 200 sites spread around the world. The uncertainty in the coordinate values is of the order of a centimetre [3.7]. Like the ICRF, the ITRF is established by the IERS. This service maintains and improves it on a permanent basis. It should also be noted that the ITRF is officially the primary geodetic standard to which all geodetic systems are gradually being related. The rotation of the ITRF relative to the ICRF is described by the orientation parameters of the Earth.

(c) Relation between geocentric coordinate time and proper time

In the non-rotating geocentric system, the relation between the proper time τ of a clock and the geocentric coordinate time t is given by (3.17) and (3.18).

In the coordinate system rotating with the Earth, the expression for $h(t)$ is

$$h(t) = c^{-2}\left[\hat{U}_{\mathrm{g}} + \Delta\hat{U}(t) + \frac{1}{2}V(t)^2\right] + 2c^{-2}\omega\frac{\mathrm{d}A_{\mathrm{E}}}{\mathrm{d}t}\,, \tag{3.21}$$

where

- $\hat{U}_{\mathrm{g}}$ is the potential U plus the potential of the axifugal force at the level of the geoid (the geoid is a total equipotential surface), with the value $\hat{U}_{\mathrm{g}} = 6.263\,686 \times 10^7\ \mathrm{m^2\,s^{-2}}$,
- $\Delta\hat{U}$ is the difference between the total potential (including that due to the axifugal force) and $\hat{U}_{\mathrm{g}}$ at the location of the clock,
- V is the coordinate speed of the clock relative to the Earth,
- A_{E} is the area of the projection onto the equatorial plane of the surface swept out by a vector extending from the centre of mass of the Earth to the clock, as measured in the rotating system and taken as positive when the clock moves towards the east.

Relations (3.18) and (3.21) are often used. We shall apply them, in particular, to compare times and frequencies of widely separated standards, and also to

set up a worldwide time reference. They are largely accurate enough for such purposes, given the uncertainties in present time standards, provided that U or $\Delta\hat{U}$ are correctly evaluated.

(d) Evaluating the Earth's gravitational potential

At the surface of the Earth, the lunisolar tidal potential leads to relative frequency shifts of the order of 10^{-17}, which remains negligible. It plays an increasing role as we move away from the Earth, leading to relative frequency shifts up to about 10^{-15} at the altitude of geostationary satellites (36 000 km). We shall not take them into account, even though it might be necessary for a clock carried aboard such a satellite. The interested reader is advised to consult [3.6].

Let us now consider the contribution made by the terrestrial potential U_T to U. Its evaluation leads metrologists into areas somewhat unfamiliar to them, and often causes them some difficulty.

At the distances typical of artificial Earth satellites, an expansion of the Earth's potential in spherical harmonics is required. Bounds on accuracy are determined by uncertainties in the satellite orbits. In the best of cases, they correspond to relative frequency inaccuracies between 10^{-17} and 10^{-18}.

Considering ground-based clocks, or those carried by aircraft, if relative frequency inaccuracies of the order of 10^{-14} can be tolerated, then it is enough to keep the first terms in the spherical harmonic expansion, taking

$$U_T = \frac{GM_E}{r} + \frac{J_2 GM_E a^2 (1 - 3\sin^2\phi)}{2r^3} , \tag{3.22}$$

where the symbols and their values (exact up to at least the last decimal place) are explained as follows:

- GM_E is the product of the gravitational constant and the mass of the Earth, with value

$$GM_E = 3.986\,004\,42 \times 10^{14}\ \mathrm{m^3\ s^{-2}} ;$$

- r is the geocentric coordinate distance in metres;
- J_2 is the coefficient of the quadrupole moment of the Earth, with value $J_2 = 1.082\,636 \times 10^{-3}$;
- a is the radius of the Earth at the equator, with value $a = 6\,378\,137$ m;
- ϕ is the geocentric latitude.

In (3.21), $\Delta\hat{U}$ is given by

$$\Delta\hat{U} = U - \hat{U}_{\mathrm{g}} + \frac{\omega^2 r^2 \cos^2\phi}{2} , \tag{3.23}$$

the last term representing the potential of the axifugal force.

If the altitude above the geoid h is known (not to be confused with the function h defined earlier), then $\Delta\hat{U}/c^2$ can be obtained from [3.8]

$$\frac{\Delta\hat{U}}{c^2} = -1.088\,21 \times 10^{-16} h - 5.77 \times 10^{-19} h \sin^2\phi + 1.716 \times 10^{-23} h^2 , \tag{3.24}$$

for h given in metres.

The inaccuracy is then less than 10^{-16} for altitudes below 15 km. However, care must be taken in evaluating h, since its definition may vary depending on the levelling networks and topographic maps used. It is safer to consult a geodesist. Reference [3.9] is useful in this context.

(e) Terrestrial Time: another geocentric coordinate time

By (3.21), if an ideal clock is fixed on the Earth at the level of the geoid, its proper time τ is related to t by

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = 1 - h(t) = 1 - \frac{\hat{U}_{\mathrm{g}}}{c^2} \approx 1 - 0.697 \times 10^{-9} . \tag{3.25}$$

It therefore lags by 22 ms per year relative to t. Higher up, the discrepancy diminishes, but only very slightly, by 3.4 μs per year and per kilometre, according to (3.24).

International Atomic Time (TAI) was first established quite spontaneously using atomic clock readings without any attempt to take relativistic effects into account. Its scale unit was therefore close to one second on the rotating geoid. When a relativistic definition of the world time scale became necessary during the 1970s, no one would have considered for a moment introducing a definition which meant that all Earth-based clocks would lag by 22 ms per year. A Terrestrial Time (TT) was thus defined, called Terrestrial Dynamical Time (TDT) from 1979 to 1991, which involves a frequency shift relative to the geocentric coordinate time, in such a way that the TT scale unit lasts one second of proper time on the rotating geoid. We shall see in Chapter 7 that TAI is a realisation of TT.

It was then noticed that the intricacy and temporal variation (e.g., tidal changes) inherent in the definition and realisation of the geoid are sources of uncertainty in the definition and realisation of TT, and hence also of TAI. In

order to avoid these difficulties, the IAU redefined TT in 2000, using the following terms: TT is a time scale differing from TCG by a constant rate

$$\frac{\mathrm{d}TT}{\mathrm{d}TCG} = 1 - L_G \,, \tag{3.26}$$

where $L_G = 6.969\,290\,134 \times 10^{-10}$ is a defining constant. The value of L_G was taken equal to the best available estimate of $\hat{U}_g/c^2$ in 2000, but will not be changed in the future, even though values for $\hat{U}_g$ may improve. This new definition of TT does not imply a change of rate of TT in 2000. It has practically no effect when processing data from ground-based clocks, but is particularly important for future spaceborne clocks. In the following discussion, we retain the symbol $\hat{U}_g$, but define it to be $c^2 L_G$.

The use of TT, although perfectly natural, has been a source of some confusion in dynamical astronomy and space geodesy. The coordinate change it involves has a rather complex effect on the evaluation of the potential, and is not always applied correctly. Moreover, up until 1991, international organisations were uninformative as to which spatial coordinates should be used.

Let $\underline{t}$ be the Terrestrial Time, defined by

$$\frac{\mathrm{d}\underline{t}}{\mathrm{d}t} = 1 - \frac{\hat{U}_g}{c^2} \tag{3.27}$$

and a choice of origin (see Chapter 7). Now leaving the x^i with their definition via the metric (3.14) for the non-rotating geocentric system, the new metric takes the form

$$\mathrm{d}s^2 = -\left(1 - \frac{2U}{c^2}\right)\left(1 + \frac{\hat{U}_g}{c^2} + O(c^{-4})\right)^2 c^2\,\mathrm{d}\underline{t}^2 + \left(1 + \frac{2U}{c^2}\right)\left[(\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2\right] . \tag{3.28}$$

Putting

$$\frac{\mathrm{d}\tau}{\mathrm{d}\underline{t}} = 1 - \underline{h}(\underline{t}) \,, \tag{3.29}$$

where, keeping only terms up to order $O(c^{-2})$,

$$\underline{h}(\underline{t}) = \frac{U - \hat{U}_g}{c^2} + \frac{\underline{v}^2}{2c^2} \,, \tag{3.30}$$

and $\underline{v}$ is the coordinate speed expressed relative to $\underline{t}$. For the applications con-

sidered in the present text, the expression

$$\underline{h}(\underline{t}) = h(t) - \frac{\hat{U}_g}{c^2} \tag{3.31}$$

will be sufficiently accurate.

In the rotating geocentric system, spatial coordinates X^i are maintained, as recommended by the International Union of Geodesy and Geophysics (IUGG) in 1991 [3.10], and hence,

$$\underline{h}(\underline{t}) = \frac{\Delta\hat{U}}{c^2} + \frac{\underline{V}^2}{2c^2} + \frac{2\omega}{c^2}\frac{\mathrm{d}A_E}{\mathrm{d}\underline{t}} , \tag{3.32}$$

calculating the corresponding coordinate speed $\underline{V}$ using $\underline{t}$.

(f) Synchronisation and time reference

In general relativity, the Einstein synchronisation loses its transitivity and hence also its interest for the problems which concern us. We shall make use of the *coordinate synchronisation* defined as follows.

In a given coordinate system, two events are *coordinate simultaneous* when they are attributed the same date in the coordinate time t of this system. Two clocks are synchronous, in the sense of being coordinate synchronised, in an interval $t_1 \leq t \leq t_2$, if their coordinate simultaneous readings are equal for all dates t within this range.

In the same spirit, the results of comparisons between clocks are always expressed by the difference in their coordinate simultaneous readings, that is, by comparing readings at the same coordinate time.

In general then, synchronisation and comparison of clocks have no meaning in coordinate systems other than the one relative to which these notions have been set up. Note, however, that a coordinate change of type

$$\begin{aligned} t &\longrightarrow t' = t'(t) , \\ x^i &\longrightarrow x'^i = x'^i(t, x^1, x^2, x^3) \end{aligned} \tag{3.33}$$

does not affect the condition of coordinate synchronisation, nor the result of comparing clocks (although it changes the dates). This is what happens if we switch from t to $\underline{t}$.

Clearly, for Earth-based metrology, or in circumterrestrial space (out to the orbits of geostationary satellites, for example), the rotating or non-rotating geocentric coordinate systems are used. As they have the same coordinate time, the choice between the two is merely a question of convenience. For ground-based clocks, the rotating system is more appropriate, whilst for orbiting clocks, the non-rotating system may well be preferred. If one day we succeed in setting

up widely separated laboratories on the Moon, it could be necessary to define a selenocentric coordinate time. Clocks synchronised on the Moon would not then be synchronous in the terrestrial system.

As far as applications are concerned, worldwide dissemination of a uniquely realised geocentric coordinate time based upon the definition of the second (of proper time) and a unanimously recognised metric achieves the following.

- It provides a basis for synchronising clocks across the Earth and in its immediate environment (i.e., for artificial satellites).
- It provides the time coordinate which, when associated with appropriate space coordinates, enables us to describe orbital motions (of artificial satellites or the Moon) around the Earth.
- Through a four-dimensional transformation, it provides a practical realisation of other coordinate times, such as the Barycentric Coordinate Time required for calculating the orbits of planets or interplanetary probes.
- Through the transformation (3.17), it gives the local proper time and also the SI second to be used for local measurements (the proper second).

3.4 Summary and conclusions

As in everyday life, ordinary Newtonian time is quite sufficient for most technical and scientific applications. However, it is the metrologist's duty to adopt a theoretical framework compatible with the precision and stability of the best standards, over the whole region of space in which they may be used. For time and length, this framework is general relativity.

In this theory, it is essential to distinguish proper physical quantities, directly measurable using standards, from coordinate quantities which depend on an arbitrary choice of spatio-temporal coordinate system. The reason is that it is impossible to find a coordinate system in which the graduation unit along the axes, called here the *scale unit*, is at each point of spacetime equal to a second or a metre, such as we might obtain them at this point by using the definitions of SI units. Let us first examine the consequences for time.

- The unit of time, the second, to be used for experiments in a small enough laboratory (the usual size of room, for example) must be established by a standard located in that laboratory and fixed relative to it. Indeed, it is just because fixed local standards are used for all physical quantities that the universality of the local laws of physics can be guaranteed, and this justifies the use of physical constants within the framework of special relativity (Einstein's equivalence principle). To put it briefly, we say that the local second achieved in this way is the *proper* second (with the implicit understanding

that we are talking about the proper time of the laboratory in which it is used).

- The time scale taken as the worldwide reference must be defined as a coordinate, since this is the only way of unambiguously dating events. In other words, it is a *coordinate time*. As far as terrestrial metrology is concerned, a geocentric spatial coordinate system rotating with the Earth is generally used. However, a geocentric system that is not rotating relative to space is sometimes useful. These two coordinate systems are defined in such a way that they share the same time coordinate, called *geocentric coordinate time*. The relation between proper time and the geocentric coordinate time is given by the integral (3.17), using the expressions (3.18) or (3.21).
- In relativity, synchronisation no longer has an absolute meaning. By convention, events are simultaneous in a specified coordinate system when they have the same time coordinate. From this convention, two clocks can be compared by the difference between their coordinate simultaneous readings. If this difference is always zero, the clocks are said to be synchronous in the sense of being coordinate synchronised. For Earth-based metrology, this coordinate is the geocentric coordinate time. We observe that a change of coordinates in which t' is a function of t alone does not affect these notions of coordinate simultaneity or coordinate synchronisation.

In practice, there are obvious reasons for constructing the unique coordinate time to be taken as standard by convention on the basis of the best possible realisations of the second. Then the transformation of this coordinate time into the proper time of any particular local environment provides a local time scale graduated in proper seconds. This is what is now achieved using International Atomic Time (TAI). It is essential to bear in mind that, although the TAI is graduated in hours, minutes and seconds, the time interval between two consecutive marks indicating TAI seconds (hereafter referred to as the TAI *scale unit*) does not last for one (local) proper second, except on the rotating geoid.

The same comments apply to the unit of length, the metre, and spatial coordinates. The metre must be viewed as the unit of proper length. In its 1987 definition, based on a conventional value for the speed of light (compatible with previous experimental values), it is obtained from the proper second by a local experiment. The scale unit for spatial coordinates, although commonly called the metre, is not everywhere equal to the local metre. In practice, length measurements are less accurate than time measurements, with the result that the spatial scale unit and proper metre need not often be distinguished. However, this distinction is beginning to have some relevance in global geodesy and it is essential in certain astronomical studies.

4

The evolution of time measurement

When measuring time, just as when measuring any other physical quantity, it would seem logical to begin by evaluating metrological needs in view of applications, and then to seek ways of satisfying them. This approach was indeed attempted for astronomical dynamics. But more often than not, standards were realised before considering the use to which they would be put; and then those uses themselves preceded official recognition of the standards by organisations set up to manage metrology on a worldwide scale. In this sense, metrology is a strange combination of pragmatism and rigour.

The main part of this book is devoted to presenting a snapshot of time measurement as it is practised at the time of writing. This contrasts with the present chapter, in which we describe how such measurements have evolved, in terms of both techniques and ideas. It is a story punctuated by hesitation, doubt and sometimes even inconsistency.

4.1 Date, calendar and hour

The date of an event in the scientific sense is a whole set of data attributing a time label to it relative to some specified time scale. Traditionally, it comprises some way of identifying the day, known as a calendar, together with a subdivision of the day which we commonly call hours.

The various calendars (Gregorian, Jewish, Islamic, etc.) are ingenious schemes for making use of the natural cycles that have precise astronomical definitions:

- the apparent *solar day* is the duration between two successive transits of the Sun at the local meridian,
- the *tropical year* is the duration between two transits of the Sun through the vernal equinox (the spring equinox),

- the *lunation* (or lunar month) is the duration between two successive new moons.

We shall restrict ourselves mainly to the well-known Gregorian calendar, which respects the tropical year to a good approximation.†

The day is subdivided into hours, minutes and seconds in a way inherited from the Babylonians [4.1]. It has resisted any attempts at decimalisation. In France, a decree was issued on 24 November 1793 with the intention of imposing a decimal division of the day and a *decimal second*, but it was suspended on 7 April 1795. It is only for intervals shorter than one second that decimal submultiples are used, i.e., the millisecond, microsecond, etc.

In this complex system, a whimsical pocket of resistance to technocratic intervention, the units do not even bear constant relation to one another. Added to the fact that a year may contain either 365 or 366 days, a further claim to originality was to come in 1971. The day, normally lasting 86 400 s, can sometimes include one second more or less in the *Coordinated Universal Time* (UTC) that governs our lives.

In order to simplify their work, astronomers sometimes use the *Julian Date* (JD). This is based on a continuous count of the days since 4713 BC. The Julian Date can be complemented by giving also the decimal fraction of the day. In this system, days are counted from midday. For example, 1 January 2000 at midday corresponds to JD = 2 451 545.0.

In time metrology, the study of the Earth's rotation, and space science, the *Modified Julian Date* (MJD) is often used. This is defined by

$$\mathrm{MJD} = \mathrm{JD} - 2\,400\,000.5\ . \tag{4.1}$$

This means that MJD = 0.0... corresponds to 17 November 1858 at time 0 h. (Note that we are following international usage as regards acronyms. MJD comes from English, whilst International Atomic Time TAI comes from the French *Temps atomique international*.)

By their definitions, the calendar and the JD and MJD systems are closely linked to alternation between day and night. They are nevertheless used with time scales based upon other phenomena, such as International Atomic Time, when these systems drift only slightly relative to the true solar day. No ambiguity can result, provided that the time scale is indicated in the record of the date.

† Between the Gregorian reform of 1582 and about the year 3200, the Gregorian calendar will gain roughly one day over the tropical calendar. The uncertainty in this gain is mainly due to unpredictable irregularities in the Earth's rotation.

4.2 Time measurement based on alternation of day and night

4.2.1 Mean solar time

In ancient times, the day, from sunrise to sunset, and the night, from sunset to sunrise, were almost invariably divided into 12 hours each. Naturally, these hours did not have the same duration in the day and at night, except at the equinoxes or on the equator, and they varied with season and latitude. These variable hours, or *seasonal hours*, were still being used in the fifteenth century, even though ancient astronomers had already invented and used *equinoctial hours* long before, dividing the apparent solar day between successive passages of the Sun across the local meridian into 24 equal periods.

More precisely, the *apparent solar time* (or apparent time) is defined as the hour angle of the Sun, i.e., the angle between the half-plane of the observer's meridian and the half-plane defined by the Earth's axis of rotation and the Sun. This angle is then counted in 'hours' of 15°. (Astronomers do not always clearly distinguish between angles and times, as is revealed by the fact that the astronomical ephemerides of the Bureau des longitudes in Paris are called *Connaissance des temps*.)

Apparent solar time is directly observable and was still used in country areas up until the beginning of the twentieth century. However, it is not a 'convenient' time in the sense that H. Poincaré would have understood the term (see Section 2.3). We may say that it is not a uniform time. Its irregularities were already known to Ptolemy (*circa* AD 150), and are mainly due to the elliptical shape of the Earth's orbit around the Sun and the tilting of the Earth's axis of rotation relative to the ecliptic (the plane of its orbit). They have a total amplitude of 30 minutes and are reproduced cyclically in an identical manner each year, as described by the so-called *equation of time* (see Figure 4.1). Astronomers took these variations into account by using a regularised time known as *mean solar time*. This time scale is, like apparent time, associated with the local meridian. However, attempts to perfect the system did not stop there. The ultimate aim was to guarantee strict proportionality between mean solar time and the angle through which the Earth turns about its axis. In Chapter 8, we shall discover some of the subtleties of this requirement.

Despite the advantages brought about by its regularity, mean solar time was long considered a mere tool in the astronomer's panoply. Once their work was done, they would carry out conversions so that the astronomical ephemerides could be expressed in terms of apparent time. As an example, the mean solar time at Greenwich, England, was only introduced in 1834, in the *Nautical Almanac and Astronomical Ephemeris*. The following year, in France, the *Connaissance des temps* adopted the mean solar time of Paris.

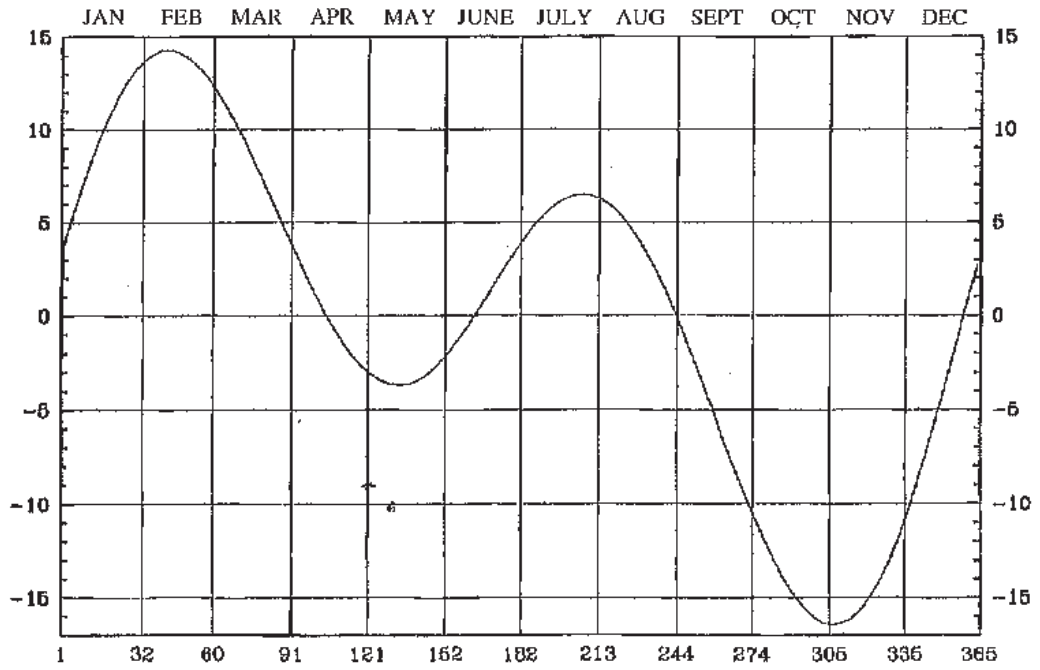


Fig. 4.1. The equation of time. Mean solar time minus apparent solar time in minutes. (Kindly communicated by the Bureau des longitudes, Paris.)

4.2.2 Universal time scales and time zones

During the second part of the nineteenth century, the introduction of rail networks made it essential to use a single definition of the hour, at least on the national level. Many countries then adopted the mean time of the meridian through their capital city, increased by 12 hours. Hence, in France, an Act of Parliament voted on 14 March 1891 imposed the hour of the Paris meridian.

However, this solution proved inadequate in certain countries which extended over a very wide range of longitudes. The idea of dividing such countries into time zones, the time differing from one to the next by a whole number of hours, in such a way that solar midday always occurs at around 12 h, is apparently due to Ch. Dowd in the USA in 1870 [4.1]. It was applied shortly afterwards by S. Fleming, an engineer on the Canadian railroads.

The use of mean solar time thus began to spread, whilst still lacking a single time standard the world over. Agreement for a worldwide unification of time came in October 1884 at an international conference held in Washington 'for the adoption of a single prime meridian and a universal time' [4.2]. The Greenwich meridian stood out as a natural choice since it was already taken as the origin for longitudes on most maritime charts, and indeed it was almost unanimously adopted by national representatives. The universal time was therefore the one determined at this meridian. Moreover, it was stipulated at the conference that the universal day should begin at midnight on the prime meridian, in

contrast to the common astronomical practice of counting days from midday, and that hours should be counted from 0 to 24.

During the following two decades, the time zone system was related to universal time and extended around the Earth. Initially, the planet was cut into 24 time zones, each occupying 15° of longitude, with the first centred on Greenwich. Each zone was centred on its *standard meridian* and *standard time* was thereby established within the zone. Any country covering a reasonably small range of longitudes was to adopt the time zone containing its capital city. Little by little, countries began to implement the new system. Since 1971, it has been associated with Coordinated Universal Time (UTC), to be defined in Section 4.5 and discussed in greater depth in Chapter 7. For example, on 9 March 1911, a French Act of Parliament introduced the time of the Greenwich meridian to France in terms intended to spare national sensitivities: 'Official time in France . . . is the Paris mean time delayed by nine minutes and twenty-one seconds.' (It is interesting to note the inappropriacy of the terms, since the mean time at midday is 0 h.) This Act remained in force until 9 August 1978, when a government decree associated official time with UTC.

The standard time system has since lost something of its original simplicity through the use of permanent shifts between official time and the time of the relevant zone, and through the use of summer time. For example, the latter appeared in France in 1916 'in order to counter an annoying tendency of a great number of town dwellers who get up and go to bed too late' [4.3]. It is extremely difficult to find out the difference between UTC and official time or the commonly used time in various countries, since no organisation is responsible for collating such information.

Astronomers were slower to accept the universal time of the 1884 conference. The *Connaissance des temps*, for example, was aligned with Greenwich Mean Time in 1916, but not with universal time which, as we have observed, differs from GMT by 12 hours.

It was only in 1925 that the ephemerides used by navigators and astronomers introduced the day which began at midnight rather than at midday. The new time scale, deduced from mean time by adding 12 hours, was referred to as *civil time*. For example, the civil time at Greenwich, which was nothing other than the universal time specified in 1884, was taken as the time argument for the *Connaissance des temps* from 1925 to 1950. However, it is commonplace, even in astronomy, to keep the appellation *Greenwich Mean Time* for this new scale, and it has been the source of much confusion. Although the name *Universal Time* arose straightforwardly from the terms used in the 1884 conference, it was not until 1948 that the International Astronomical Union, after a good deal of hesitation, finally made firm recommendations in its favour. This has

not prevented continued incorrect use of the acronym GMT, where UT should be used.

Even today, there still remains a vestige of the system whereby days are counted from midday, viz., the Julian Date mentioned in Section 4.1.

4.2.3 Towards a unique realisation of Universal Time

Let us begin with a convention aimed at simplifying language. We shall use the name *universal time* to speak of any time scale based on the Earth's rotation and referred to the prime meridian. This therefore includes Universal Time (UT) with its modern definition, as well as GMT, shifted by 12 hours relative to UT, and also *Greenwich Sidereal Time*. The latter is the hour angle of the vernal equinox measured from the prime meridian and it is mathematically related to GMT and UT.

From the middle of the nineteenth century up until around 1970, national timekeeping organisations operated along more or less unchanged lines. The basic equipment included clocks that were as stable as possible, together with refracting telescopes that could be used to observe the transits of stars across the meridian, or occasionally at equal altitudes. This was complemented by devices for emitting time signals that were originally electrical, optical or sound signals. Then, in about 1910, emitters and receivers of radio time signals were used. In periods of clear weather, astronomers spent their evenings, or even the whole night, timing star transits by means of one of the local clocks taken as master clock. They would then go to bed before setting about 'reducing their observations', which basically meant deducing *clock corrections*. These corrections were to be added to the readings of the master clock in order to obtain, at the instant of astronomical observation, the local time values or, after adding the longitude, the universal time. The purpose of the clock was therefore to provide a sort of average for observations, smoothing them out, so to speak, and to keep time between evenings when observations were made. By extrapolation, the clock thus gave an approximation to universal time in real time. With the help of this extrapolation, time signals could be emitted in the form of pulses at agreed nominal times. Figure 4.2 shows how this works, highlighting some of the uncertainties associated with the method. The point in having a good *timekeeper* is clear. However, by pushing forward progress in clock timekeeping, astronomers have finally lost their hold upon the management of time.

When radio time signals began to cross the Atlantic in around 1910, it was found that they could differ by as much as one or two seconds. This discovery came at a time when random uncertainties in local measurements of universal

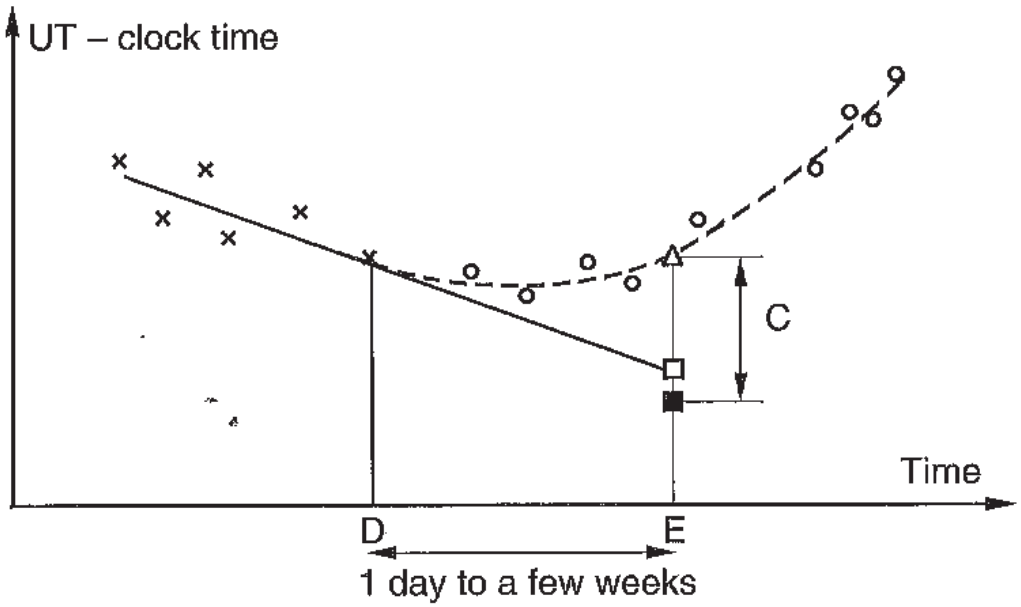


Fig. 4.2. Emission of time signals in UT. The astronomical observation × of UT minus clock time, available at the time E when the time signal is emitted, only extends up to the time D because of computation and transmission delays, and bad weather. A linear extrapolation is used to predict the value (marked □) at E. Subsequent observations (marked o) belie this prediction. In addition, the emission time ■ may not actually coincide with the prediction. The total error in the emission time is C.

time were as low as a few hundredths of a second. The main causes of systematic error, to be combined with random errors, were errors in observed star positions, instrumental delays and, above all, errors in longitudes, the three causes adding together in unknown measures. Universal time is indeed unique by its definition, but its realisations could diverge in a quite unacceptable manner.

At the Bureau des longitudes in Paris, it was clear that this situation would have to be remedied. Therefore, in 1911, on the initiative of Ferrié, they suggested that the French government convene an international conference to create a Bureau international de l'heure (BIH). Its basic task would be to unify world time. The meeting took place in 1912, but war was to intervene before the articles of the BIH could be established. The BIH nevertheless began to carry out its functions immediately, in a semi-official manner, with assistance from the Paris Observatory. It acquired official status in 1919, when it was placed under the responsibility of the International Astronomical Union, and carried out its mandate until 1988. At this point its activities were split into two parts. Atomic time measurement was transferred to the International Bureau of Weights and Measures (Bureau international des poids et mesures, BIPM), whilst its astronomical and geodetic activities were reorganised and

extended to form the International Earth Rotation Service (IERS). The IERS Central Bureau was set up in the Paris Observatory until 2001, before moving to the Institute of Cartography and Geodesy in Frankfurt, Germany. We shall comment further on the activities of the IERS later in the book. The references [4.4, 4.5, 4.6] give a historical review of the BIH and summarise its activities.

The aim of the BIH was to provide a single approximation to the theoretical universal time, known as the *definitive time* and then to give the discrepancy between this definitive time and the nominal emission times of time signals. Calculations were extremely complex, although based on simple enough principles. We may outline them in the following way. Each time service referred astronomical observations and time signal emissions to its master clock. Then by exchanging time signals, the BIH could put together a unique worldwide master clock, referring the whole set of observations and emissions to it. It was then a simple matter to construct the definitive time, by averaging astronomical measurements, and refer time signal emissions to it. The results were published in the *Bulletin horaire* every month or every other month, depending on the period.

Let us consider the results for October 1936 as an example, published in April 1937. The definitive time of fifty-six daily emissions is given to the nearest millisecond. We may expect a fuzziness to the extent of a few milliseconds in the definitive time, due to uncertainties over the estimated propagation delays of the radio signals. With regard to this fuzziness, the offsets of emission times with respect to their nominal values as estimated by the BIH were considerable, reaching as much as 0.2 s.

We may well wonder to what extent the BIH's definitive time represented the theoretical universal time defined by astronomers. It is hardly possible to give a precise answer to this question because of the systematic errors already mentioned, but it would be wise to expect uncertainties of up to several hundredths of a second. This shows how important it is to distinguish synchronisations which can be very accurately guaranteed on some conventionally agreed basis, and the realisation of a theoretical time scale using this basis. We shall see that this distinction explains a certain reluctance towards setting up an atomic measure of time.

The work carried out at the BIH barely changed until around 1960, although technical improvement of the clocks used gradually made it possible to reduce synchronisation errors in time signal emissions. We may say that, by this time, the unification of time could be ensured to within one or two milliseconds, with a time lag of about one year before results were announced. In real time, by extrapolating clock readings and using local observations, the discrepancies in the unification of time were of the order of 10 ms.

4.2.4 Definition of the second before 1960: the second of mean solar time

The measurement of time based on the Earth's rotation originally rested upon the reproducibility postulate, as discussed in Chapter 2. This was first applied to the mean duration of the day, and then to the period of rotation. When Euler showed in 1737 that the Earth, considered as an undeformable solid, was spinning at a uniform rate, the mean solar time took on the character of a Newtonian dynamical time. However, two centuries later, when it was realised that the Earth's rotation was not so regular after all, it turned out to be much more difficult to establish a mathematical model, and thereby define a time system. Only a small empirical correction was applied from 1955, in order to remove an annual fluctuation (the corrected scale being denoted UT2). It had been realised that mean solar time, and subsequently, Universal Time, were not a good measure of time, and a better clock was sought to replace the Earth's rotation.

The second, unit of time, or rather, duration, was tacitly and universally defined as the duration of 1/86 400 of a mean solar day. It is surprising that this definition of the second of mean time, which remained in force until as late as 1960, had never been officially ratified by the organisations responsible for world metrology since the 1875 Metre Convention.

4.3 Time based on Solar System dynamics: Ephemeris Time

4.3.1 First doubts on the uniformity of the Earth's rotation

Whilst Copernicus accepted the ancient Greek dogma concerning the uniformity of the Earth's rotation, Kepler mentioned the possibility of some irregularities. The first attempt to demonstrate the existence of such phenomena appears to be due to Flamsteed. Shortly after the foundation of the Greenwich Observatory in 1677, he set up gigantic pendulums in the hope of revealing some anomaly, but found none whatever [4.7]. Later, in 1752, the Berlin Academy of Arts and Sciences, presided by Maupertuis, raised the following question: 'Has the daily motion of the Earth always occurred at the same rate or not? What means do we have at our disposal for answering this question? And if there were some irregularity, what might be its cause?' In his reply, Kant said that there could be a slowing down effect due to dissipation of energy in tidal movements of the oceans. He was right, but the idea was only confirmed by observation two centuries later.

The conviction that the Earth's rotation must be perfectly uniform was deeply rooted, and these first doubts were barely heard. In 1825, Laplace wrote: 'It is therefore certain that, since Hipparcos, the day's duration has not changed by as much as one hundredth of a [decimal] second [i.e., 0.008 64 ordinary

seconds].' The claim is unfounded, as we shall see. Whilst Ferrel in 1864 and then Delaunay in 1865 both asserted that discrepancies observed between the lunar ephemeris and observation were due to a lengthening of the day, Fleming declared in 1864 that there was no motion more uniform than that of the Earth about its axis.

4.3.2 Acceptance of irregularities in the Earth's rotation

Proof of irregularities in the Earth's rotation first came from a study of the orbital motions of the planets and the Moon. To the degree of accuracy required by observations, the orbital motions of the planets can be studied under the assumption that both they and the Sun are point objects. The purely gravitational interactions between them are simple enough for their motions to be worked out to a high level of precision, using for example the model provided by Newtonian mechanics which was common practice at the time. (Relativistic theories have been in use for the past few years, but the differences will not be relevant in the context of this discussion.) Generally speaking, satellite motions about a central body are more complex to treat. This is due to the fact that the central object may not be exactly spherical and also to the presence of phenomena leading to energy dissipation. Such effects are nevertheless small in the case of the Moon, which we shall be concerned with now, and we shall ignore them for the moment. (Further explanations of this point will be given in Chapter 8.) The next step is to fit the theory to observations. In other words, numerical values must be attributed to unknown parameters so that an ephemeris can be established, giving the positions of the bodies in tabular form as a function of time, e.g., in terms of the mean solar time. But what happens if the time used to date observations differs from the uniform time of Newtonian theory? Then discrepancies arise between observed positions and those given by the ephemeris at the instant of observation. Such discrepancies, if large enough, cannot be explained by the random errors of observation. Clearly, they are minimised to a certain degree by the fitting of orbital parameters, which are therefore biased. Even then, they still remain to some extent, and because of the errors in the orbital parameters, they are liable to increase rapidly with time when the ephemeris is used to predict what will happen.

This is precisely what Newcomb found when using the lunar ephemeris established by Hansen in 1857. Having checked that the theory had been used correctly, Newcomb considered the possibility of irregularities in the mean solar time. In other words, he put the Earth's rotation to question. However, it had first to be checked that the ephemerides of the other planets revealed the concomitant discrepancies, and in a way compatible with the hypothesis of an

irregularity in the Earth's rotation. This was a difficult task, since the effects are much smaller for the planets than for the Moon. Indeed, they are proportional to the apparent speed of the orbital motion. In 1927, following research by Brown and de Sitter [4.8], no doubt could remain. The Earth's rotation was not a good clock.

Almost immediately, in 1929, Danjon suggested using the dynamics of orbital motions to measure time, in an article that proved remarkable for its clarity and the correctness of its predictions [4.9]. Let us quote some short extracts (translation).

> ... it is legitimate to consider the [rotation of the] Earth as the sole cause for the apparent disorder that still reigned in the Solar System. Although Newton's law has been saved, it is experiencing a quite extraordinary adventure: henceforth called upon to gauge the passage of time, it becomes in part unverifiable and ceases to be what could strictly be termed a law. [...] Since we would ask these [Kepler's] laws to provide a measure for the passage of time, we could no longer subject them to experimental control without entering into a vicious circle. [...] Let us simply hope that we shall one day discover a good terrestrial time standard, so that we may leave these purely logical difficulties behind us.

The good terrestrial time standard appeared in 1955 in the form of the atomic time standard. Danjon may well have been thinking of it.

Unfortunately, Danjon's article was published in an amateur astronomical review and did not receive the attention of his peers. It was not until 1950, during an international colloquium on astronomical constants held in Paris [4.10], that Clemence made a precise proposal for defining a dynamical time. This was christened *Ephemeris Time*, following a suggestion by Brouwer. The definition of Ephemeris Time, to be discussed in Chapter 8, was ratified by the IAU in 1952.

We shall see in particular that Ephemeris Time suffers from a serious drawback, the wretched quality of its readings, with uncertainties of the order of 0.1 s. It may be compared with certain luxury wrist watches which, although provided with an excellent mechanism, have no graduations marked on the face to help us gauge the position of the hands. It is thus only possible to exhibit the large scale and long term variations in the Earth's rotation. In 1936, an irregularity in the Earth's rotation was revealed for the first time and independently in two laboratories, relative to artificial clocks, including both pendulum clocks and quartz clocks. The annual total amplitude was later found to be 60 ms [4.11, 4.12]. However, as can be seen from Figure 4.3, the annual irregularity was at the limit of what could be revealed by the clocks in use at the time. Later, atomic time proved adequate to demonstrating a great many other irregularities (see Chapter 8).

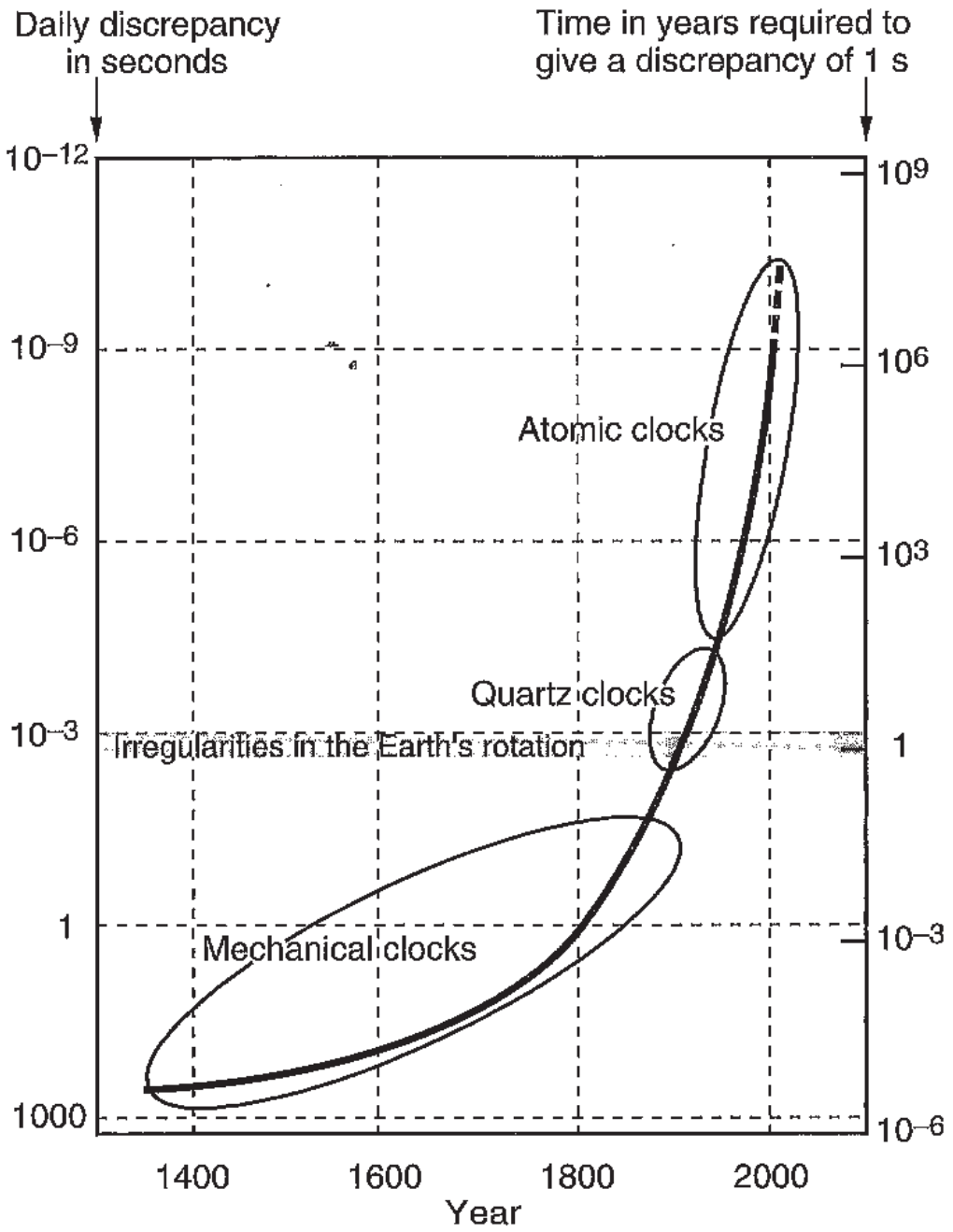


Fig. 4.3. Improvement in the quality of artificial clocks and comparison with the clock provided by terrestrial rotation.

All these irregularities lead to variations in the duration of the second of mean time. In order to eliminate some short term fluctuations and the annual irregularity, whilst also reducing the role of measurement uncertainties, we may consider the average annual value of this second. Unpredictable variations nevertheless remain, reaching almost 10^{-7} s.

4.3.3 Definition of the second from 1960 to 1967: the ephemeris second

As explained in Section 2.4.2, if we aim to base the definition of the second on dynamical considerations, it suffices to fix the duration of some particular phenomenon numerically. The 1950 colloquium recommended basing this numerical value on Newcomb's *Tables of the Sun*, a proposal which raised no objections. However, the choice of reference phenomenon did give rise to some hesitation. The one chosen by the International Astronomical Union at its 1952 General Assembly was the sidereal year of 1900, the duration measured between two solar transits of a fixed equinox (i.e., without taking the precession of the equinoxes into account), because this duration is virtually constant. It was soon realised, however, that the more directly observable tropical year was preferable, despite its slow variation. For this reason, and quite exceptionally, the IAU resolution was corrected by correspondence.

In 1956, the International Committee for Weights and Measures used the powers conferred upon it by the 10th General Conference on Weights and Measures (CGPM) in 1954 to decide upon the following definition:

> The second is the fraction 1/31 556 925.9747 of the tropical year for 1900 January 0 at 12 h Ephemeris Time.

This definition was ratified by the 11th CGPM in 1960. Shortlived, it was abrogated in 1967 in favour of an atomic definition for the second.

The definition of the *ephemeris second* proved puzzling to those who were not familiar with astronomy. Its strange definition arises from the fact that the mean longitude of the Sun is, according to Newcomb, a quadratic function of Ephemeris Time (ET). The tropical year for 1900 January 0 at 12 h ET is a fictitious year corresponding to the speed of the mean longitude at the given date.

Let us look more closely at the numerical value used in the definition of the ephemeris second. Newcomb's tables were based upon astronomical observations from the nineteenth century. Naturally, these were dated using mean solar time. The duration of the ephemeris second is therefore roughly equal to the average duration of the second of mean time over this century. However, in 1960, when the ephemeris second was adopted as the unit of time for SI units, the new second was shorter than the currently used second of mean time (averaged over the year 1960) by 1.4×10^{-8} s, and this offset was roughly known at the time of the decision. This goes against common metrological practice of adjusting successive definitions of a unit in such a way that they lead to an invariable quantity during changes of definition, at least, within the

limits of experimental error. It could not have been predicted in 1960 that this would have unfortunate consequences for the Coordinated Universal Time system that is now in use.

The ephemeris second is uniquely defined, but it is rooted in the past, in 1900. Its physical realisation at a given date depends on the astronomical ephemerides. It is estimated that it could be achieved to within $\pm 2 \times 10^{-9}$ s around 1960 by the analysis of astronomical observations extending over several years. Compared with the second of mean time, the gain in reproducibility was by a factor of 50.

4.3.4 Ephemeris Time: a scale reserved for astronomers

The beauty and simplicity of Ephemeris Time raised some enthusiasm. In 1966, when atomic clocks had already been running for eleven years, Woolard and Clemence wrote [4.13]: 'The traditional practice of measuring time by the apparent motions of celestial bodies may now be based upon an exact dynamical foundation by adopting as a primary standard the measure of time implicitly defined by the laws of motion.' The story of Ephemeris Time is nevertheless plagued by a long series of difficulties and misunderstandings, as we shall see in Chapter 8.

Ephemeris Time was never made available, nor used, in everyday life. In fact, its application was limited to the needs of astronomical dynamics. There has never been an official organisation expressly set up to centralise and process its measurements so as to produce a unique realisation that could have been adopted by convention.

Let us emphasise that, until 1960, the official time scale, Universal Time, and the definition of the second (of mean time) were both based in a consistent way upon the Earth's rotation. Hence, the dissemination of the time scale effectively provided the SI second. With the adoption of ET began a period of incompatibility between the time scale, still UT, and the unit of time. The atomic measurement of time was to put an end to this confusion in 1971.

4.4 Atomic time measurement

4.4.1 Atomic frequency standards and the atomic definition of the second (1967)

Science has gradually accepted that the various chemical species are composed of molecules which are themselves made from atoms. A relation was eventually established between the structure of emission spectra and the atomic and

molecular composition of excited gases. The assumed universality of atomic properties could then be transposed to these spectra. Maxwell in 1873 and Kelvin in 1879 both suggested that the wavelength of a spectral line and the period of the corresponding radiation could be used to define the unit of length and the unit of time, respectively [4.14]. These proposals were truly prophetic if we remember that the definition of the metre was only based upon the wavelength of a spectral line from 1960 (and until 1983). Moreover, optical emissions can now be used in this way by highly specialised laboratories, and optical time and frequency standards may soon compete with current atomic standards using transitions at centimetre wavelengths.

As in many other areas, the understanding we now have of electromagnetism, quantum physics, atomic physics and spectroscopy, which have led to the invention and development of time and frequency standards, was acquired at the end of the nineteenth century and in the first half of the twentieth century. We shall summarise these years of progress by citing those physicists who made key contributions. Planck laid the foundations for quantum theory, whilst Einstein introduced the idea of the photon and stimulated emission, amongst many other significant contributions, including the theory of relativity. Bohr applied quantum theory to the explanation of atomic structure and introduced the idea of energy levels. Hertz was the first to produce and detect radio waves. De Broglie, Heisenberg and Schrödinger created and developed wave mechanics. Stern contributed to the kinetic theory of gases and discovered with Gerlach the magnetism of atoms and its spatial quantisation.

Technological advances which began in the 1930s and were pushed ahead by the need for radio communication and radar detection during the Second World War also played a key role. In 1945, it became possible to produce radio waves with frequencies up to 30 GHz and to measure this frequency. These developments were the starting point for a major leap forward in precision Hertzian spectroscopy, led by Townes and Pound.

The first atomic clock was made in 1948 by Lyons at the National Bureau of Standards in the USA, now called the National Institute of Standards and Technology [4.15]. The reference was an intense absorption line of the ammonia molecule, lying at around 24 GHz. The molecular resonance controlled the frequency of a quartz oscillator which then provided pulses to mark time. However, the resonance was broadened by the Doppler effect and collisions. Its long term stability was no better than that of a quartz clock and this approach was dropped.

The caesium clock was the result of work carried out by Rabi and his students [4.16, 4.17]. At the end of the 1930s, Rabi was working on the magnetic resonance technique for atomic or molecular beams. This method, originally

aimed at measuring magnetic moments, was soon adapted to the study of the radio frequency spectra of atoms and molecules. The first experimental evidence for the hyperfine transition in the caesium atom was obtained in 1940. The possibility of using the magnetic resonance method with an atomic beam to make an atomic clock had already been discussed within the Rabi group in 1939. It seems well-established that Rabi presented this idea publicly at a conference in 1945, during a meeting of the American Physical Society. In fact, he suggested using the caesium atom. In 1950, Ramsey improved conditions for the interaction between the electromagnetic wave and the atoms by introducing two separate oscillating fields. Putting together the experimental methods developed by Rabi and Ramsey, Essen and Parry built the first reliable caesium beam clock in 1955 at the National Physical Laboratory in the United Kingdom [4.18]. It was regularly used to calibrate the frequency of quartz oscillators. At the same time, one of Rabi's students, Zacharias, developed a prototype commercial caesium clock at the Massachusetts Institute of Technology, in which the atomic resonance controlled a quartz oscillator through a servo loop. The first caesium clocks based on this prototype were commercialised in 1956 by the National Company under the name of Atomichron [4.14].

Experimental confirmation of the principle of amplification by stimulated emission of radiation was first obtained in 1954–55 by Townes in the USA and by Basov and Prokhorov in the USSR. They built the first self-oscillating masers (Microwave Amplification by Stimulated Emission of Radiation), using the ammonia molecular transition at 24 GHz. They were considered for use in atomic clocks but their long term instability of the order of 10^{-10} was too great and the idea was not taken further [4.19]. These achievements were nevertheless useful in showing the way to a solid state maser amplifier devised by Bloembergen in 1956, and then the laser, developed by Schawlow and Townes in 1958 [4.17].

The hydrogen maser, elaborated by Ramsey, results from an evolution of Rabi's magnetic resonance method. The original aim was to narrow the resonance line of the caesium atom by increasing dwell time between the two oscillating fields [4.17]. To this end, the caesium atoms were held back for a certain time, allowing collisions to occur in a cell coated with paraffin wax and equipped with baffles. It was then realised that a Teflon coating was better. The idea was put to the test with hydrogen atoms, allowing them to undergo a great many collisions without significant perturbation. The detection of hydrogen atoms by stimulated emission was also achieved and the hydrogen maser was thereby invented in 1959–60. Its success rests upon the very long dwell time of the atoms in the Teflon coated cell (about 1 s), and hence upon the very long period of interaction with the electromagnetic field in a resonant cavity, which

favours stability of the frequency. A commercial version was soon developed by Vessot at the manufacturer Varian Associates.

Since many applications required compact atomic time and frequency standards, the possibility of using optical pumping was studied from the moment of its discovery. In 1950, Kastler suggested replacing the intense inhomogeneous magnetic field, until then used to select and detect the atomic states, and involving bulky magnets in those days, by an optical method. Dicke showed in 1953 that Doppler broadening of microwave resonance lines could be practically eliminated by mixing the atoms with a neutral gas that limited their scattering speeds. In 1956, Dehmelt discovered that the microwave resonance was manifested by a reduction in intensity of the pumping light transmitted by the cell containing the atoms and the buffer gas [4.15, 4.16]. It thus became possible to build compact rubidium clocks, and they were soon being produced by various manufacturers. These are still the most widespread atomic clocks. From the 1980s, optical pumping methods have been applied to caesium beam clocks, thanks to the development of semiconductor lasers.

The two- or three-dimensional confinement of ions in a non-uniform alternating electric field was developed by Paul from 1955. The long term trapping obtained in this way is favourable for observing narrow lines in the microwave region. Dehmelt was the first, in 1965, to use such traps in the precise measurement of a hyperfine transition frequency, namely that of $^3He^+$. In 1969, Major suggested combining the optical pumping and ion confinement methods in a time and frequency standard using $^{199}Hg^+$. Such a standard was first achieved in 1979 at the Laboratoire de l'horloge atomique in France. Several improved operating models of this standard were produced by Hewlett-Packard under the guidance of Cutler in 1983. Studies of trapped-ion frequency standards continue, using transitions in the microwave and optical regions. Some of these involve the laser cooling method proposed by Wineland and Dehmelt in 1975.

In the same year, Hänsch and Schawlow described an efficient way of slowing down and reducing the thermal motions of neutral atoms. The method was first used to slow down atom beams. Then, in 1989, Wieman and his team at the University of Colorado succeeded in applying it to caesium atoms held at low pressures in a cell. A ball of atoms with a kinetic energy of thermal excitation corresponding to a temperature of the order of 1 microkelvin was trapped at the crossing point of six light beams. Several teams in the USA and France, among others, have made experimental and theoretical contributions towards improving techniques for cooling, trapping and manipulating cold atoms. It is now possible to launch cooled caesium atoms with a speed of the order of 1 $m\,s^{-1}$. It has thus been possible to use them in caesium clocks, where the advantages

of slow atoms are decisive. The first primary standard based on cooled caesium atoms was built in France in 1996 by Clairon at the Laboratoire primaire du temps et des fréquences, Salomon at the Ecole normale supérieure, and their respective research teams.

The Nobel committee has rewarded some of those physicists who directly contributed to the invention and development of atomic clocks: I. Rabi (1944), C. Townes, N. Basov and A. Prokhorov (1964), A. Kastler (1966), N. Ramsey, W. Paul and H. Dehmelt (1989), S. Chu, C. Cohen-Tannoudji and W. Philips (1997), the dates corresponding to the attribution of the prize.

Chapter 6 explains how the various atomic frequency standards work. We shall see that it is the caesium standard that presently leads to the greatest accuracy. Accuracy is a key idea in metrology. It refers here to the capacity to deliver a frequency bearing a perfectly known relationship to the transition frequency of the unperturbed atom. It excludes all frequency drift. At the time of writing, the caesium standard thus plays a privileged role in fundamental time measurement. Let us now review how its accuracy has evolved and outline the consequences of this evolution for the definition of the second.

The first caesium standards involved frequency inaccuracies with a relative value of the order of 10^{-9}. (Uncertainties and small variations are generally expressed as relative values. We shall no longer specify this in the following, since an absence of physical unit will be a sufficient indicator.) Progress was made in leaps and bounds and, as we shall see, standards could be compared with one another in order to assess their agreement. Clearly, the transition frequency ν_{Cs} had to be expressed in terms of the unit of time then available. In their early research work, Essen and Parry used the second of mean time which could be supplied in real time. The ephemeris second was already defined, although not yet the recognised SI unit. However, as we have already noted, it could only be realised as a result of long analysis. This analysis was undertaken by Markowitz *et al.* [4.20], giving in 1958 the frequency value

$$\nu_{Cs} = 9\,192\,631\,770 \pm 20\ \text{Hz}\ . \tag{4.2}$$

The unit Hz mentioned here is based on the ephemeris second. The uncertainty of ±20 Hz was almost entirely due to uncertainty in the realisation of the ephemeris second.

Naturally the various laboratories studying atomic frequency standards were keen to agree upon a single frequency value for caesium. The value due to Markowitz *et al.* was immediately adopted by convention, in the sense that it was regarded as a strictly accurate reference for expressing frequencies. This implicitly defined an unofficial atomic second. Indeed, it was recognised as such by the 12th General Conference on Weights and Measures (CGPM) in

1964, although it was felt that the moment had not yet come to adopt a new official definition of the second in view of further significant progress that might soon be made. However, aware of the urgency of the situation, the hyperfine transition of caesium 133 was designated for use as standard and the above value was assigned to its frequency.

In 1967, the frequency inaccuracies of the atomic standards were reduced to 10^{-12} and it was decided to move ahead with an atomic definition for the unit of time. This was accomplished by the 13th General Conference on Weights and Measures in 1967:

> The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom.

The definition of the so-called ephemeris second was thereby abrogated. The atomic second was fitted as far as possible with the ephemeris second as it stood at the time. Subsequent measurements of Ephemeris Time confirmed this fitting. The current second thus has a duration that corresponds to that of the second of mean time as it was averaged over the nineteenth century.

The improvement of caesium standards continued. In 1976, inaccuracies in the best standard, then in Germany, were reduced to a few multiples of 10^{-14}. The situation remained at this point until 1995 when a new stage in the ascension towards improved accuracy was set in motion by the discovery that atoms could be cooled by lasers. The technique has already been integrated into some caesium clocks, as we shall see in Chapter 6.

Applications of time measurement are on the increase because industry is able to manufacture high quality timing equipment. In particular, caesium standards are commercially available and, although they are less accurate than the best laboratory instruments, they are less bulky, have high frequency stability and can operate continuously with great reliability. Strictly speaking, these are caesium clocks, stringing together the seconds. The existence of such clocks has made it feasible to base the world time scale on atomic transitions rather than the motions of celestial bodies. This brings us to the birth of International Atomic Time.

4.4.2 From Universal Time to International Atomic Time

(a) The age of frequency comparisons and integrated atomic time

The first caesium frequency standards operated sporadically. They were used from time to time to calibrate the frequency of independent quartz clocks

which, between such measurements, retained a memory of the atomic frequency. It was then possible to establish a time scale based upon the atomic transition frequency, that is, an atomic time. Let us discuss some of the details of this operation. We may thereby bring out the properties of the atomic time realised by this method and also introduce useful definitions in what follows.

An oscillator is characterised by a *nominal frequency* ν_0, stated or claimed by its manufacturer and referred to the definition of the second. Its actual frequency is slightly different and varies. This is denoted $\nu(\theta)$, where θ is the date in a time scale θ taken as reference. We shall see that few metrological qualities are required of θ which may, for example, be taken as Universal Time. Since the nominal frequencies of various oscillators are not necessarily the same, it is useful to define:

- the *normalised frequency* $\Phi(\theta)$ by

$$\Phi(\theta) = \frac{\nu(\theta)}{\nu_0} , \tag{4.3}$$

  which is close to unity;
- the *relative frequency offset* $y(\theta)$ by

$$y(\theta) = \Phi(\theta) - 1 , \tag{4.4}$$

  which is small compared to unity.

In order to construct an atomic time scale, the value of the nominal frequency of the transition $\nu_{\mathrm{Cs},0}$ must first be fixed. This can be the value that has been used to define the second, but in the beginning this was not known and different values were used. If the frequency of the external quartz clock is ν_{Q}, the episodic comparisons of frequencies at dates θ_i can be expressed in terms of the difference between normalised frequencies, viz.,

$$y_{\mathrm{Cs}}(\theta_i) - y_{\mathrm{Q}}(\theta_i) \equiv [y_{\mathrm{Cs}} - y_{\mathrm{Q}}](\theta_i) . \tag{4.5}$$

These are in fact mean values around the date θ_i. Starting from these measured values, the function $[y_{\mathrm{Cs}} - y_{\mathrm{Q}}](\theta)$ must be estimated as accurately as possible. The atomic time τ_{A} is then obtained by correcting the readings τ_{Q} of the quartz clock:

$$[\tau_{\mathrm{A}} - \tau_{\mathrm{Q}}](\theta) = [\tau_{\mathrm{A}} - \tau_{\mathrm{Q}}](\theta_0) + \int_{\theta_0}^{\theta} [y_{\mathrm{Cs}} - y_{\mathrm{Q}}](\theta)\, \mathrm{d}\theta . \tag{4.6}$$

This explains why time scales produced in this way are often referred to as *integrated time scales*. The origin of τ_{A} is arbitrary. Whilst it was at first chosen differently by the various laboratories, it was eventually fixed in 1960 by the

condition that τ_A should equal the Universal Time UT2 on 1 January 1958 at 0 h UT2. As far as the value of $\nu_{Cs,0}$ was concerned, it was universally agreed to use the value obtained by Markowitz *et al.* It is clear that the atomic time scale depends on what assumptions are made to interpolate the frequency, and also on the numerical method adopted to carry out the integral. In other words, it is not uniquely determined.

When the first atomic time scales were being built up in this way in the years 1955 to 1960, very low frequency (VLF, 10–30 kHz) radio broadcasts were already being used for long distance radio communications and as navigational aids. The carrier frequencies were stabilised by quartz oscillators. These frequencies could be very accurately received by phase tracking. They thus constituted a good reference to which frequencies available in different laboratories could be compared. Frequency comparisons thus became possible over intercontinental distances. Time comparisons remained uncertain, however, since, even though an initial time comparison was available, it was virtually impossible to avoid losing some periods of the carrier frequency, so that anything from a few microseconds to a few tens of microseconds might slip away each year.

The pioneers of atomic time therefore only had available frequency comparisons between standards, tainted with a good many uncertainties. These included weaknesses in their local mode of frequency interpolation, a poor knowledge of radio wave propagation, and phase losses. Through computations, they were nevertheless able to construct a mean atomic frequency standard that was probably better than each of the individual standards, and produce mean integrated atomic times that were materialised locally by correcting readings from one of their clocks. This method was used in particular by the Bureau international de l'heure until 1969 [4.21].

Such integrated atomic times had originally been set up to study irregularities in Universal Time, that is, in the Earth's rotation. They were also used to stabilise time signal broadcasts, allowing as they did a better extrapolation of Universal Time than earlier clocks. In contrast, their application to other fields of activity, such as celestial dynamics in particular, raised little enthusiasm. The general opinion of astronomers is well expressed in a work published in 1966 [4.13]:

> An atomic clock provides only a standard of *frequency*; it determines a unit of time, but not the continuous count of units that is necessary to determine the interval elapsed since any initial epoch in the past. Astronomical time determinations are essential for defining an epoch and referring instants of time to it, as no artificial clock can be indefinitely sustained in continuous operation in the manner of the celestial motions.

(b) The advent of modern atomic time

It is true to say that the rise of atomic time is primarily due to technical advances made in the design of atomic clocks and their general proliferation. But it would not have taken place at all without precise methods for comparing their readings, for *time comparisons*, and without being able to transfer times in such a way as to carry out *synchronisation*. Conventional radio time signals as they appeared at the beginning of the twentieth century, and which are still used, are subject to some uncertainty as regards the time required for their propagation. This uncertainty, estimated to be around 1 ms, is quite unacceptable if we hope to exploit the qualities of atomic clocks over large distances. For example, at the time when the atomic definition was adopted (1967), thirty years of averaging would have been required to benefit from afar from the full accuracy of the best caesium clocks by the use of such time signals.

In order to fix the orders of magnitude relevant to the metrological problems treated in this work, we may reasonably assume that the frequency degradation introduced by remote time comparisons over one day of measurements should not exceed the inaccuracies in the frequency standards. For standards accurate to 10^{-12}, uncertainties in comparisons should be less then 0.1 μs, whilst at 10^{-14}, an uncertainty of 1 ns should be achieved. These are truly demanding requirements. It has almost always been impossible to harmonise the accuracy of time comparisons with the stability and accuracy of atomic clocks.

Nevertheless, in 1962, an experimental transmission of signals by the telecommunications satellite Telstar provided a time transfer link accurate to 1 μs (for the spatial segment) between the United Kingdom and the USA [4.22]. Although this experiment was followed by many others, the exercise remained at the experimental stage. As recalled in the quotation at the end of the last section, one feature of the quantity time is that its measurement requires a continued effort. Methods were therefore needed to compare times that were simple and cheap enough to apply on a day-to-day basis.

A practical solution was found around 1968 using the navigation system Loran-C. The signals travel by ground waves and thus have highly stable propagation delays (with possible variations of ±1 μs), although they are not known a priori, and a range of 1000–2000 km. Locally, the simultaneous reception of ordinary television signals was also used successfully, to even greater accuracy. The idea was therefore to begin by determining the propagation delays and then check them episodically. This was done by transporting operating caesium clocks, by road and on commercial flights. These methods will be described in Chapter 5. For the moment, let us just say that large regions of the globe could be covered in this way, including North America and

Europe, with a routinely operating network of time comparisons accurate to ±1 μs.

From this moment on, mean atomic time scales could be constructed from clock readings. These atomic times were realised by calculating corrections to the readings of the participating clocks. This led to a new problem. Optimal algorithms had to be established in order to build time scales from data of varying quality (see Chapter 7). The Bureau international de l'heure replaced its mean frequency standard by a mean atomic clock on 1 January 1969, without introducing any discontinuity into its atomic time scale. In 1973, it set up an algorithm and a world coordination that have survived to this day with only minor modifications (see Chapter 7) in order to keep up with technical progress in standards and time transfers. A considerable improvement in remote clock comparisons was provided in 1983 by the satellite system known as the Global Positioning System (GPS). The GPS has gradually superseded Loran-C in timing laboratories and reduced uncertainties to a few nanoseconds. Recently, time comparisons by exchange of signals relayed by telecommunications satellites began to be used operationally, with improved results, especially for frequency comparisons.

National organisations, in particular the US Naval Observatory in Washington, were also producing atomic time scales. The need for agreement upon a single reference led several key international organisations to recommend the use of the BIH scale, including the International Astronomical Union (IAU) in 1967, the International Union of Radio Science (URSI, Union radioscientifique internationale) in 1969, and the International Radio Consultative Committee (CCIR) in 1970. Ultimate consecration came from the organisations in the Metre Convention, with the definition of *International Atomic Time* (TAI, Temps atomique international) formulated in 1970 by the International Committee for Weights and Measures (CIPM) in the following terms:

> International Atomic Time (TAI) is the time reference coordinate established by the Bureau international de l'heure on the basis of the readings of atomic clocks operating in various establishments in accordance with the definition of the second, the unit of time of the International System of units.

This definition, or rather, this official recognition, was ratified by the 14th General Conference on Weights and Measures in 1971. After a transition period that had extended from 1955 to 1971, time measurement had finally become coherent again, with the unit of time and the world time scale based on the same physical phenomenon, a specified atomic transition. Time, like

all other fundamental physical quantities, was now measured by laboratory standards. Nothing could prevent it from taking its place amongst those other quantities, within the International Bureau of Weights and Measures at Sèvres in France. This happened in 1988.

Since the atomic time scale at the BIH had been maintained continuously since 1955, it was retroactively called TAI, even though computational methods had evolved.

Whilst the accuracy and stability of atomic time standards and the TAI were improving, it became clear that a fuller definition within the context of general relativity would soon be needed. We should mention the extremely apposite ideas put forward on this subject by Becker as early as 1967 [4.23]. But it was not until 1980 that the Consultative Committee for the Definition of the Second (CCDS, a CIPM committee which became in 1997 the Consultative Committee for Time and Frequency) gave a relativistic definition of the TAI, a definition that was completed by the IAU in its resolution A4 in 1991, and which should be slightly modified following the redefinition of TT in 2000 (see Section 3.3.2e). An important consequence of progress in time measurement for the history of science is that general relativity has become an essential tool in metrology and in practical applications. Indeed the topic of 'metrology and relativity' is the subject of a report prepared under the auspices of the CCDS [4.24].

4.5 Coordinated Universal Time

During the period when International Atomic Time was being set up, the idea of coordinating the broadcast of radio time signals (which had continuously sent out Universal Time) was gradually maturing.

The initiative was taken by Great Britain and the United States. In each of these countries, from August 1959, signals were broadcast to mark locally obtained atomic time scales. At this time, the value of the caesium transition frequency that would later serve to define the second was already in general use. As noted, this value had been related to Ephemeris Time. Consequently, atomic time scales were observed to run fast relative to Universal Time. In order to build an atomic time in approximate agreement with UT, a relative frequency offset y_U had to be introduced. The same value of y_U was chosen in both countries and time signals initially synchronised. Since UT involves unpredictable variations, time adjustments had to be made in order to avoid changing y_U too often. These adjustments were made in steps of 50 ms at agreed times in the light of observations of UT.

As many time signals had their second markers strictly linked to the carrier frequency (e.g., for an emission at 10 MHz, the second marker begins every 10^7 cycles), the atomic frequency could be found from y_U, either by using the spacing of the markers, or by measuring the carrier frequency.

This coordinated effort was soon joined by other countries and as early as 1960 the International Union of Radio Science asked the BIH, which was responsible for unifying UT, to fix the offset y_U each year. Its value would then remain constant throughout the year. Subsequently, the BIH also decided the dates when time steps should be introduced.

This system evolved fairly rapidly. In 1963, the time steps went from 50 ms to 100 ms. Until 1965, the time scale that was more or less common to all the coordinated signals, and which had been spontaneously christened *Coordinated Universal Time* (UTC), still had not been given a rigorous definition. For example, the BIH considered UTC as an average of the emission times of the coordinated signals. An important step in the development of UTC was the BIH's decision to define it, from 1 January 1965, by a mathematical relation with the atomic time of the BIH, which later became the TAI. This relation was

$$UTC - TAI = y_U(TAI - TAI_0) + B \ , \tag{4.7}$$

where TAI_0 is an arbitrarily fixed date origin and B a constant modified by steps so as to maintain the condition

$$|UTC - UT2| < \varepsilon \ , \tag{4.8}$$

with UT2 being Universal Time corrected for its quasi-periodic annual irregularity and ε the agreed tolerance.†

This definition of UTC was used until 1972. However, the need for the frequency offset y_U, sometimes changed at the beginning of the year, became more and more burdensome. (The largest modification was made on 1 January 1966, when it went from -1.5×10^{-8} to -3.0×10^{-8}.) One of the drawbacks was that certain emissions had their carrier frequency linked to the shifted frequency of UTC, so that technical alterations had to be carried out on the emitters. For this reason, the International Radio Consultative Committee had accepted in 1966 that certain time signals would be broadcast with $y_U = 0$, at the expense of frequent time steps. The size of these steps was fixed at 200 ms and they were only imposed at the beginning of a month when needed to keep

† Note that a time scale such as UT1 is designated by its acronym written in upright capitals. Its reading at an instant specified in terms of a time scale θ is written in italics, e.g., $UT1(\theta)$. The argument θ, value of θ, is not always shown explicitly.

ε smaller than 200 ms. This *Stepped Atomic Time* was only adopted by the German signals (Federal Republic of Germany) and one emitter in the USA. It is significant mainly because it prepared the way for the currently used UTC system, adopted in 1972.

We may wonder why time signals, and hence world time, were not adjusted to International Atomic Time when this time scale was officially recognised in 1971. Of course, TAI runs fast relative to UT. In the very long term, 12 o'clock would have occurred well before the Sun had transited the Greenwich meridian. However, this deviation is of the order of one hour every thousand years. When we realise that in Britain (and other countries), we accept changing the clocks by one hour twice a year and hence living through the summer an hour or so ahead of solar time, the idea of putting the clocks back one hour every ten centuries would appear a rather minor adjustment.

Objections to a general use of TAI are partly sentimental, harking back to the cherished belief that time is determined by the celestial motions, and partly technical. Sailors who use astronomical means of navigation were firmly opposed to the idea of no longer broadcasting Universal Time, which they needed for their fix. The correction required to obtain UT from TAI could be predicted to the nearest second several years ahead, and this is accurate enough for astronomical positioning at sea. However, maritime organisations considered that the risk of error (e.g., a correction made with the wrong sign) was too great. It was also risky to broadcast two types of time signal, one in UT and the other in TAI. In 1970, after long debate and much disagreement [4.25], the CCIR was ready to recommend that UTC be defined with $y_U = 0$ and B equal to a whole number of seconds, modified by leap seconds. At first, the tolerance ε (now referred to UT1 and not UT2) was fixed at 0.7 s, before being set at 0.9 s in 1974. Moreover, time signals had to carry coded information that could be used to obtain UT1 to within 0.1 s.

The new UTC was established according to rules whose details were laid out by the CCIR (see Chapter 7), and was officially recognised by the 14th CGPM in 1971. It was put into practice on 1 January 1972, requiring a time step that was not a round fraction of 1 s. It has been in general use ever since. The system is being increasingly criticised, but attempts to form a continuous time scale like TAI have failed. The leap seconds introduced at intervals of one to a few years, depending on the whims of our planet's rotation, do indeed have their drawbacks, but they must be accepted if we wish to have a single world time. Humankind is perhaps not ready to move to a purely atomic time, and so the Earth's rotation continues to regulate world affairs, at least as far as common usage is concerned.

4.6 Final comments

Since 1972, the measurement of time has officially rested upon the hyperfine transition of caesium, even though the Earth's rotation is still followed in an approximate way. Some believed that astronomical measures of time were to make a comeback when the first millisecond pulsar was discovered in 1982. We shall see in Chapter 8 that such bodies cannot provide a definition of the second, no matter how regular their rotation may be.

It will be noted how one definition of the second succeeds another. Under the pressure of scientific and technical requirements, the use of a new definition may even precede its official adoption.

The situation for time scales is quite different. New ones are created, associated with the current definition of the second, but formerly used scales do not disappear. This can be explained to some extent by their historical role. They have been used to date events and consequently remain in the archives. Naturally, there is every reason to maintain such archives in their original form. This is true for astronomical observations, dated in the apparent time of the place in which they were made. If we wish to make use of them, the dates can be converted into Universal Time or Ephemeris Time, although this involves some degree of interpretation and the result is not therefore unique, whereas the original document remains authentic. As far as atomic time is concerned, it has only existed since 1955.

Another reason is that the various time scales, based on different physical phenomena, retain an intrinsic meaning. Universal Time represents the Earth's rotation. Ephemeris Time and the various coordinate times defined in the context of general relativity are dynamical times, associated with dynamical theories. Who knows whether their long term comparison with atomic time might not lead to crucial data? In the same spirit, a *Pulsar Time* is now under construction.

The survival of formerly used time scales is not a manifestation of conservatism. Neither is the definition of new scales a sign of perfectionism, for it is required by applications. We must accept that we have moved into an era of multiple time scales. Unfortunately for the reader, we shall be unable to simplify without being guilty of omission. At least, it will be appreciated that the idea of a universal time has been safeguarded for public use.

5

Clock time

5.1 Introduction

The base unit of the International System of Units (SI units) in the time domain is a unit of duration, the second. It would therefore have been appropriate to refer to the atomic standards which provide this unit as time standards, or time interval standards. In practice, however, these terms are rarely used, with the term *frequency standard* being preferred, for it is indeed the frequency that they actually supply.

According to the official prescription [5.1], we must distinguish a *primary standard* 'that is designated or widely acknowledged as having the highest metrological quality and whose value is accepted without reference to other standards of the same quantity' from a secondary standard 'whose value is assigned by comparison with a primary standard of the same quantity'. We shall follow this prescription even though the qualitative aspects in the definition of primary standards are somewhat embarrassing. Indeed there exist frequency standards that are certainly not secondary, and yet which could in no sense be considered as primary.

Anyone building a frequency standard generally attempts to produce a wave as close as possible to a sine wave, thereby aiming to ensure spectral purity, and with frequency as close as possible to the nominal value. But the distinct periods of this wave are not identified and numbered. By definition, a clock supplies further information, precisely by identifying these periods. This is usually achieved by numbering pulses at a frequency of 1 Hz which are called *second markers*, or more familiarly, *second pulses*. The word 'clock' also implies the idea of continuous operation over long periods, e.g., of several years. A clock therefore produces a time scale (its *proper time*, in relativistic terminology).

We begin this chapter by discussing ways of describing instability in the frequency of oscillators and fluctuations in clock time scales. We will then be able to characterise the stability of the various standards to be presented, including in particular quartz oscillators which play a key role in time measurement at all levels of quality.

It will be convenient to extend the results of our study to time scales which are not produced by real clocks, as well as to those that have an astronomical basis. We must then consider that such a scale arises from a fictitious clock. For example, we will see that the (frequency) instability of International Atomic Time is about 10^{-15} over sampling periods of 40 days, although the word 'frequency' is quite often, and wrongly, omitted here.

When we consider a coordinate time scale, we must imagine that it arises from the proper time of a fictitious clock, applying the theoretical relationship between the proper time and the coordinate time in the coordinate system used. The characterisation of the instability and inaccuracy then refer to this clock.

Although work on instability has been stimulated by advances made with frequency standards, it has applications in all those domains involving long time series of measurements, notably in geophysics. The parameters describing the Earth's rotation may be cited as an example, one of these being Universal Time itself.

When time measurement was based upon astronomy, the primary standard was provided by celestial motions. These supplied us with time, that is, with *the* time as we think of it when we read our watch. (The vagueness of terminology used to speak of time is quite daunting!) Frequency was a quantity derived from this, so that all frequency standards were then secondary. But the arrival of atomic frequency standards raised a new problem, concerning the accuracy of their frequency. Accuracy is often perceived as a qualitative notion. However, in the study of frequency it has a precise definition that is somewhat peculiar to this quantity. Inaccuracy can be quantitatively characterised, as we shall explain in Section 5.5.

In the remainder of the chapter, we consider how frequency standards and clocks can be compared. Another type of fictitious clock will be encountered in this context, that is, the clock associated with a time signal. These signals, either broadcast, or disseminated by cable or any other method, are made up of time markers carrying a date. They can therefore be viewed as clocks. However, the transmission delays mean that we must specify whether we consider them at emission or in a given place of reception.

5.2 Frequency and time instability

5.2.1 Definitions

Fluctuations in the frequency of an oscillator in the neighbourhood of its nominal frequency result from perturbations such as thermal noise in electronic components, intrinsic noise of the resonator providing the oscillation frequency, instrument ageing, and environmental variations. These fluctuations constitute what is called *frequency instability.*†

If the oscillator is operating as a clock, its frequency instability leads to fluctuations in the time scale it produces compared with the ideal time constructed on the basis of its nominal frequency. This is known as *time instability.*

For both frequency and time, we may distinguish random and deterministic instabilities. The former refers to fluctuations that would seem to be random. The latter refers to those that appear to be functions of external parameters, such as age and temperature, although it is not known a priori what form the functions take (for then, a corresponding correction could be made). In the following, we shall be mainly concerned with the random component, assuming that the deterministic component, if one is observed, has been measured and subtracted off.

The properties of time and frequency instabilities can be characterised either in the *time domain* or in the *spectral domain.* In the time domain, we study the behaviour of average frequency samples of varying duration. Naturally, a minimal sampling interval is set by measurement conditions. In the spectral domain, we exploit properties of the Fourier transform of the function representing the frequency variations or time deviations. Here, too, we only have a discrete series of measured values, which implies the existence of a (Fourier) cutoff frequency.

Since the beginning of the 1960s, a great many theoretical and experimental studies have been carried out to characterise frequency instability and the time instability of oscillators. This research has enabled specialists to reach agreement concerning the use of variances suited to such a characterisation. The results obtained have been adopted by the various professional and international organisations [5.2, 5.3].

5.2.2 Notation

Readers must now prepare themselves for a difficulty that the authors have not been able to spare them, viz., differences of notation. In this section, devoted

† Although it is not strictly correct, use of the expression 'frequency stability' rather than 'frequency instability' is tolerated.

to the statistical characterisation of oscillators, we have adopted the notation established by general usage. It is unfortunate that it sometimes differs from the notation normally used to refer to the same physical quantities in other theoretical contexts, such as relativity physics, time scales and frequency standards. The problem is that the alphabets available are not sufficient to represent the range of parameters and mathematical models needed to describe physical reality. It is likely that even greater confusion would have been caused by modifying the traditional notation and the notation specific to each area has therefore been maintained. The reader will be warned when a symbol changes meaning in such a way that it might lead to misinterpretation.

The signal v_C delivered by an oscillator C can be written

$$v_C(t) = V_{0,C} \sin\left[2\pi \nu_{0,C} t + \phi_C(t)\right] , \tag{5.1}$$

where we have assumed that the amplitude is constant. The quantity $\nu_{0,C}$ is the constant nominal frequency and t the ideal proper time based on the definition of the second. The latter is equivalent to τ, used in this sense in previous chapters. The phase ϕ_C contains the deviations, both random and systematic, relative to the ideal periodic variation.

The instantaneous frequency $\nu_C(t)$ is defined by

$$\nu_C(t) = \nu_{0,C} + \frac{1}{2\pi} \frac{d\phi_C(t)}{dt} . \tag{5.2}$$

The relative frequency offset of the oscillator, first introduced in Section 4.4.2a, is given by

$$y_C(t) = \frac{\nu_C(t)}{\nu_{0,C}} - 1 . \tag{5.3}$$

It represents as a relative value the instantaneous departure of the oscillating frequency from its nominal value. This is very small for the highly stable oscillators we shall be considering.

An oscillator often supplies a range of frequencies with specified nominal values, e.g., 1 MHz, 1 kHz, 1 Hz. By its construction, the actual frequencies remain in strictly the same ratio as the nominal frequencies, so that there is only one normalised frequency for all these frequency values.

The clock reading is incremented by the nominal value $1/\nu_{0,C}$ of the period every time the total phase $2\pi \nu_{0,C} t + \phi_C(t)$ increases by 2π. The time indicated since the time origin is therefore

$$t_C = t + x_C(t) , \tag{5.4}$$

where we have set

$$x_C(t) = \frac{\phi_C(t)}{2\pi \nu_{0,C}} . \tag{5.5}$$

The quantity t_C is the proper time of the clock C. The proper time deviation resulting from imperfection in the oscillator is thus equal to $x_C(t)$. This is the time instability of the clock with respect to an ideal reference time scale.

From the above definitions, it is easy to show that

$$y(t) = \frac{dx(t)}{dt} , \tag{5.6}$$

where $y(t)$ represents the relative frequency offset of an oscillator or a clock as a function of t and $x(t)$ its proper time deviation.

5.2.3 Measurements in the time domain

Many methods have been devised to measure frequency and time instability [5.4, 5.5, 5.6]. Each requires two frequency generators or two clocks, one being taken as reference. We begin by assuming that the latter has negligible frequency instability compared with the instrument under investigation. We shall consider the contrary case in Section 5.2.8.

We restrict ourselves to illustrating the measurement of time instability on the one hand, and frequency instability on the other, by two commonly used and easily interpreted methods.

(a) Measurement of time instability

The experimental setup is shown schematically in Figure 5.1. The clock under investigation and the reference clock supply periodic pulses, e.g., every second. The clocks are connected to the two inputs of an intervalometer. This instrument registers the moment when the rise of an applied pulse reaches a predetermined level. Counting is triggered by the arrival of a pulse at input 1 and it is stopped by the first pulse to follow at input 2. The intervalometer is generally controlled by the reference clock.

The measurement gives the value x_k of $x(t)$ modulo one second at the time t_k when a pulse generated by the clock under test arrives at input 2. Using the best instruments, the uncertainty in the measured value is of the order of 10 ps. A systematic error arises from the propagation delay in the connecting cables. If it is constant, it plays no role in the statistical analysis of the time instability since the latter appeals to differences between values of x_k, as we shall soon see.

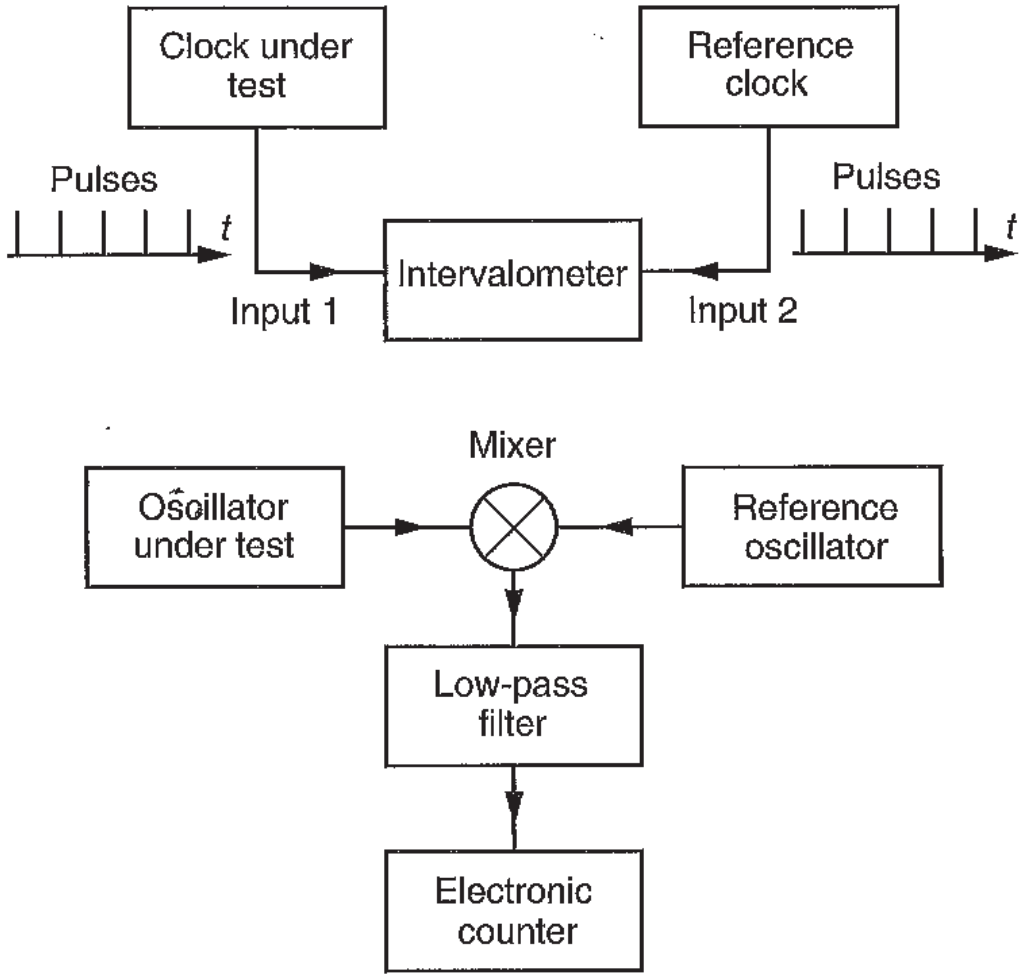


Fig. 5.1. Schematic view of the measurement of time instability (top) and frequency instability (bottom).

The set of values x_k obtained in this way constitutes a time series defined at equally spaced times t_k. Following common usage in the field of instability characterisations, the time interval between successive values of t_k will be denoted τ. Strictly speaking, this is an interval of proper time and, according to the notation introduced in Chapter 3, should be denoted $\Delta\tau$.

(b) Measurement of frequency instability

The method to be described is suited to the case where the two oscillators, or more generally, the two generators, produce different but closely spaced frequencies. For example, their values may be close to 10 MHz, whilst their difference could be of the order of 1 Hz.

Let $\nu_C(t)$ be the frequency of the oscillator under investigation, and ν_R that of the reference oscillator, supposed constant. As shown in Figure 5.1, the oscillators are connected to a mixer. This is a non-linear device which, to a first approximation, generates the product of the two signals it receives. Its output signal contains spectral components at frequencies $|\nu_C - \nu_R|$ and $\nu_C + \nu_R$. The

low frequency component is isolated by filtering and we thus detect the beat frequency $\nu_b = |\nu_C - \nu_R|$.

The relative change in ν_b can be written

$$\frac{\Delta\nu_b}{\nu_b} = \frac{\nu_C}{\nu_b}\frac{\Delta\nu_C}{\nu_C} . \tag{5.7}$$

This is considerably greater than the relative change in ν_C, because of the presence of the multiplicative factor ν_C/ν_b. Using the numerical values proposed above, the latter is of the order of 10^7.

The frequency ν_b is measured with the help of an electronic counter whose time base is controlled by the reference oscillator. It measures the duration of n periods of the input signal. This duration is equal to $\tau = n/\nu_b$. The counter thus provides the mean value of ν_b over the time interval τ, denoted by $\bar{\nu}_b^\tau$.

The uncertainty in this measurement depends on the signal-to-noise ratio of the beat signal. With appropriate equipment, this uncertainty is close to $10^{-7}/\tau$ when ν_b lies in the vicinity of 1 Hz and the pass band of the low-pass filter is equal to 10 Hz. It follows that the uncertainty in the value of $\Delta\nu_b/\nu_b$ is of the order of $10^{-14}/\tau$ in the example considered here. This can often be neglected.

To simplify, let us suppose that the dead time between two measurements is very small compared with the time interval τ. Successive measurements thus produce a time series of values of $\bar{\nu}_b^\tau$, at the discrete times t_k such that $t_{k+1} - t_k = \tau$. We then extract the average value of $\bar{\nu}_b^\tau$ and hence of ν_b. Using (5.3) and (5.7), we may deduce a time series of values of the relative frequency offset of the oscillator under test, but averaged over the time interval τ.

As the value of $\bar{y}_k^\tau$ is defined between times t_k and t_{k+1}, equation (5.6) shows that the following relation holds between the measured values of the frequency instability $\bar{y}_k^\tau$ and the time instability x_k :

$$\bar{y}_k^\tau = \frac{1}{\tau}[x_{k+1} - x_k] . \tag{5.8}$$

5.2.4 Measurements in the spectral domain

Measurements in the spectral domain [5.4, 5.5, 5.6] provide us with the power spectral density† of the time and frequency fluctuations under investigation. The latter are denoted $S_x(f)$ and $S_y(f)$, respectively. The frequency f is the analysis frequency, also known as the Fourier frequency, and not to be confused

† The power spectral density of a random process at frequency f is equal to $\sigma^2/\Delta f$, where σ^2 is the mean squared fluctuation of the process filtered through a very narrow window of width Δf centred on the value f.

with the oscillation frequency. By (5.6),

$$S_y(f) = 4\pi^2 f^2 S_x(f) \,. \tag{5.9}$$

5.2.5 *Modelling frequency and time fluctuations*

Experimentally, it is only possible to obtain a limited number of samples of $\bar{y}_k^\tau$ since the total measurement time is necessarily finite. There may only be a handful in the case where the time or frequency instability is to be characterised over a period τ that exceeds a few days. It is therefore only possible to calculate an estimator of the variance of the fluctuations under investigation (see Section 5.2.6). This estimator is itself a random variable, in the sense that a different value would be obtained if the measurement were to be repeated. The uncertainty is all the greater when the sample number decreases. It is then difficult, even impossible, to deduce the power spectral density of the frequency or time fluctuations for low values of the Fourier frequency. However, it is often needed when studying the properties of systems that use atomic frequency and time standards, such as certain navigation and telecommunications systems to be discussed in Chapter 9.

In order to get around this problem, a mathematical model of time and frequency fluctuations has been developed. The validity of the model is assessed by checking that experimental results agree with the model or its predictions, within the bounds of experimental uncertainty.

The model which gives a satisfactory account of instabilities of random origin in the frequencies of oscillators that we are considering here has the following characteristics. The process is ergodic. The frequency instability, represented by $y(t)$, has zero mean value. The probability density of $y(t)$ is Gaussian. Finally, the power spectral density of the relative frequency fluctuations is represented by a polynomial in f which contains up to five terms. For $f > 0$,

$$S_y(f) = \sum_{\alpha=-2}^{2} h_\alpha f^\alpha \,, \tag{5.10}$$

where α is an integer in the range -2 to $+2$. The values of the constants h_α depend on the source. They may be zero for certain values of α. The corresponding value of $S_x(f)$ follows from (5.9) and (5.10).

Table 5.1 gives the usual terms for the various kinds of independent noise that may contribute to time and frequency instability.

The physical origins of white phase and frequency noise are understood. They are associated with thermal noise in the radio region or shot noise from

Table 5.1. *Terms used to describe the five noise processes contributing to time and frequency instability.*

α	Noise
2	White phase noise
1	Flicker phase noise
0	White frequency noise
−1	Flicker frequency noise
−2	Random walk frequency noise

a particle flux. The origins of the other types of noise are less clear. However, the great number of experimental results acquired are witness to their reality.

5.2.6 Characterising frequency instability

In general, the variance of a random process is related to the integral of its power spectral density over the whole frequency range. For $y(t)$, we find that

$$E\left\{y^2(t)\right\} = \int_0^\infty S_y(f)\,\mathrm{d}f\ , \tag{5.11}$$

where $E\{\ \}$ denotes the mathematical expectation. It is immediately obvious that, when $S_y(f)$ has the form (5.10), the integral diverges. For $\alpha \geq -1$, the divergence occurs for $f \to \infty$. However, frequency fluctuations are always filtered through a low-pass filter. This mathematical difficulty does not therefore correspond to a real physical problem. More serious is the divergence for $f \to 0$ when $\alpha = -1, -2$. Of course, a frequency of zero can never be attained, but it can be shown that the mathematical divergence implies that the estimated value of the variance depends on the total measurement time. This is unsatisfactory. The situation is even worse when we consider $E\left\{x^2(t)\right\}$ because the power spectral density $S_x(f)$ contains terms in $f^{\alpha-2}$. The corresponding integral therefore diverges at the lower bound for four out of the five values of α, i.e., those with $\alpha \leq 1$.

It can be shown that for $\alpha \leq 1$, it is the first or second order increments in $x(t)$, and hence in the phase fluctuations (see (5.5)), that are stationary. The above difficulties are then removed if the statistical treatment of the time and frequency fluctuations appeals to the mean square of these increments. Many variances can then be considered. We shall discuss only the most commonly used amongst them.

(a) Definition of the Allan variance

We begin by discussing the Allan variance, also known as the two-sample variance. Its definition presupposes an unlimited number of samples. It is equal to half the mean value of the two-sample variance of the quantities $\bar{y}_k^\tau$, where the samples are taken in an adjacent series, i.e., with no dead time between successive values.† We have [5.7]

$$\sigma_y^2(\tau) = \frac{1}{2} E\left\{\left[\bar{y}_{k+1}^\tau - \bar{y}_k^\tau\right]^2\right\} . \tag{5.12}$$

Using (5.8), we obtain

$$\sigma_y^2(\tau) = \frac{1}{2\tau^2} E\left\{\left[x_{k+2} - 2x_{k+1} + x_k\right]^2\right\} , \tag{5.13}$$

which brings in the second order increment in x_k. Equations (5.12) and (5.13) can be used to characterise frequency instability in the time domain.

The Allan variance can also be expressed in terms of the power spectral density $S_y(f)$. It can be shown that

$$\sigma_y^2(\tau) = \int_0^\infty |H_A(f)|^2 S_y(f)\, \mathrm{d}f , \tag{5.14}$$

where $|H_A(f)|^2$ is the squared modulus of a transfer function which represents, in the frequency domain, the properties of the digital filtering associated with calculation of the Allan variance. This transfer function is given by

$$|H_A(f)|^2 = 2\frac{\sin^4(\pi f \tau)}{(\pi f \tau)^2} . \tag{5.15}$$

When $f \to 0$, $|H_A(f)|^2$ varies as f^2, guaranteeing the convergence of the integral at the lower limit for all the typical values of α.

The low-pass filter with cutoff frequency f_h, which is always present at the input or inside the equipment measuring $\bar{y}_k^\tau$, is taken into account by multiplying $|H_A(f)|^2$ by the power transfer function of this filter, denoted $|H_F(f)|^2$. We assume here that

$$|H_F(f)|^2 = \begin{cases} 1 & \text{for } f < f_h , \\ 0 & \text{for } f \geq f_h . \end{cases} \tag{5.16}$$

† The factor of 1/2 arises because the two-sample variance is a special case of a more general one, the N-sample variance. When the random variables are relative frequency fluctuations averaged over the time interval τ, the N-sample variance is

$$\sigma^2(N, \tau) = \frac{1}{N-1} \sum_{n=1}^{N} \left(\bar{y}_n - \frac{1}{N} \sum_{k=1}^{N} \bar{y}_k \right)^2 .$$

Table 5.2. *Correspondence between the various noise components perturbing an oscillator and the Allan variance of the normalised frequency fluctuations when* $2\pi f_h \tau \gg 1$.

$S_y(f)$	$\sigma_y^2(\tau)$
$h_2 f^2$	$3h_2 f_h/4\pi^2\tau^2$
$h_1 f$	$h_1[1.04 + 3\ln(2\pi f_h \tau)]/4\pi^2\tau^2$
h_0	$h_0/2\tau$
$h_{-1} f^{-1}$	$2h_{-1}\ln 2$
$h_{-2} f^{-2}$	$2\pi^2 h_{-2}\tau/3$

The value of the Allan variance is only affected by this filtering for $\alpha = 1, 2$, provided that

$$f_h \gg \frac{1}{2\pi\tau} . \tag{5.17}$$

(b) Variation of Allan variance with sampling period τ

Calculating the integral on the right hand side of (5.14) with $|H_A(f)|^2$ replaced by $|H_A(f)|^2 |H_F(f)|^2$, we can establish the expression for $\sigma_y^2(\tau)$ associated with each noise component perturbing the oscillator, when the condition (5.17) is satisfied. The correspondence is shown in Table 5.2.

The functional dependencies of $S_y(f)$ and the Allan deviation $\sigma_y(\tau)$ are usually represented on a log–log graph, as in Figure 5.2, which shows their asymptotic behaviour. When a given type of noise dominates over at least one decade of f or τ values, such graphs reveal the value of α, by observing the slope of the curve. They can thus be used to identify the type of noise involved, if $\alpha \leq 0$. For example, the section of the graph of $\sigma_y(\tau)$ with zero slope is associated with flicker frequency noise and is commonly called the flicker floor. The Allan deviation varies as τ^{-1} for white and flicker phase noise. They cannot therefore be distinguished in this way. One way to remedy this situation is to take advantage of the different dependence of $\sigma_y^2(\tau)$ on the pass band f_h of the low-pass filter. Another is to consider the modified Allan variance discussed below.

In the above, it has been assumed that there is no dead time between successive samples of $\bar{y}_k^\tau$. If this is not the case, as happens in practice, bias is introduced into the calculation of $\sigma_y^2(\tau)$, except for white frequency noise. The biases have been tabulated [5.5].

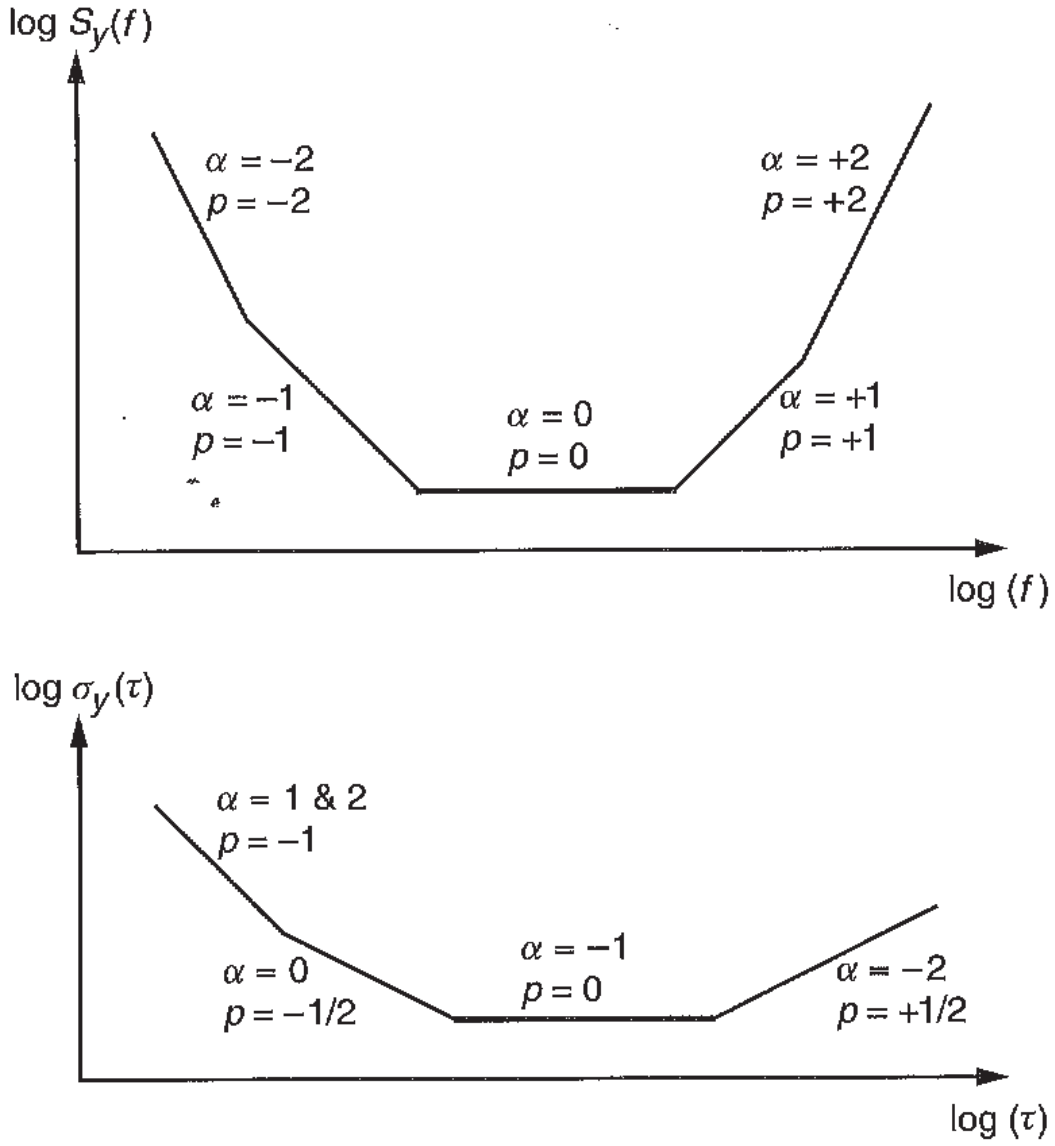


Fig. 5.2. Asymptotic log–log representation of the variation of $S_y(f)$ and $\sigma_y(\tau)$ with their respective arguments f and τ, for the different values of α. Slopes are given by the values of p.

Figures 5.6 and 6.11 show examples of how the Allan deviation varies with τ.

(c) Practical calculation of the Allan variance

In the real world, the number of available values of x_k or $\bar{y}_k^\tau$ is finite. Consequently, it is only possible to obtain an estimate of the Allan variance. We shall now describe how the estimator for $\sigma_y(\tau)$ can be calculated in practice.

Generally speaking, measurements are not repeated for different values of τ. In fact, we determine a time series of $M + 1$ values of x_j at times separated by τ_0, or a time series of M values of $\bar{y}_j^{\tau_0}$ over the sampling period τ_0. We assume

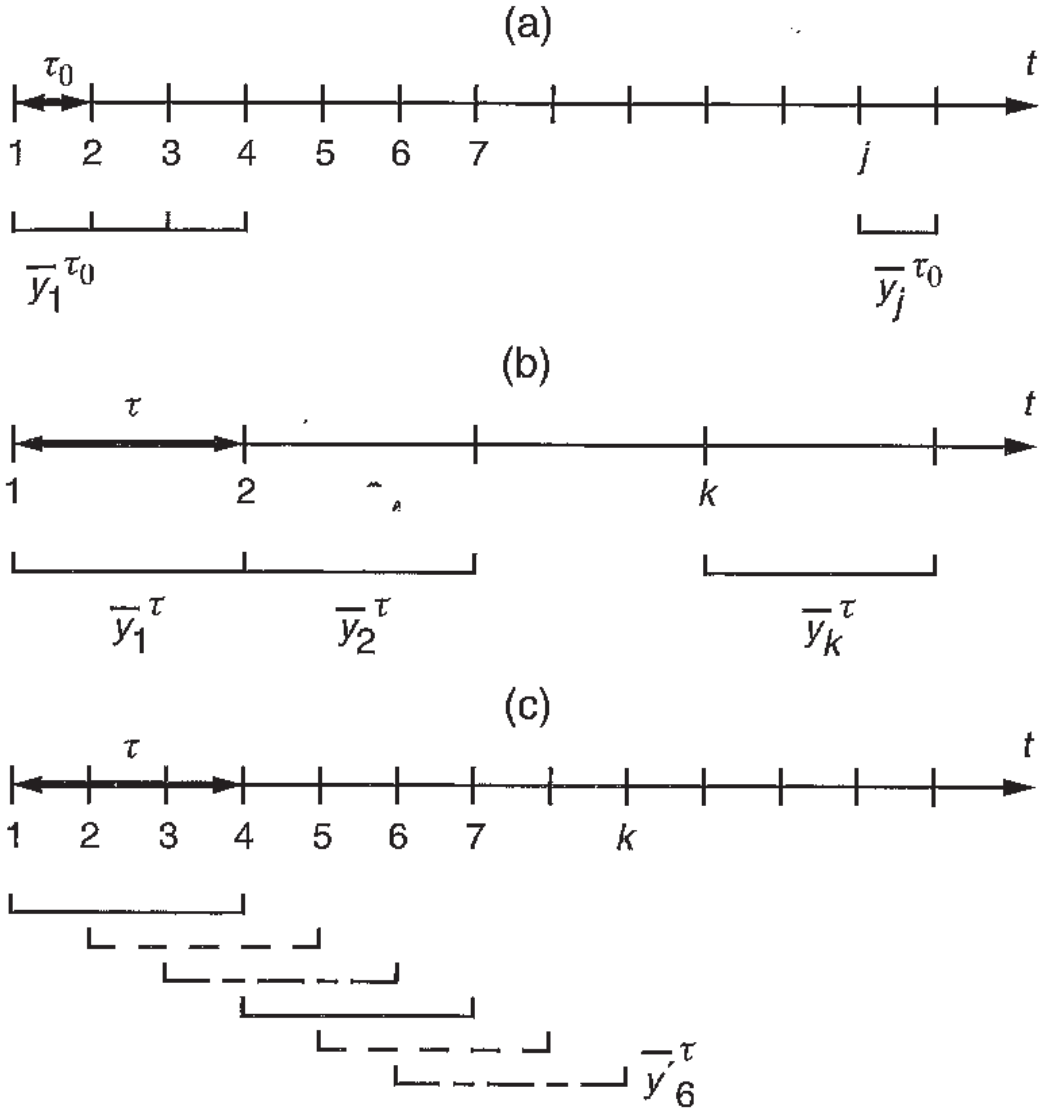


Fig. 5.3. Practical calculation of the Allan variance. (a) Time intervals over which initial data set is defined. (b) Illustration of the first method for processing the data. (c) Illustration of the second method for processing the data.

once again that the dead time is negligible. Figure 5.3a shows the basic time intervals of duration τ_0 over which the $\bar{y}_k^{\tau_0}$ are defined.

The data is then grouped together in order to deduce the Allan variance for $\tau = n\tau_0$, where n is a positive integer.

The simplest and most direct combination is shown in Figure 5.3b, taking $n = 3$. A time series is formed from adjacent values of $\bar{y}_k^{\tau}$ by calculating the average variation in the normalised frequency over the time interval $\tau = n\tau_0$ with one of the following relations:

$$\bar{y}_k^{\tau} = \frac{1}{n\tau_0}\left[x_{nk+1} - x_{n(k-1)+1}\right] , \tag{5.18}$$

or

$$\bar{y}_k^{\tau} = \frac{1}{n} \sum_{j=n(k-1)+1}^{nk} \bar{y}_j^{\tau_0} \,. \tag{5.19}$$

The number of pairs of values of $\bar{y}_k^{\tau}$ that can be formed is

$$p_1 = \text{int}\left(\frac{M}{n}\right) - 1 \,. \tag{5.20}$$

The estimated value $\hat{\sigma}_y^2(\tau)$ of the Allan variance is then calculated by one of the following relations:

$$\hat{\sigma}_y^2(\tau) = \frac{1}{2p_1(n\tau_0)^2} \sum_{k=1}^{p_1} \left[x_{n(k+1)-1} - 2x_{nk+1} + x_{n(k-1)+1}\right]^2 \,, \tag{5.21}$$

$$\hat{\sigma}_y^2(\tau) = \frac{1}{2p_1} \sum_{k=1}^{p_1} \left[\bar{y}_{k+1}^{\tau} - \bar{y}_k^{\tau}\right]^2 \,. \tag{5.22}$$

The characterisation of the same oscillator via different time series of the same length yields different values of the Allan variance estimator. The average value of the estimators is equal to the mathematical expectation defined by (5.12). The uncertainty in $\hat{\sigma}_y(\tau)$ characterised by its own standard deviation is equal to $K_\alpha/(p_1 + 1)^{1/2}$ when p_1 is greater than a few units [5.8]. The parameter K_α ranges between 0.75 and 0.99 when α varies from -2 to $+2$.

Another way of combining the experimental data is more commonly used in practice because it makes better use of them, leading to smaller uncertainties [5.9]. Instead of forming the set of values of $\bar{y}_k^{\tau}$ as indicated above, a different set can be obtained from

$$\bar{y}_k'^{\tau} = \frac{1}{n\tau_0}(x_{k+n} - x_k) \,, \tag{5.23}$$

or equivalently from

$$\bar{y}_k'^{\tau} = \frac{1}{n} \sum_{j=k}^{k+n-1} \bar{y}_j^{\tau_0} \,. \tag{5.24}$$

Figure 5.3c shows the time intervals over which the new $\bar{y}_k'^{\tau}$ are defined, taking $n = 3$. These intervals overlap. However, we consider the values of $\bar{y}_k'^{\tau}$ belonging to adjacent intervals in order to estimate the Allan variance. Such values would be $\bar{y}_{k+n}'^{\tau}$ and $\bar{y}_k'^{\tau}$, for example. The number of such pairs is p_2 given by

$$p_2 = M - 2n + 1 \,. \tag{5.25}$$

The new estimator of the Allan variance over the sampling period τ is then calculated using one of the following relations:

$$\hat{\sigma}_y^2(\tau) = \frac{1}{2p_2(n\tau_0)^2} \sum_{k=1}^{p_2} \left[x_{k+2n} - 2x_{k+n} + x_k\right]^2 , \tag{5.26}$$

$$\hat{\sigma}_y^2(\tau) = \frac{1}{2p_2} \sum_{k=1}^{p_2} \left[\bar{y}_{k+n}^{\prime\tau} - \bar{y}_k^{\prime\tau}\right]^2 . \tag{5.27}$$

It can be shown that this too has an average value equal to the mathematical expectation defined by (5.12).

The method of estimating the uncertainty in the Allan deviation thus calculated involves more complex calculations than the previous one [5.10, 5.11]. The main result is that, for given M and n, the number of pairs p_2 is considerably greater than p_1 (unless we approach $M/2$) and it follows that the relative uncertainty in $\hat{\sigma}_y(\tau)$ is reduced.

(d) Allan variance in the presence of systematic effects

There is nothing to stop us calculating the Allan variance when the relative frequency offset $y(t)$ features systematic variations. For example, the perturbation might be a linear frequency drift described by

$$y(t) = at , \tag{5.28}$$

or a sinusoidal modulation described by

$$y(t) = a' \cos 2\pi f_M t . \tag{5.29}$$

Parameters a and a' represent the slope of the frequency drift and the relative amplitude of the frequency modulation, respectively. The frequency of the modulation is f_M such that $f_M < f_h$.

In the case of a linear drift, the definition (5.12) implies

$$\sigma_y^2(\tau) = \frac{1}{2} a^2 \tau^2 . \tag{5.30}$$

On a log–log graph representing the variation of $\sigma_y(\tau)$, we then obtain a straight line with unit slope. Added to the other usual noise effects, it begins to play a dominating role for large values of τ.

If the perturbation is sinusoidal, the definition gives

$$\sigma_y^2(\tau) = a'^2 \frac{\sin^4(\pi f_M \tau)}{(\pi f_M \tau)^2} . \tag{5.31}$$

In the graph of $\sigma_y^2(\tau)$, its effect is manifested by an alternating variation about

a straight line of slope -1. This occurs now for values of τ of the order of $1/f_M$.

Therefore, characterising frequency instability by means of the Allan variance or deviation, we can identify the presence of certain types of systematic perturbation to the oscillator frequency.

(e) Modified Allan variance

One drawback of the Allan variance is that it varies in the same way, as τ^{-2}, when the noise affecting the oscillator is white phase noise ($\alpha = 2$) or flicker phase noise ($\alpha = 1$). In order to use it to distinguish these two types of noise, measurement series must be carried out for different values of the pass band f_h of the measurement system.

It has been shown that this disadvantage can be overcome by introducing a new variance, known as the modified Allan variance, denoted by $\mathrm{Mod}\,\sigma_y^2(\tau)$. The defining algorithm makes it possible to vary the pass band in a virtual sense [5.5].

With $\tau = n\tau_0$, we have

$$\mathrm{Mod}\,\sigma_y^2(\tau) = \frac{1}{2}E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{k=1}^{n}\bar{y}_{i+k+n}^{\tau_0} - \frac{1}{n}\sum_{k=1}^{n}\bar{y}_{i+k}^{\tau_0}\right)\right]^2\right\}, \qquad (5.32)$$

or

$$\mathrm{Mod}\,\sigma_y^2(\tau) = \frac{1}{2}\frac{1}{(n\tau_0)^2}E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(x_{i+2n} - 2x_{i+n} + x_i)\right]^2\right\}. \qquad (5.33)$$

For $n = 1$, we retrieve the expression for the Allan variance. Table 5.3 gives the functional dependence of $\mathrm{Mod}\,\sigma_y^2(\tau)$ on τ for the various types of noise.

It will be observed that this dependence differs for white phase noise ($\alpha = 2$) and flicker phase noise ($\alpha = 1$). These types of noise can thus be unambiguously identified and distinguished. For example, it is possible to measure the level of white phase noise associated with the finite resolution of the intervalometer used in the direct measurement of time instability (see Section 5.2.3a).

For the other types of noise ($\alpha \leq 1$), there is no special advantage in using the modified Allan variance. Note that for $\alpha \leq 0$, the numerical coefficients are not the same for the Allan variance and the modified Allan variance.

Table 5.3. *Correspondence between the various noise components perturbing an oscillator and the modified Allan variance of the normalised frequency fluctuations when* $2\pi f_h \tau \gg 1$. *The expression for* $\mathrm{Mod}\,\sigma_y^2(\tau)$ *is exact for white phase noise whatever the value of* $n = \tau/\tau_0$. *For the other noise types, the expressions given are valid to within a few percent for* $n \geq 10$.

$S_y(f)$	$\mathrm{Mod}\,\sigma_y^2(\tau)$
$h_2 f^2$	$3h_2 f_h \tau_0^2/4\pi^2\tau^3$
$h_1 f$	$h_1[1.04 + 3\ln(2\pi f_h \tau)]/4\pi^2\tau^2$
h_0	$h_0/4\tau$
$h_{-1} f^{-1}$	$0.936 h_{-1}$
$h_{-2} f^{-2}$	$5.42 h_{-2}\tau$

5.2.7 *Characterising time instability*

Time instability is generally characterised by appealing to the Allan variance or modified Allan variance. We shall give an example of the use of each of these variances.

(a) Uncertainty in time predictions

In the study of time scales and the operation of digital telecommunications networks, it is often necessary to predict a value of the time deviation between two clocks at some later date. The simplest approach consists in carrying out a linear extrapolation. We set

$$\hat{x}(t+\tau) = x(t) + \tau \bar{y}_k^{\tau}(t) \,, \tag{5.34}$$

where $x(t)$ and $\hat{x}(t+\tau)$ are the time deviation between the two clocks at the present time t and the predicted value at the later time $t+\tau$, respectively. The quantity $\bar{y}_k^{\tau}(t)$ is the value of the relative frequency offset between the two clocks, time-averaged over the period from $t-\tau$ to t. The prediction is impaired by a random error $\Delta x(t, \tau)$ equal to

$$\Delta x(t, \tau) = x(t+\tau) - \hat{x}(t+\tau) \,, \tag{5.35}$$

where $x(t+\tau)$ is time deviation which will actually be observed at the future time $t+\tau$. The quantity $\Delta x(t, \tau)$ is the *time prediction error.*

The variance of $\Delta x(t, \tau)$ characterises the uncertainty in the prediction [5.12]. From (5.8) and (5.13), we obtain

$$E\left\{\Delta x^2(t, \tau)\right\} = 2\tau^2\sigma_y^2(\tau) \ . \tag{5.36}$$

The standard deviation of $\Delta x(t, \tau)$ is thus independent of τ for white and flicker phase noise. It varies as $\tau^{1/2}$ for white frequency noise, it is proportional to τ for flicker frequency noise and it grows as $\tau^{3/2}$ for random walk frequency noise.

(b) Variance of time instability

In order to characterise the properties of time transfer and dissemination systems, as well as telecommunications systems, it was considered useful to introduce the variance of the time instability, denoted by TVAR. (Note that the Allan variance and modified Allan variance are sometimes denoted by the acronyms AVAR and MVAR, respectively.) This is defined by [5.13]

$$\text{TVAR} = \frac{\tau^2}{3}\text{Mod}\,\sigma_y^2(\tau) \ . \tag{5.37}$$

5.2.8 Determining contributions from each oscillator

The measurement of time and frequency instability involves two generators. We assume that the measurement system does not contribute to observed fluctuations.

If the instability of oscillator 1 is known to be much smaller than that of oscillator 2, all measured instability is attributed to the latter. If the two generators are built in the same way, then for want of a better hypothesis, we may suppose that they are of the same quality. If in addition there is no reason to suppose that their fluctuations are correlated, the variance of the measured fluctuations is equal to twice that of each individual oscillator.

It is possible to separate the individual contributions if more than two generators are available. Let us consider the case of three generators as an example. Let σ_i^2, σ_j^2, σ_k^2 be the intrinsic Allan variances of each for a given τ, and σ_{ij}^2, σ_{jk}^2, σ_{ki}^2 the Allan variances resulting from pairwise comparison for the same value of τ. Provided that the fluctuations are not correlated, we find

$$\begin{cases} \sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2 \ , \\ \sigma_{jk}^2 = \sigma_j^2 + \sigma_k^{2'}, \\ \sigma_{ki}^2 = \sigma_i^2 + \sigma_k^2 \ . \end{cases} \tag{5.38}$$

We may then deduce the individual variances:

$$\begin{cases} \sigma_i^2 = \dfrac{1}{2}\left(\sigma_{ij}^2 + \sigma_{ik}^2 - \sigma_{jk}^2\right) , \\ \sigma_j^2 = \dfrac{1}{2}\left(\sigma_{jk}^2 + \sigma_{ij}^2 - \sigma_{ik}^2\right) , \\ \sigma_k^2 = \dfrac{1}{2}\left(\sigma_{jk}^2 + \sigma_{ik}^2 - \sigma_{ij}^2\right) . \end{cases} \tag{5.39}$$

This method gives satisfactory results if the following conditions are fulfilled:

- the generators have comparable quality;
- their fluctuations are not correlated;
- measurements are made simultaneously;
- uncertainty in the estimation of variances σ_{ij}^2, σ_{jk}^2 and σ_{ki}^2 is sufficiently small.

This method is known familiarly as the *three-corner hat method.*

5.3 Mechanical oscillators

The highest achievement in mechanical oscillators came from pendulum clocks in astronomical observatories, which reached their ultimate level of perfection at the beginning of the twentieth century. We shall not describe these beautiful instruments. An interesting reference is [5.14]. Let us just recall some of the precautions taken to ensure the regularity of their operation. The clocks were closed in airtight tanks and maintained at a pressure slightly below atmospheric pressure. These were kept in rooms at constant temperature and protected as far as possible from vibrations. Their motion was produced by weights that were raised electrically, thus guaranteeing a constant couple and hence a constant oscillation amplitude of the pendulum. They often had no dial, sending out only electrical pulses each second. Despite all these precautions, their frequency instability over sampling periods of one month could reach anything between 5×10^{-8} and 10^{-7}.

At the Paris Observatory, a 'mean pendulum clock' was put together from a selection of the best pendulum clocks, located in the observatory cellars at the depth of the catacombs. This answered the needs of the Time Service and the Bureau international de l'heure. From 1950, quartz clocks were integrated into the mean pendulum clock (sic), the two types of clock being used in tandem until 1954, when the last pendulum clock left this association and the transformation was completed.

5.4 Piezoelectric oscillators

Quartz piezoelectric oscillators were the first time and frequency oscillators to benefit from advances made in electronic technologies. Their operation is based on the phenomenon of piezoelectricity, discovered by P. Curie in 1880, and the invention of the first electronic amplifier, the triode, by Lee and Forest in 1907. These oscillators have been used in communication systems and in time and frequency metrology since the 1920s. Today, high performance quartz oscillators are used as reference frequency generators and timekeepers in a great many radio navigation and telecommunications systems. In addition, they are always present in atomic clocks, where their frequency is controlled by the atomic resonance, and signals accessed by the user are derived from them.

In the context of increasingly demanding electronic systems and continued improvement in the performance of atomic clocks, much effort is being directed at improving resonator design and optimising the association with its electronic circuit. Much other research is aimed at understanding the origin of short-term frequency instabilities, sensitivity to environmental conditions and ageing effects [5.15, 5.16].

5.4.1 Resonators

The resonator is a vibrating plate. Various types of deformation (e.g., bending, extension, shear) can propagate through the bulk of the material. A resonance is produced when the acoustic waves satisfy the boundary conditions. Its frequency thus depends on the dimensions of the plate, the type of deformation that is excited and properties of the material used.

The mechanical resonance can be maintained by an electronic circuit via the piezoelectric effect. The latter couples electric fields and mechanical deformations in crystals with no centre of symmetry, such as quartz. The electric field is produced using electrodes deposited on the faces of the plate or located very close to them, as shown schematically in Figure 5.4.

Natural quartz crystals are not used. Instead, the crystals are commercially manufactured in large quantities and at relatively low cost in autoclaves where they are grown from seed. The material obtained is extremely pure chemically and its crystal structure exhibits very few flaws. Dielectric losses are also very low.

Figure 5.5a shows the equivalent electric circuit of a piezoelectric resonator. The quartz plate placed between the electrodes acts as a capacitor with capacitance C_0. The inductance, capacitance and resistance associated with the dynamical behaviour of the plate are represented by L_1, C_1 and R_1, respectively.

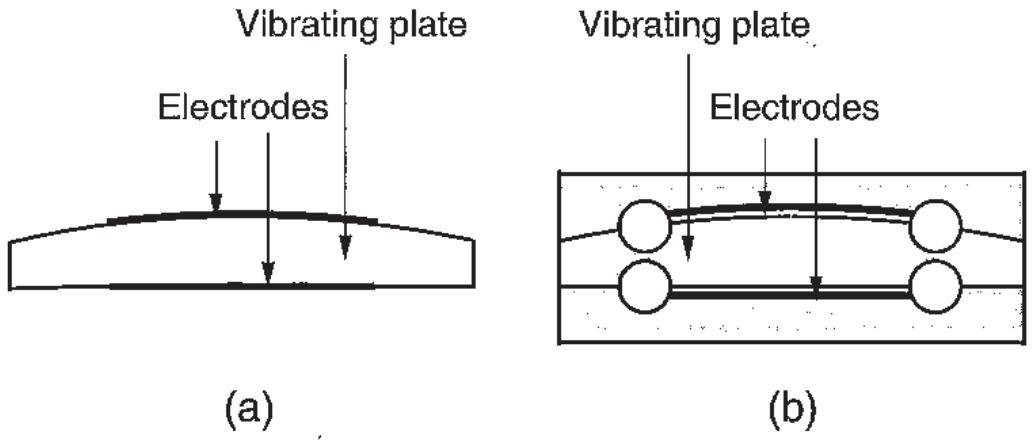


Fig. 5.4. Cross sections of quartz resonators. (a) Electrodes deposited on the vibrating plate. (b) Electrodes not fixed on the vibrating plate. Connections are not represented.

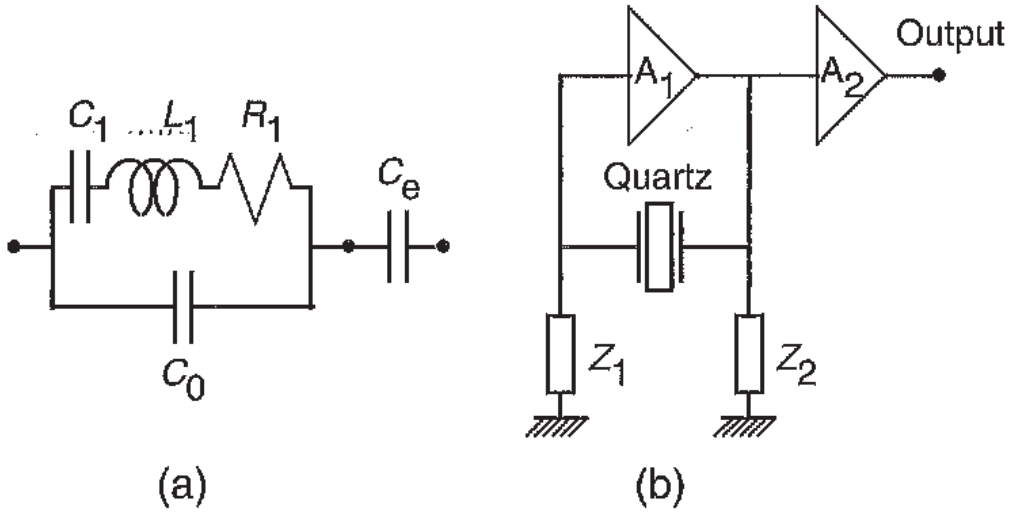


Fig. 5.5. Quartz oscillator and resonator. (a) Equivalent circuit diagram for a quartz resonator. (b) Schematic of a quartz oscillator.

An external capacitor with capacitance C_e placed in series with the resonator is used to adjust the resonant frequency.

The angular frequency ω_0 of the electrical resonance is given by

$$L_1 C_1 \omega_0^2 = 1 \,, \tag{5.40}$$

and the quality factor is

$$Q = \frac{L_1 \omega_0}{R_1} \,. \tag{5.41}$$

For precision resonators in metrological applications, $\omega_0/2\pi$ is generally equal to 5 or 10 MHz, whilst the quality factor is of the order of 10^6.

Because of the capacitor of capacitance C_0 in the circuit of Figure 5.5a, this circuit has two resonances, the series resonance and the parallel resonance,

with very similar frequency values. In the first case, the resonator exhibits low impedance, close to R_1. In the second, the impedance is high. One or other of these resonances is excited depending on whether the resonator is connected to low or high impedances Z_1 and Z_2, shown in Figure 5.5b.

The resonator is designed to minimise the effects of perturbations applied to the vibrating part of the quartz plate. The orientation of the plate relative to the crystallographic axes is thus chosen in such a way that the resonant frequency has an extremum with respect to temperature changes. The resonator is placed in a thermostat which keeps the temperature at that value for which the frequency variation vanishes. The choice of orientation of the plate is also guided by a desire to reduce the variation of the resonant frequency caused by stresses which may be due either to the mechanical mounting of the resonator or to accelerations suffered by the oscillator.

In the most up-to-date quartz resonators, the electrodes are not deposited on the faces of the vibrating plate [5.17]. They are located on auxiliary plates as shown in Figure 5.4b. These are placed a few micrometres from the active plate. Any damping of the resonance by the electrodes themselves is thereby suppressed. Moreover, the vibrating plate no longer suffers the ageing effects associated with slow relaxation of stresses and migration of materials that occur at the interface between the quartz and the metal.

5.4.2 Oscillators

The resonator is inserted in a feedback loop, between the input and output of an amplifier A_1, as shown schematically in Figure 5.5b. When the oscillator takes advantage of the series resonance, impedances Z_1 and Z_2 have low resistances.

The total phase displacement in the feedback loop is always equal to $2k\pi$, where k is an integer. Consequently, if a variation occurs in the phase displacement between the input and output voltages of the amplifier, this is compensated in the quartz resonator. Such a change can be produced by the noise that amplifier A_1 superimposes on the electrical oscillation. When there is a variation in the applied frequency, close to the resonant frequency f_0, a phase change $d\phi$ results at the terminals of any resonator, given by

$$d\phi = -2Q\frac{df}{f_0} . \tag{5.42}$$

The perturbation is thus accompanied by a change in the oscillation frequency. However, the corresponding change decreases as the quality factor increases. This explains why it is preferable to use resonators with a high quality factor when building oscillators with low frequency instability.

The additional capacitor of capacitance C_e is used to fine tune the oscillation frequency. It generally comprises two capacitors, one of fixed capacitance and the other a varicap, connected in series. The latter uses the change in capacitance of a diode with varying applied polarisation voltage. The oscillation frequency can then be controlled by an electrical voltage. This feature is used when the frequency of a quartz oscillator is servo controlled by the frequency of an atomic resonance (see Sections 6.4.1c and 6.5.3d).

The electronic circuit associated with the quartz oscillator is also thermostatically controlled. The amplifier A_2 increases the signal level and insulates the oscillation circuit from perturbations that might arise from application circuits.

5.4.3 Performance levels

For applications to high level metrology, the frequency of quartz oscillators is equal to 5 or 10 MHz, or else it is close to one of these values.

For sampling periods somewhere between about 1 and 10^3 s, the frequency instability of precision quartz oscillators is determined by flicker frequency noise. The Allan deviation is then less than 1×10^{-13} for the best devices. For longer periods, there is a frequency drift due to ageing effects. It is often less than 10^{-11} per day. A frequency instability of 1×10^{-14}, for τ in the range 1 to 10^3 s, together with a long term drift of 1×10^{-12}, would appear to be accessible in the near future.

5.5 Atomic time and frequency standards. The notion of accuracy

In Section 4.4.1, we explained how atomic time and frequency standards came into being, whilst in Chapter 6 we shall present their operating principles.

At the present stage, we shall just give a precise definition of the idea of accuracy already used in Section 4.4.1. In an atomic standard, the reference frequency ν_0, postulated to be naturally unchanging, is provided by the transition of an unperturbed atom or ion (or even a molecule, but we intend the word 'atom' to cover all these cases). It is a proper frequency (in the relativistic sense) in a very small spatial region, in principle, infinitesimal, which accompanies the atom in its motion. The set of apparatus used to observe this atomic frequency, the frequency standard, brings perturbations with it, as we shall explain in Chapter 6. It also transforms the frequency into other frequencies that are easier to handle. The output frequency from the standard is denoted ν_{out}. This is also a proper frequency in a small spatial region that surrounds, for example, the connector at which ν_{out} is available. In order to guarantee the

accuracy, the ratio

$$R = \frac{\nu_{\text{out}}}{\nu_0} \tag{5.43}$$

must be determined as closely as possible. The relative uncertainty σ_{E} that characterises the inaccuracy of the standard is equal to the relative uncertainty in the estimate of R.

Like any physically realised frequency, ν_{out} suffers from instability. A first precaution consists in taking its average value over a period ensuring minimal instability. We have seen that this period corresponds to the regime dominated by flicker frequency noise. From now on, we shall take ν_{out} to be this frequency average.

The art of the physicist is to identify all possible sources of perturbation. The next task is to evaluate each of them, often making use of auxiliary measurements, and to assess the uncertainty in this evaluation. Having done this for each perturbation i, a correction factor r_i is produced that can be written in the form

$$r_i = (1 + c_i) \pm \sigma_i \ , \tag{5.44}$$

where σ_i is the corresponding uncertainty. The estimate $\hat{R}$ of R is then

$$\hat{R} = F r_1 r_2 \ldots r_n \ , \tag{5.45}$$

for the whole set of n perturbations. F is the frequency transformation factor, assumed to be precisely known, as is generally the case. Since the c_i are much smaller than 1, we may adopt the following form

$$\hat{R} = F(1 + c_1 + c_2 + \cdots + c_n) \ . \tag{5.46}$$

It is usually assumed that the various perturbations are mutually independent, so that the inaccuracy in the normalised frequency is

$$\sigma_{\text{E}} = \left(\sum \sigma_i^2\right)^{1/2} \ . \tag{5.47}$$

For convenience when using the standard, the corrections c_i can be introduced physically at the level of the output frequency. This has no effect on the inaccuracy of the standard.

This notion of inaccuracy is not related specifically to atomic frequency standards, but rather it is a general notion that applies to any physical quantity defined with respect to some natural phenomenon that is assumed to be invariable.

5.6 Time and frequency comparisons

5.6.1 General observations

For all physical quantities, comparisons between standards constitute one of the cornerstones of metrology. Their main purpose is to check agreement between the primary standards in various laboratories, and then to contribute towards improving them by identifying the cause of errors that might otherwise have escaped notice. They may also be used to produce a mean primary standard, better than each of its participants. As far as the user is concerned, they provide a means of accessing primary standards, through the organisation of calibration chains.

In the area of atomic time measurement, the key quantity for those building standards is the frequency. Within the laboratory, it is easy to compare the various frequency standards on a continuous basis and thereby to check that they are functioning correctly. It is then possible to make occasional frequency comparisons between remote laboratories, as is done for other physical quantities.

However, the application of frequency standards to time measurement is considerably more demanding, as was emphasised in Section 4.4.2. Almost continual time comparisons are required to produce and disseminate a unique time scale with which users can synchronise their clocks and thereby procure the unit of time, the second. We shall also consider the example of systems which require an internal synchronisation as precise as possible that cannot be maintained without resetting at intervals of less than one day.

In the following sections, we begin by examining methods of time comparison, and then methods of frequency comparison. We shall be mainly concerned with the principles, since techniques evolve so rapidly and depend heavily on the opportunities that present themselves. This last claim deserves some explanation. Of course, certain systems are specially assigned to time dissemination, and to time and frequency comparisons, e.g., the time signals broadcast since the beginning of the twentieth century at frequencies ranging between a few kHz and several tens of MHz. Such emissions still exist and their carrier frequencies are often stabilised by atomic standards, in such a way that they also disseminate an excellent frequency reference that could easily be put to use with the help of a cheap receiver. But to reach better accuracy, it would have been much too costly to set up operational systems entirely devoted to clock comparisons. It was thus decided to take advantage of existing systems, created for other purposes, and this at no extra cost! Such systems include radio navigation, positioning and communication systems. We shall examine the way in

which methods of time and frequency comparison are most commonly applied using these systems.

In order to make remote comparisons between the best frequency standards and the best clocks, a relativistic treatment of the measurements is generally a prerequisite. There is no simple way of specifying the limits of accuracy involved in a classical treatment, because they depend on so many factors. We may nevertheless say that, for standards and clocks fixed at ground level, a classical treatment is adequate provided that we can accept uncertainties of a few times 10^{-13} in the normalised frequency and one microsecond in the time, for all the methods described below.

5.6.2 Relationship between frequency and proper time increment

The relationship between the readings of a clock and its frequency often involves mental gymnastics of a potentially hazardous nature as far as the results are concerned, even when handled by experts! If τ_C represents the proper time (readings) of a clock expressed in seconds, and if τ is the proper time of an ideal clock located in the same place, i.e., a clock whose second markers are separated by an interval of one second that is in strict conformity with the SI definition, then (5.6) implies

$$\frac{dx_C(\tau)}{d\tau} = y_C(\tau) \,. \tag{5.48}$$

As explained in Section 5.2.2,

$$\tau_C(\tau) = \tau + x_C(\tau) \,, \tag{5.49}$$

and hence, introducing the normalised frequency $\Phi = 1 + y$ defined in Section 4.4.2,

$$\Phi_C = \frac{d\tau_C}{d\tau} \,. \tag{5.50}$$

Hence, for two clocks C and D,

$$\frac{\Phi_C}{\Phi_D} = \frac{d\tau_C}{d\tau_D} \,. \tag{5.51}$$

5.6.3 Time comparisons

(a) Proper time of a clock and instrumental delays

A clock is a complex set of equipment and the question arises as to exactly where its proper time is located. In fact, it will be the time scale constructed using the pulses it produces each second, as they occur at a specified connector.

These pulses take the form of a rise and fall of voltage that is kept as brief as possible but must necessarily have some non-negligible *rise time*. The pulse height must therefore be specified.

A pulse arriving from elsewhere is dated by appealing to local measurements, as explained in (b) below. But before it arrives at the clock, this pulse may have undergone transformations that introduce an *instrumental delay*. This happens, for example, in radio receivers. Since it is always the arrival of the signal before transformation that is useful, dates provided by the clock must be corrected for instrumental delay.

The equations established in the following for time comparisons will not take instrumental delays into account. Their role will be mentioned in applications. We should emphasise that they are difficult to determine and that the global uncertainty in clock comparisons is generally dominated by uncertainties originating in this way.

(b) Local time comparisons

Let us recall from Section 5.2.3a that *time interval counters* measure, in their own time scale τ_{TIC}, the time interval $\Delta\tau_{\mathrm{TIC}}$ elapsed between a second pulse of a clock C and the first second pulse of a clock D to arrive after it. In order to interpret this measurement, one approach is to arrange for $\Phi_{\mathrm{TIC}} = \Phi_{\mathrm{C}}$. This can be achieved by driving the counter at the frequency of C. Another approach involves calculation based on frequency deviation measurements. We then have

$$\Delta\tau_{\mathrm{TIC}} = [\tau_{\mathrm{C}} - \tau_{\mathrm{D}}]\ (\theta_2) \quad \text{modulo 1 second}\ , \tag{5.52}$$

where θ_2 is the date in some reference time scale θ when the counter is stopped by the second pulse from D.† The whole number of seconds is then added. Modern instruments can carry out these measurements with minimal uncertainties of 10 ps. For the present application, we must of course measure the delays due to the cables.

(c) Remote time comparisons

When the clocks are too far apart to be connected by cables, time comparisons are carried out using one of the following three modes, or a combination of them:

- physical transportation of an operating clock;

† Recall that the expression $[\tau_{\mathrm{C}} - \tau_{\mathrm{D}}]\ (\theta)$ is understood to mean the reading from C minus the reading from D at the date θ.

- one-way transmission of electromagnetic signals between an emitter and a receiver;
- quasi-simultaneous transmissions of electromagnetic signals along almost identical trajectories joining clocks A and B, from A to B and from B to A (the so-called two-way time transfer method).

The equations, in relativistic form, appeal to the conventional definition of time comparisons explained in Section 3.3.2f. We observe that they are always made up of a 'classical part', the same as would be obtained in the absolute time framework, together with small additional terms. This separation is due to the form of the metric and depends on the degree of approximation accepted. It is convenient in the sense that the so-called relativistic terms can simply be left out when tolerance levels render them negligible. However, there is one serious drawback, for a separation of this kind may lead us to believe that general relativity is somehow just an add-on to the classical theory, whereas it in fact represents a globally new treatment. It is worthless and ill-advised to distinguish a priori a classical effect, a special relativistic effect and a gravitational effect. This may lead to error and, in particular, the unwarranted inclusion of 'acceleration effects'.

A relativistic treatment always requires the increment in the coordinate time (in some specified coordinate system) to be evaluated, either during transportation of the clock, or during transmission of the signal. The geocentric coordinate systems, either non-rotating or rotating with the Earth, are used in this context. The following discussion, based upon [5.18], is applicable when time uncertainties of a few hundred picoseconds and normalised frequency uncertainties of 10^{-16} are tolerated, for ground-based or orbiting clocks, out to the distance of the geostationary satellites (42 000 km from the centre of the Earth), apart from a few exceptions that will be pointed out. A more precise analysis can be found in [5.19, 5.20]. We may use either the geocentric coordinate time t, or its modified form $\underline{t}$ defined in Section 3.3.2e. Formulas will be given in terms of t. They remain valid when t is replaced by $\underline{t}$, and $h(t)$ is replaced by $\underline{h}(\underline{t})$, functions analysed in (3.18), (3.21), (3.30) and (3.32). When the measurement mode requires it, expressions in terms of $\underline{t}$ will be used.

According to (3.17) and (5.51), the increment Δt in t between the two events constituted by the readings $\tau_C(t_1)$ and $\tau_C(t_2)$ of a clock C is given by

$$\Delta t = \frac{\tau_C(t_2) - \tau_C(t_1)}{1 + y_C} + \int_{W(C)} h(t)\, dt \ , \tag{5.53}$$

where y_C is the relative frequency offset with respect to an ideal frequency standard and the integral is calculated along the worldline W(C) of the clock from t_1 to t_2.

Equation (5.53) may seem somewhat surprising since we should have $\Delta t = t_2 - t_1$. In order to understand its use, we must assess the role of t in the following way. Uncertainty in Δt directly contributes to uncertainty in the time comparison. In readings such as $\tau_C(t_2)$, it is the precision in τ_C which is at stake, whilst t_2 serves only to label the event 'second reading of C', which we could just as well have described in these terms. In the integral, we certainly need t and the limits t_1, t_2, but a physical realisation of t is sufficient, provided that its frequency offset from t is not too large. For example, if this offset is 10^{-9}, since $h(t)$ is of the order of 10^{-9}, the error introduced into Δt will be less than $10^{-18}\Delta t$, which is quite negligible for today's measurements. If we assume that the reference scale θ, already used previously, satisfies the latter condition, t can be replaced by θ everywhere on the right hand side of (5.53). The scale might be Coordinated Universal Time (UTC), for example, easily available to within a few tenths of a second (taking possible leap seconds into account) or to within 10^{-9} as regards normalised frequency. However, in the following we shall generally stick to expressions in terms of t, leaving it to the user to work out the degree of approximation to t that would be acceptable, never a difficult task.

(1) Time comparisons by clock transportation. Let A and B be the two clocks under comparison. The method here involves carrying a clock M from the vicinity of A to the vicinity of B, where local time comparisons $[\tau_A - \tau_M]\ (t_1)$ and $[\tau_B - \tau_M]\ (t_2)$ are made. Interpretation is clear in the absolute time model. M serves to extrapolate the readings from A. This application requires a preparatory stage in which the difference in normalised frequencies $y_A - y_M$ is evaluated. It also requires a closure check: M is brought near A and it is confirmed that the observed time and frequency difference with A is compatible with what is expected, within the limits of estimated uncertainty. If this is indeed the case, the observed *closing error* must nevertheless be taken into account. This involves the common sense of the experimenter rather than a statistical treatment.

The relativistic interpretation was one of the first applications of general relativity to commonly made measurements. It follows from (5.53), with $\Delta t = t_2 - t_1$, that

$$\tau_M(t_2) - \tau_M(t_1) = (1 + y_M)\left[\Delta t - \int_{W(M)} h(t)\,dt\right]. \qquad (5.54)$$

The same equation is written to obtain $\tau_A(t_2) - \tau_A(t_1)$. Then subtracting one

from the other, and bringing in the measured quantities,

$$[\tau_B - \tau_A]\,(t_2) = [\tau_B - \tau_M]\,(t_2) - [\tau_A - \tau_M]\,(t_1) - (y_A - y_M)\,\Delta t$$
$$-(1 + y_M)\int_{W(M)} h(t)\,dt + (1 + y_A)\int_{W(A)} h(t)\,dt\ . \tag{5.55}$$

The first three terms on the right hand side represent the classical expression. Relativistic effects are represented by the remaining terms, where the integrals are calculated along worldlines W(M) and W(A) of clocks M and A, between dates t_1 and t_2. With negligible inaccuracy compared with measurement uncertainties, we may usually take $y_M = 0 = y_A$ in the factors multiplying these integrals.

First undertaken in 1967, at the initiative of Hewlett-Packard and the Smithsonian Astrophysical Observatory, time comparisons by clock transportation were common practice until around 1988 for the purpose of synchronising the control centres of navigation systems such as Loran-C, tracking stations for artificial satellites, and national timekeeping laboratories. It was by far the best way of guaranteeing the accuracy of time scale comparisons since all other methods involved propagation and instrumental delays which could not be directly estimated without introducing an excessive dose of uncertainty.

Normally, commercially manufactured caesium clocks were carried by road or plane. Their size and weight (around 40 kg) meant that they could be placed on a passenger seat. When their own electricity supply was not sufficient, they were allowed to use on-board resources, with the agreement of the airline. In 1978, for instance, the Annual Report of the Bureau international de l'heure lists forty-three time transfers of this nature, ranging across all the continents, with uncertainties of 10 to 300 ns. Clock transportation became rarer from 1986, with more and more time comparisons being carried out via the Global Positioning System. Moreover, at around this time, air traffic safety regulations began to cause problems, for it was realised that caesium is liable to spontaneous combustion when exposed to the air. Attempts to carry rubidium clocks were soon abandoned for their lack of stability in such conditions.

It is interesting to note that this time transfer method is the source of a great many anecdotes, and some misadventures. We might mention the airline tickets made out to Mr Clock and the general surprise of the flight personnel, the complications caused by appearance of the word 'atomic' and its rapid replacement by 'electronic', or the endless confusion and misunderstanding of customs officials.

In order to account for relativistic effects, the flight path of the airplane had to be known. Such effects may reach a few tens of nanoseconds over intercon-

tinental transfers. For example, in a Paris–Ottawa transfer, it transpired that 32 ns had to be subtracted off the results of a classical calculation, whilst for the return flight, assumed to follow the same route, only 7 ns had to be taken off. This could have been used to show that the Earth is indeed rotating.

(2) Time comparison by one-way transfer of electromagnetic signals

Principles and theory. This method, used since the beginning of the twentieth century (see Chapter 4) is based upon the theory of signal propagation. It is often chosen when transmitting times to a low level of accuracy (time signals for navigational purposes, speaking clocks, and so on). It can also satisfy stringent accuracy requirements, such as those we are concerned with here, operating by direct view radio transmission. For large distances, this means using a space-based emitter.

We begin with the theory behind the basic operation, which is comparison of two clocks by direct view, and then we shall consider some applications, in particular those that make use of the Global Positioning System (GPS).

In the vacuum, electromagnetic signals propagate in straight lines relative to inertial systems (i.e., systems in uniform translational motion, without rotation). The non-rotating geocentric system defined in Chapter 3 constitutes such a system to a quite sufficient approximation for time comparisons. If photons travel from the event E of signal emission, with coordinates $(t_{\mathrm{E}}, \mathbf{x}_{\mathrm{E}})$, to the event R of signal reception, with coordinates $(t_{\mathrm{R}}, \mathbf{x}_{\mathrm{R}})$, the increment Δt in t during propagation is, to order c^{-3} which suffices here,

$$\Delta t = \frac{d}{c} + (\Delta t)_U \ , \tag{5.56}$$

where d is the coordinate distance between E and R, calculated from

$$d = \left[(x^1_{\mathrm{R}} - x^1_{\mathrm{E}})^2 + (x^2_{\mathrm{R}} - x^2_{\mathrm{E}})^2 + (x^3_{\mathrm{R}} - x^3_{\mathrm{E}})^2 \right]^{1/2} \ , \tag{5.57}$$

which we shall abbreviate to $|\mathbf{x}_{\mathrm{R}} - \mathbf{x}_{\mathrm{E}}|$, even though $\mathbf{x}_{\mathrm{R}} - \mathbf{x}_{\mathrm{E}}$ does not have the properties of a vector in the spacetime. The term $(\Delta t)_U$ is found by integration along the propagation path, taking $U = GM_{\mathrm{E}}/r$, with $r = |\mathbf{x}|$, bearing in mind that $GM_{\mathrm{E}} = 3.986 \times 10^{14}\ \mathrm{m^3\,s^{-2}}$. We obtain

$$(\Delta t)_U = \frac{2GM_{\mathrm{E}}}{c^3} \ln \frac{r_{\mathrm{E}} + r_{\mathrm{R}} + d}{r_{\mathrm{E}} + r_{\mathrm{R}} - d} \ . \tag{5.58}$$

This term is small. For transmission from a geostationary satellite down to a ground-based clock on the same vertical, it has a value of 56 ps.

Using the geocentric system rotating with the Earth and considering a receiving station fixed on Earth, with coordinates $\mathbf{X}_R$, Δt is obtained from

$$\Delta t = \frac{|\mathbf{X}_R - \mathbf{X}_E|}{c} + \frac{2\omega A_E}{c^2} + (\Delta t)_U \ , \tag{5.59}$$

where $\mathbf{X}_E$ is evaluated at date t_E and $(\Delta t)_U$ is still given by (5.58). The term $2\omega A_E/c^2$ describes the Sagnac effect. The area A_E, defined just after (3.21), refers here to the photon trajectories. To a good enough approximation, it is the area of the equatorial projection of the triangle with vertices the centre of mass of the Earth and the points with coordinates $\mathbf{X}_E$ and $\mathbf{X}_R$, taken positively when propagation occurs towards the east. The Sagnac effect is often significant. For example, in transmission from a point on the equator up to a geostationary satellite at an elevation of 45° towards the east, its value is 137 ns. The delay caused by refraction, which is evaluated separately, must then be added to Δt.

Suppose that, at the signal emission, we note the reading $\tau_S(t_E)$ of a local clock S. This reading can be transmitted to an experimenter by the usual means. (Alternatively, we may arrange for the emission to be attached to the readings of S so that the nominal date of the signal is known.) We follow the same procedure with a local clock A at the signal reception, taking a reading $\tau_A(t_{RA})$. Applying (5.53), we obtain the time difference between the clocks S and A at the date t_E from the relation

$$\begin{aligned}[\tau_A - \tau_S]\ (t_E) = \tau_A(t_{RA}) - \tau_S(t_E) \\ -(1 + y_A)(\Delta t)_A + (1 + y_A) \int_{W(A)} h(t)\, dt \ ,\end{aligned} \tag{5.60}$$

where $(\Delta t)_A$ is given by (5.56) or (5.59), and the integral is calculated along the worldline W(A) of A between the dates t_E and t_{RA}. As usual, values of t are replaced by readings from a realised time scale.†

When two clocks A and B are to be compared but they are not in direct view, we can use the reception of the same signal from an emitter that is in

† Exceptionally, let us check that the change $t \to \underline{t}$ does not alter the value of the left hand side. From (3.27),

$$\Delta \underline{t} = \Delta t \left(1 + \frac{\hat{U}_g}{c^2}\right) , \tag{5.61}$$

and from (3.31),

$$h(t) = \underline{h}(t) + \frac{\hat{U}_g}{c^2} \ ; \tag{5.62}$$

The term $\Delta \underline{t}\, \hat{U}_g/c^2$ coming from the integral does indeed compensate for the change in the evaluation of t.

direct view of each of the clocks. Writing the analogous equation to (5.60) for clock B and then subtracting (5.60), we obtain

$$[\tau_A - \tau_B]\,(t_E) = \tau_A(t_{RA}) - \tau_B(t_{RB}) - (1 + y_A)(\Delta t)_A + (1 + y_B)(\Delta t)_B + (1 + y_A)\int_{W(A)} h(t)\,dt - (1 + y_B)\int_{W(B)} h(t)\,dt\ . \tag{5.63}$$

Amongst the four last terms on the right hand side of (5.63), it is often only necessary to retain $(\Delta t)_B - (\Delta t)_A$. For example, for an emitter carried by a geostationary satellite, Δt is less than 0.14 s and if we have $|y| < 10^{-8}$, then $|y\Delta t| < 1.4$ ns. Since $h(t) < 7 \times 10^{-10}$, the integrals never exceed 100 ps, and y can be neglected in the multiplicative factors. It is nevertheless advisable to check each case separately to make sure that significant components have not been omitted.

Note that the fact of receiving the same signal means that the clock at S can be dispensed with. This signal is quite arbitrary. However, the presence of a stable clock at S avoids the constraint of a strictly common view. In this case, time comparisons between A and S, and B and S, made at different dates must then be referred to a common date via frequency estimates. Note also that we only need to know the differences in delay due to refraction at the receivers.

Application to Loran-C. Time comparison by one-way transmission was widely used from 1967, receiving signals from the Loran-C system. During the 1980s, this system was gradually abandoned for the purposes of fundamental metrology, in favour of GPS. Positioning by Loran-C is based on the synchronism of (caesium) clocks in ground stations emitting at 100 kHz. The positioning candidate measures the time difference caused by signal propagation delays. At sea level, each pair of stations supplies a locus for its position. The fix is obtained by receiving emissions from three stations. Of course, the navigation receivers carry out all the calculations. The precision of the system is due to the stability of propagation delays for *ground waves*, which can be observed at distances of 1000 to 2000 km. For a fixed observer, the variations in these delays may be around 1 μs. The actual delay itself cannot be evaluated to the same level of accuracy. In time comparison applications, it was measured by clock transportation, and the measurement included receiver delays. Specialised receivers for time comparisons were manufactured commercially. With great reliability and low cost, they contributed enormously to the task. It is estimated that overall uncertainties in Loran-C time comparisons were in the neighbourhood of 1 μs. Synchronisation of the various chains of Loran-C control stations across the North Atlantic at the end of the 1960s made it possible

to set up permanent time transfer links between North America and Europe. Indeed this played a major role in the development of time measurement.

Application to commercial television. The reception of commercial television signals contained in test lines or used to synchronise images was also widely used in conjunction with Loran-C [5.21]. At close range, within sight of the same emitter, uncertainties in time comparisons could be reduced to around ten nanoseconds, after calibrating transmission and instrumental delays by clock transportation. Links were established over several hundred kilometres using relays and radio beams. Accumulated uncertainties reached a few hundred nanoseconds, whilst many difficulties arose from the varied equipment and the problem of routing emissions.

Application to GPS and GLONASS. We now turn to the USA's Global Positioning System (GPS) which is at present widely used in remote clock comparisons. There is a Russian counterpart named *Global Navigation Satellite System* (GLONASS), which produces results of comparable accuracy and is also used. As the two systems only differ through minor technical features, we shall simplify by restricting explanations to GPS.

Like Loran-C, these systems are based upon the synchronism of emitted signals, the main difference being that the emitters are now space-based. In Chapter 9, we explain how they work. For our present purposes, it suffices to view the GPS satellites as emitters of time pulses. Apart from these pulses, the satellites broadcast data allowing their position to be determined in a terrestrial coordinate system at the time of emission of the pulses. Signals are emitted at frequencies close to 1.6 GHz and 1.2 GHz. Ionospheric refraction leads to delays that can vary greatly between day and night, or in periods of solar activity. The delay may be as much as a hundred nanoseconds in normal operating conditions. As it is a simple function of the frequency, it can be evaluated to the nearest nanosecond from the time difference at reception at each of the frequencies. If only one frequency is received, it is also possible to use a model of the ionospheric refraction, transmitted by the satellites, accurate to the nearest ten nanoseconds.

The satellites carry either caesium or rubidium atomic clocks, synchronised to GPS Time to within a few nanoseconds. Although it is not essential for the operation of the GPS system, GPS Time is maintained in synchronisation with $TAI + 19$ s to within a hundred nanoseconds or so. Receivers specialising in time measurement exist commercially. Those from the first generation run on a single frequency. They are programmable and successively track the assigned satellites for periods of 13 minutes at a time (we shall see why in Chapter 9).

Then, toward 1999, multichannel receivers came into use, tracking all visible satellites simultaneously. All these receivers carry out calculations in real time and supply the difference in mean readings between the local clock and GPS Time.

Let us consider the method in which the same pulse is received at the two clocks to be compared, a method often called the *common view method*. After exchange of data, time comparisons are obtained simply by subtraction. Uncertainties can be reduced to a few nanoseconds at any distance, by averaging over one day. This presupposes that the stability of the clocks justifies such an average. In addition, in order to reach this accuracy, precautions are necessary which considerably complicate the task.

The coordinates of the receiving antennas must be known to within ten centimetres or so, in the same geodetic coordinate system as the satellite coordinates (in one nanosecond, the signal travels 30 cm). The frame of reference for GPS is the *World Geodetic System*, version WGS 84, which agrees to within a few decimetres with the ITRF frame maintained by the International Earth Rotation Service (see Section 3.3.2b). The latter is adopted for very accurate work.

Delays introduced by the receivers are measured by the manufacturer using a local emitter that emits signals similar to those from the satellites. These delays are assumed to be invariable, but it is safer to check. For time comparisons, it suffices to calibrate the difference in delays between receivers. This can be done by transporting a receiver used as a transfer standard. The method is applied, but without much conviction, for the uncertainties are dominated by the sensitivity of receivers to environmental conditions. These uncertainties are of the order of a few nanoseconds.

Corrections due to the ionosphere must be measured, either by means of dual-frequency timing receivers, or else using specially designed equipment.

One difficulty, which fortunately disappeared in May 2000, was the use of the so-called *Selective Availability* (SA, see Section 9.2.2b). This deprived unauthorised users of the full accuracy of GPS. For time comparisons, the degradation caused by SA could be overcome by a rigorous observation schedule [5.22] and the use of precise ephemerides computed in retrospect. These constraints on schedules can now be relaxed, but precise ephemerides are still needed to obtain the utmost precision. They are established by a scientific service, the International GPS Service (IGS), mainly to supply the needs of geodesy and to monitor the Earth's rotation.

The development of multichannel receivers improves the precision of time comparisons by increasing the amount of data. However, the limiting factors in accuracy are due to weak points in the receiving systems, such as temperature

sensitivity (observed variations of several nanoseconds), and perturbation by reflections in the vicinity of antennas. Attempts are being made to overcome these and gain in accuracy by following the phase of the carrier, as is done in geodetic applications of GPS, rather than simply receiving the coded signal.

The Russian positioning system GLONASS operates according to the same principles as GPS. The fact that it is not so widely used for time comparisons can be explained by the later appearance of suitable receivers.

The success of GPS suggests setting up a civilian system along the same lines (Chapter 9). Time comparisons and the dissemination of time should benefit.

Application to quasars. Another type of emitter is also subject to the common view method, although this time much more remote. These are quasars. Very long baseline interferometry (VLBI) is a truly outstanding application of atomic clocks (see Section 8.1.4a and Chapter 9). In its use for geodesy and studies of the motions of the Earth, it involves measuring the differences in arrival times of random signals from quasars at sites lying several hundred to several thousand kilometres apart. Clocks at these sites cannot be synchronised to the desired accuracy (uncertainties of around 10 ps). Their offset is one of the many parameters that are globally fitted when processing measurements. But this fitting does not distinguish the clock offset from certain instrumental delays and errors in the estimation of atmospheric refraction. Delay calibrations using zero baseline interferometry could lead to time comparisons with uncertainties as low as a hundred picoseconds. Unfortunately, VLBI stations are often located a long way from timing laboratories.

(3) Time comparisons by two-way transfer of electromagnetic signals. Consider two clocks A and B. One-way transfers from A to B and from B to A are carried out almost simultaneously, with propagation along only slightly differing paths. We define

t_{EA} emission time of signal from A to B,
t_{RB} reception time of this signal at B,
t_{EB} emission time of signal from B to A,
t_{RA} reception time of this signal at A,

and

$$t_{av} = \frac{1}{2}(t_{EA} + t_{EB}).$$

Applying (5.60) and assuming that we can justify linearly interpolating $\tau_B - \tau_A$

between t_{EA} and t_{EB}, we obtain by averaging

$$[\tau_B - \tau_A]\ (t_{av}) = \frac{1}{2}\,[\tau_B(t_{EB}) + \tau_B(t_{RB})] - \frac{1}{2}\,[\tau_A(t_{EA}) + \tau_A(t_{RA})] - \frac{1}{2}\,[(\Delta t)_{AB} - (\Delta t)_{BA}] + \frac{1}{2}\varepsilon\ , \tag{5.64}$$

where

$$\varepsilon = y_A(\Delta t)_{BA} - y_B(\Delta t)_{AB} + (1 + y_B)\int_{W(B)} h(t)\,dt - (1 + y_A)\int_{W(A)} h(t)\,dt\ , \tag{5.65}$$

the integrals being taken over the worldlines of the clocks during propagation. The term ε is normally very small.

The coordinate time increments $(\Delta t)_{AB}$ for the path from A to B, and $(\Delta t)_{BA}$ for the path from B to A, are calculated using (5.56) or (5.59). Only their difference occurs in the formula. Consequently, when we are able to evaluate this difference, the method dispenses with the problem of obtaining precise knowledge of the clock coordinates, propagation paths and atmospheric delays.

The method, designated by TWSTFT, is usually applied to compare ground-based clocks using signals relayed by a geostationary satellite. It is then convenient to use the coordinate system rotating with the Earth. It is easy enough to account for a residual motion of the satellite relative to the Earth [5.19]. In (5.64), the contribution from terms $(\Delta t)_U$ as given by (5.58) is much less than 1 ps. The terms in (5.65) containing integrals tend to balance out so that their contribution is less than 1 ps provided that $|y| < 10^{-2}$. It may be necessary to evaluate the terms $y\Delta t$. The Sagnac effect is important because it changes sign with the direction of propagation. For example, if A is in Paris and B in Tokyo, and if the geostationary satellite is at longitude 67.5° east, then $(\Delta t)_{AB} = -(\Delta t)_{BA} = 299$ ns.

When applying this method, we must ascertain the sum of instrumental delays at emission and reception at each of the stations. In one station, it is straightforward enough to measure the difference in these delays by local experiments. If it remains constant, we may assume that each of the two delays is constant. However, measurement of these delays themselves has proved to be a delicate matter. As a last resort, the two-way time transfer can be globally calibrated via another type of time comparison. Results may also depend on variations in delays due to the satellite transponder. Still further difficulties occur if time links involve several antennas on the satellite.

Administrative obstacles are sometimes encountered when seeking permission to emit. A licence fee may be charged by the organisations managing the satellites.

These problems, together with the great convenience of the GPS system, have unfortunately slowed down development of the two-way time transfer method, despite its promise of improved accuracy. Nevertheless, following the first trial of this method in 1962, mentioned in Section 4.4.2 [4.22], many experiments followed. We may cite the first operational time transfer link between France, Canada and then Germany, from 1978 to 1982, thanks to the experimental telecommunications satellite Symphonie [5.23]. The overall uncertainty in these transfers, around a few hundred nanoseconds, was less than that incurred when using Loran-C, so that the Bureau international de l'heure used them to set up the TAI.

For the last few years, a working group at the International Bureau of Weights and Measures (BIPM) has been responsible for defining a system of timing links operating by the two-way transfer method. This involves choosing the satellites and frequency bands, and defining the particularities of the ground monitoring stations, the calibration method, the measurement schedule, and the exchange and processing of data. Several links are operated on a regular basis, some of them used by the BIPM for computing TAI. The instability of time comparisons with one day averaging is about 100 ps, but the inaccuracy, depending on delay calibration, lies in the range 1 to 10 ns. As it is, TWSTFT has no equal for long distance frequency comparisons (uncertainty 10^{-15} over one day) and is an excellent tool for the study of primary frequency standards. For time comparisons, competition with GPS is still open.

There are other applications of the two-way transfer method that do not involve geostationary satellites. At close range, the method can be implemented using laser signals transferred through the air or in optical fibres. At medium range, Russian scientists have used signal reflections from meteorite trails, at a frequency of 57 MHz. The range can reach about 1000 km and uncertainties are around 20 to 30 ns after calibrating instruments by clock transportation [5.24].

(4) A combined method for time comparisons: the LASSO experiment. The LASSO experiment (Laser Synchronization from Stationary Orbit) is an example of a mixed method aimed at improving accuracy. It uses laser ranging facilities which normally fire at artificial satellites and the Moon for studies in global geodynamics, oceanography, astronomy and so on.

A satellite S, geostationary in the experiment as it was carried out, carried a clock, also denoted by S. It was equipped with corner-cube retroflectors and a detector that could date the arrival of laser pulses at S.

The link between a clock A in some ranging station and the clock S may be compared with a classic two-way transfer link. Without taking into account the various instrumental delays, it is formulated via (5.64) and (5.65) with B replaced by S and the emission and reception times at S set equal. The link between a clock B and S is brought about in the same way. Time comparison between A and B involves 'transportation' of the clock S between the arrival dates of signals emitted by A and B. By good organisation of the laser shots, the duration of this transportation can be made as small as desired, so that the key quality required of S is short term frequency stability. A quartz clock is suitable. The arrival dates of the signals are transmitted to a coordinating facility.

This experiment was a project of the European Space Agency. It was originally set up on the Sirio 2 satellite but was lost due to launch failure in 1982. It was taken up again in 1988 with the Meteosat P2 satellite. In the meantime, ruby lasers had been replaced by neodymium–YAG lasers at the ranging facilities. Although these lasers have the advantage of shorter pulses, they transmit less energy and many of the stations lost all chance of receiving an echo from a geostationary satellite equipped with the reflectors designed for LASSO. In addition, as the satellite was stabilised by rotation (period 0.6 s), beams had to be fired in narrow windows (33 ms), which greatly complicated measurement arrangements.

As in all time comparisons, determination of instrumental delays is both essential and difficult. Delays at S cancel out. To find those at the ground stations, a portable laser ranging device was built. This calibrating instrument was positioned next to the one to be calibrated and fired towards the same local target (at a distance of about one kilometer). Each instrument was capable of receiving on return the emission of the other in addition to its own emission. By intersecting the shots in all possible ways, enough equations were produced to be able to determine all the delays involved.

The links established by LASSO between the lunar laser ranging stations in Grasse (France) and McDonald (Texas) in 1992–93 demonstrated that time transfer uncertainties were below 100 ps before calibrations were taken into account. The latter introduce a further uncertainty of 1.5 ns which predominates [5.25].

This method is complex and heavily dependent on cloud conditions. However, technical improvements since its original design offer some prospects. It could serve as a calibration method for other techniques, for research or

for downlinking data from high quality atomic clocks in orbit. The equipment needed is light and compact and can easily be installed either on geostationary or low Earth orbiting satellites already provided with clocks and remote-sensing devices.

5.6.4 *Frequency comparisons*

Let us just recall that all frequency measurements are average measurements, even if the word 'average' is often omitted. The frequency that concerns experimenters is the *proper frequency* referred to the local second. Quantities Φ and y should be understood in this sense.

When two remote frequency standards are compared, it is intuitively clear that the quantity to be determined is the difference between their normalised frequencies under the assumption that these standards have been brought side by side. Unfortunately, this intuitive idea does not give rise to a precise definition. This is because, strictly speaking, the result may depend on the route followed by the standards. The same could be said for any physical quantity. By convention, the frequencies are compared over the same coordinate time interval in a specified coordinate system, in this case the geocentric system.

(a) Frequency comparisons based on time comparisons

(1) Direct comparison of standards. Consider two clocks A and B whose normalised proper frequencies are to be compared. Using one of the methods described in Section 5.6.3, we measure the proper time differences $[\tau_B - \tau_A]\ (t_1)$ and $[\tau_B - \tau_A]\ (t_2)$ at dates t_1 and t_2. In principle the measurement dates are expressed in coordinate time t, but a rough approximation θ is sufficient here. We calculate

$$M_B - M_A = \frac{[\tau_B - \tau_A]\ (t_2) - [\tau_B - \tau_A]\ (t_1)}{t_2 - t_1} . \tag{5.66}$$

We then obtain the normalised proper frequency difference, averaged over (t_1, t_2), from

$$\langle y_B - y_A \rangle = M_B - M_A + \langle h(t, \mathbf{x}_B) \rangle - \langle h(t, \mathbf{x}_A) \rangle , \tag{5.67}$$

where the symbol $\langle\ \rangle$ denotes the time average. The function h is given by (3.18) or (3.21). It depends on t alone (and this can be replaced by θ for all practical purposes) along the worldlines $\mathbf{x}(t)$ of the clocks. Clearly, in a local experiment, the two terms in h cancel out.

Applied to ground-based clocks, the equations give

$$\langle y_B - y_A \rangle = M_B - M_A - 1.091 \times 10^{-13}(h_B - h_A)/\text{km} , \tag{5.68}$$

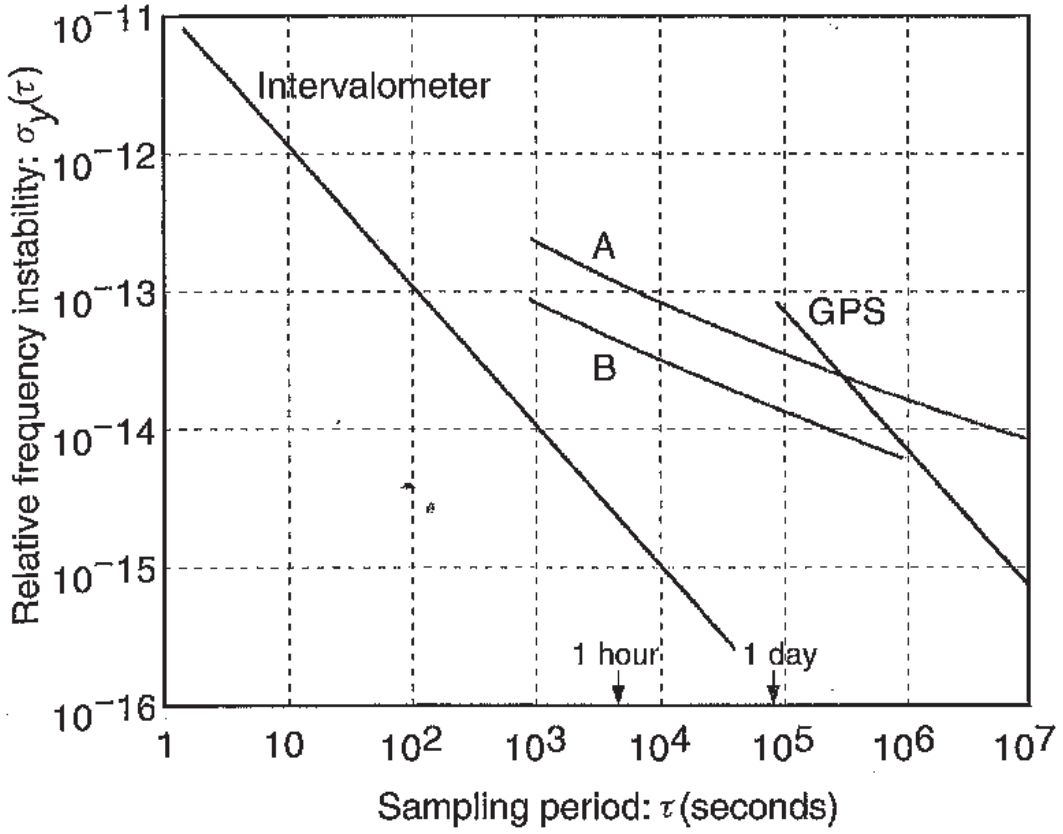


Fig. 5.6. Frequency instability introduced by time comparisons: by GPS (±5 ns) and by intervalometers (±10 ps). As a comparison, the mutual instability of the primary standards CS1 and CS2 at the PTB (Germany) is shown in curve A and that of CS2 and a hydrogen maser in curve B.

where h_A and h_B are the altitudes in kilometres above the geoid (sea level). The inaccuracy in (5.68) is less than 10^{-15} in this case.

When frequency comparisons are deduced from time comparisons, they are free of instrumental delays, provided the latter remain constant, and this can sometimes be checked.

Although uncertainties from delay calibrations may be omitted, noise remains in the time comparisons. This noise can be considered as white noise characterised by the standard deviation σ_C. The corresponding uncertainty in the frequency difference deduced from two time comparisons at dates t_1 and t_2 is

$$\sigma_y = \frac{\sqrt{2}\sigma_C}{t_2 - t_1} . \tag{5.69}$$

In the usual graph showing the frequency instability of oscillators, the instability introduced by time comparisons between two ideally stable oscillators takes the form of a straight line with slope −1 (see Figure 5.6). We can thus see immediately the minimal measurement interval to take so that the oscillators do not appear to be degraded.

The accuracy of the frequency comparisons is increased by extending the interval (t_1, t_2). But note that an uncertainty of 1 ns in time comparisons carried out at a one day interval leads to an uncertainty of 1.6×10^{-14} in the normalised frequency. Now at the present time (2000), it is difficult to compare remote clocks to better than 100 ps. For the best primary standards, this requires average frequency comparisons lasting for several days and rules out studies of shorter term frequency variations. Yet there is no way of doing better. Here is an example of the importance of time comparisons and the limitations imposed by uncertainties involved in them.

These methods are thus well suited to studying accuracy and long term stability. The two-way transfer method, with its excellent time resolution, produces the best results. Reception of GPS signals is a less precise method, but reliable.

(2) Comparisons via International Atomic Time. In laboratories equipped with GPS timing receivers, International Atomic Time can be ascertained to within a few nanoseconds, but with a delay of 30 to 60 days depending on the date which interests us in the month. (This arises because of the monthly organisation of calculations used to determine TAI.) If there is no great hurry, this is a simple means of comparing clocks in these laboratories. TAI can be replaced by Coordinated Universal Time (UTC), taking care to include possible leap seconds. These two time scales are representations of $\underline{t}$, not of t (see Section 3.3.2e). In the following formulas, we shall denote TAI or UTC by $\underline{t}^*$.

We calculate the rate N_A of clock A relative to TAI by

$$N_A = \frac{\left[\tau_A - \underline{t}^*\right](\underline{t}_2^*) - \left[\tau_A - \underline{t}^*\right](\underline{t}_1^*)}{\tau_A(\underline{t}_2^*) - \tau_A(\underline{t}_1^*)} , \tag{5.70}$$

so that

$$N_A = 1 - \left\langle \frac{\mathrm{d}\underline{t}^*}{\mathrm{d}\tau_A} \right\rangle = 1 - \left\langle \frac{\mathrm{d}\underline{t}^*}{\mathrm{d}\underline{t}} \right\rangle \left\langle \frac{\mathrm{d}\underline{t}}{\mathrm{d}\tau} \right\rangle \left\langle \frac{\mathrm{d}\tau}{\mathrm{d}\tau_A} \right\rangle . \tag{5.71}$$

We shall see in Chapter 7 that the duration of the TAI scale unit, in seconds on the rotating geoid, can be obtained from BIPM publications. Let D_{TAI} be the numerical value of its time average over the interval considered. Then,

$$\left\langle \frac{\mathrm{d}\underline{t}^*}{\mathrm{d}\underline{t}} \right\rangle = (D_{TAI})^{-1} . \tag{5.72}$$

Using (3.29) and (5.50), we obtain

$$\langle \Phi_A \rangle = (D_{TAI})^{-1} + N_A + \langle \underline{h}\left(\underline{t}^*, \mathbf{x}(\underline{t}^*)\right) \rangle , \tag{5.73}$$

to a sufficient level of accuracy.

Proceeding in the same way with clock B, over the same interval of $\underline{t}^*$, we obtain $(\Phi_A - \Phi_B)$, which is independent of the value of D_{TAI}.

This method, like the one before, is only really suitable for accuracy and long term stability, at least for high quality standards. The stability of TAI is good enough for strict simultaneity of measurements to be unnecessary. It is worth noting that (5.73) gives direct access to the proper second with the frequency accuracy of TAI. We shall return to this point in Chapter 7.

(b) Frequency comparisons based on phase measurements

Local comparisons of the frequencies of two oscillators can be carried out using the methods described in Section 5.2.3.

For remote frequency comparisons, a radio broadcast frequency can be used as intermediary. As explained in Chapter 4, this method is no longer of much interest to fundamental metrology. The broadcasting of stable frequencies is a simple and economical way of sharing frequency standards with users who do not need ultimate accuracies.

Let us nevertheless record the relationship between the normalised proper frequency $[\Phi_E]_E$ of an emitter E measured at the location of E, and the normalised proper frequency $[\Phi_E]_A$ at reception at A, viz.,

$$\frac{[\Phi_E]_A}{[\Phi_E]_E} = \frac{1 - c^{-1}(\mathbf{v}_A \cdot \mathbf{n}) + c^{-2}\left[U(\mathbf{x}_A) + v_A^2/2\right]}{1 - c^{-1}(\mathbf{v}_E \cdot \mathbf{n}) + c^{-2}\left[U(\mathbf{x}_E) + v_E^2/2\right]} . \tag{5.74}$$

In this relation, velocities are expressed in the non-rotating geocentric coordinate system. $\mathbf{v}_E$ is the velocity of the emitter and $\mathbf{v}_A$ the velocity of the receiver, at the moments of emission and reception, respectively. The potential U is evaluated at the locations of E and A and $\mathbf{n}$ is the unit vector along EA, constructed with the space coordinates in the direction of propagation. This equation is correct to about $\pm 10^{-14}$ in the vacuum for emitters carried by satellite, out to the distance of geostationary satellites.

Applying the same relation to a receiver B at another location, we obtain the ratio of the normalised proper frequencies received for the same emission:

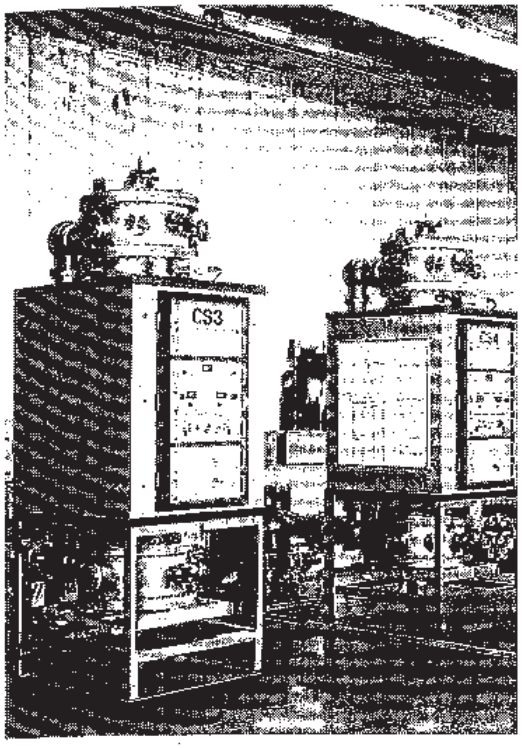
$$\frac{[\Phi_E]_A}{[\Phi_E]_B} = \frac{1 - c^{-1}(\mathbf{v}_A \cdot \mathbf{n}) + c^{-2}\left[U(\mathbf{x}_A) + v_A^2/2\right]}{1 - c^{-1}(\mathbf{v}_B \cdot \mathbf{n}) + c^{-2}\left[U(\mathbf{x}_B) + v_B^2/2\right]} . \tag{5.75}$$

Comparison between normalised proper frequencies Φ_{HA} and Φ_{HB} of standards located at A and B is given by

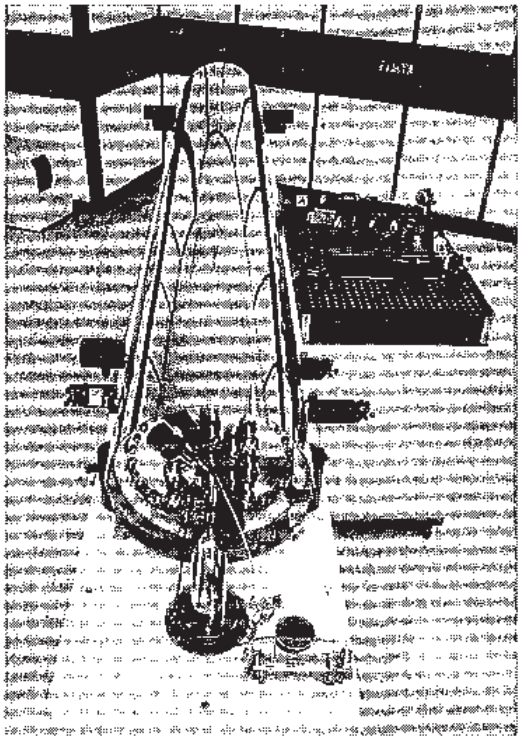
$$\frac{\Phi_{HA}}{\Phi_{HB}} = \frac{\Phi_{HA}}{[\Phi_E]_A} \frac{[\Phi_E]_B}{\Phi_{HB}} \frac{[\Phi_E]_A}{[\Phi_E]_B} . \tag{5.76}$$

The first two ratios on the right hand side are found from local measurements whilst the third is calculated using (5.75).

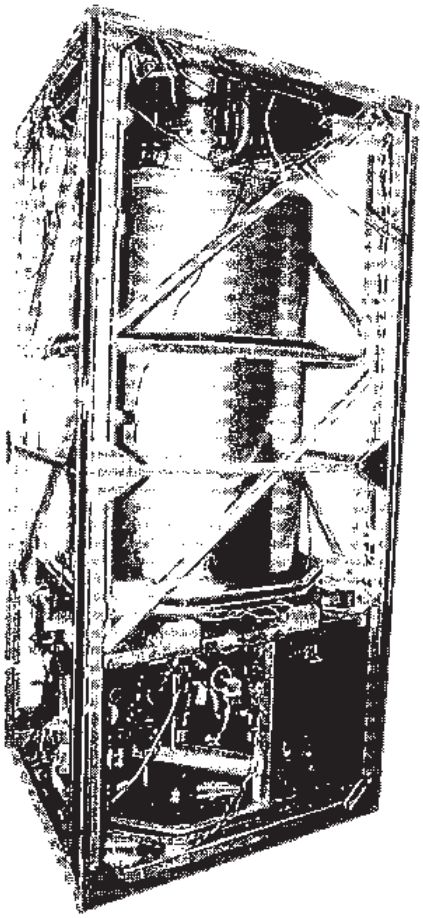
In applications to one-way transfers from a satellite, accuracy is limited to about 10^{-13} because of refraction effects.



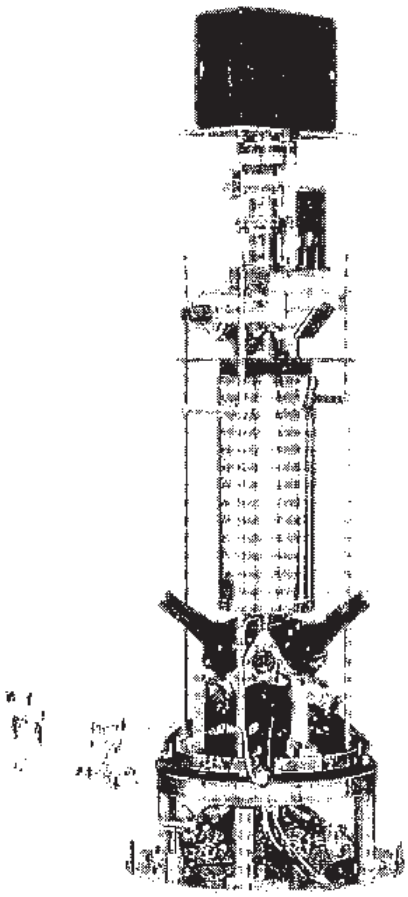
Magnetically deflected caesium beam primary standards CS3 and CS4. (Courtesy of A. Bauch, PTB.)



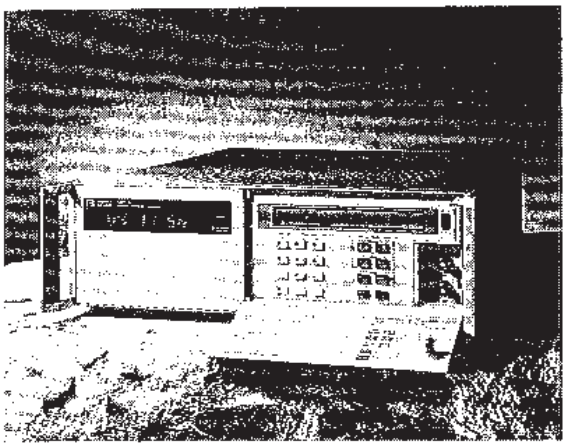
Optically pumped caesium beam primary standard NIST 7. (Courtesy of R.E. Drullinger, NIST.)



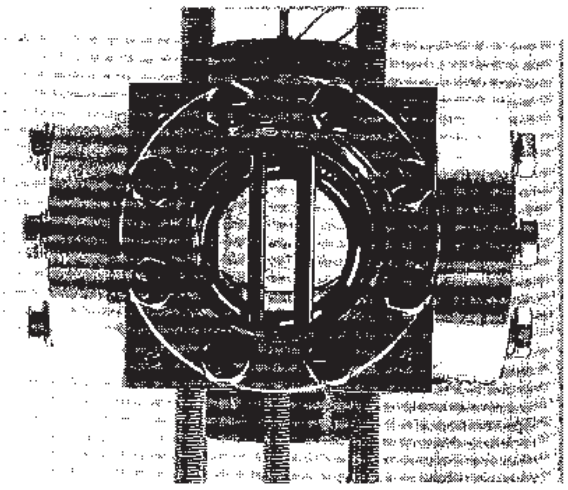
Primary standard FO1 based on a fountain of laser-cooled caesium atoms. (Courtesy of A. Clairon, LPTF.)



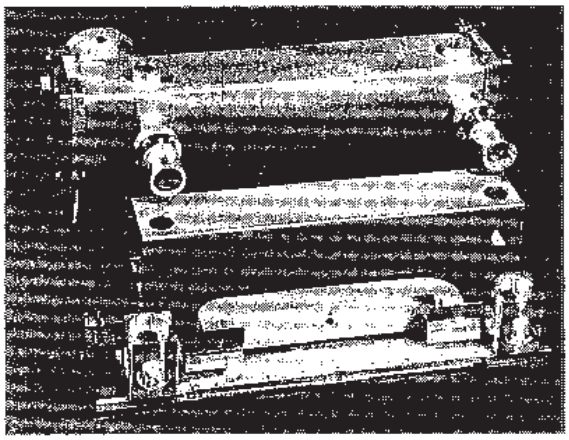
Prototype orbiting atomic clock based on cooled atoms. (Courtesy of A. Clairon and P. Laurent, LPTF.)



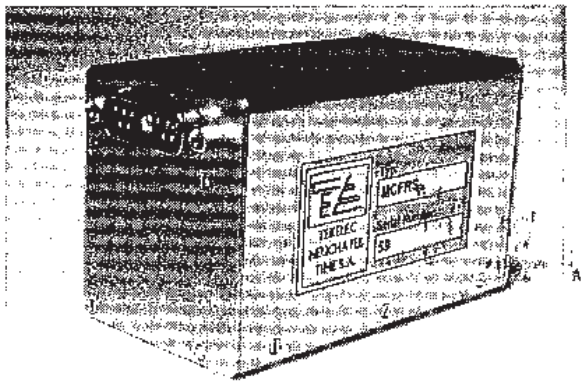
Commercial caesium clock.
(Courtesy of D. Ferrand,
Hewlett-Packard.)



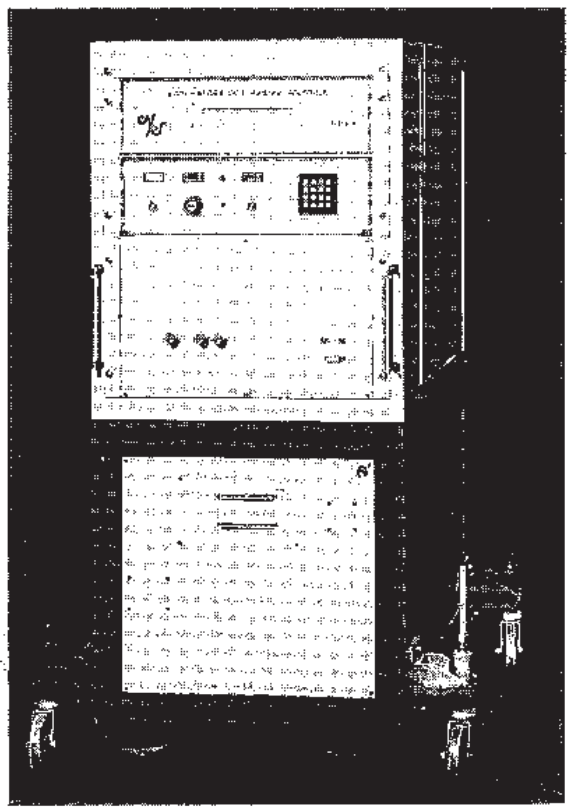
Linear radio frequency trap
for Hg^+ ion confinement.
(Courtesy of J.D. Prestage, JPL.)



Components of a compact prototype
for an optically pumped caesium clock.
(Courtesy of P. Cerez, LHA.)



Compact rubidium clock.
(Courtesy of P. Rochat, TNT.)



Hydrogen maser.
(Courtesy of R. Barillet and P. Petit,
LHA.)

6

Atomic frequency standards

The aim of this chapter is to describe the principles, operation and features of the main types of atomic clock, i.e., caesium clocks, hydrogen masers, rubidium clocks and confined ion clocks. The greater part of the discussion will be devoted to the caesium clock since it is used to realise the present definition of the second.

The reader will also find a general description of these frequency standards in [6.1, 6.2, 6.3], for example, and an in-depth study in [6.4].

6.1 Spectroscopic basis for atomic frequency standards

6.1.1 Invariance and universality of atomic properties

Up to the present day, all experimental observations in spectroscopy have proved compatible with the hypothesis that atomic properties are the same at all times and in all places, when they are assessed by an observer situated close to the atom and accompanying it in the same motion. This postulate is clearly a specific instance of Einstein's equivalence principle, stated in Section 3.2.1e.

It is therefore possible to build instruments which, using a specified atomic transition, are all able to deliver a signal in real time with the same frequency, anywhere and at any time, provided that relativistic effects due to the non-coincidence of the atom and the observer have been properly accounted for. This frequency only depends on fundamental physical constants.

The fact that the transition frequencies of different atoms remain in constant ratio to one another, up to the limits of experimental error, constitutes a very precise corroboration of the equivalence principle in the area of spectroscopy.

6.1.2 Spectral properties of alkali atoms

The main atomic clocks take advantage of a hyperfine structure transition in hydrogen, rubidium and caesium atoms, or in ionised mercury. The alkali metals, including rubidium and caesium, have an unpaired electron in the outer electron shell. Their inner electron subshells are either full or empty. The Hg^+ ion has a similar electronic configuration and we shall call it 'alkali' by extension. Likewise for the hydrogen atom, which has only one electron.

Let us now review the main properties of the alkali atoms insofar as they are necessary for describing the workings of atomic clocks [6.5].

(a) Fine structure

The outermost electron occupies an electron shell characterised by the principal quantum number n. Its value is given in Table 6.1 for the energy levels we are considering.

An orbital angular momentum denoted by $\hbar\mathbf{L}$ is associated with the motion of the electron around the nucleus, where $\hbar$ is Planck's constant h divided by 2π. The associated quantum number is L. In the alkali atoms, the total orbital angular momentum reduces to that of the single outer electron. We consider only two values of L, viz., $L = 0$ and $L = 1$.

For $L = 0$, the energy of the outer electron is minimal. The atom is said to be in its *ground state*. Using the standard spectroscopic notation†, this state is denoted $^2S_{1/2}$. Under the effect of a suitable excitation energy, the atom can be shifted into an *excited state*. The first such state corresponds to $L = 1$. The energy difference between the ground state and the first excited state is of the order of a few electron volts (eV). It follows that, if the transition between ground state and first excited state occurs through emission or absorption of electromagnetic radiation, the wavelength of this radiation will lie in the optical part of the electromagnetic spectrum.

All electrons possess an intrinsic property known as spin. The spin quantum number S takes the value 1/2. An angular momentum $\hbar\mathbf{S}$ is associated with the electron spin. The spin angular momentum of an alkali atom is the same as that of the unpaired electron.

† The spectroscopic notation arises from long-standing conventions. The upper case letter characterises the value of the quantum number of the orbital angular momentum. For example, S and P correspond to $L = 0$ and $L = 1$, respectively. The letter S should not be confused with the spin quantum number. The upper index placed before the letter gives the value of $2S + 1$, called the multiplicity, where S is now the spin quantum number. The lower index placed after the letter specifies the value of the quantum number J for the total angular momentum. When this notation is preceded by a number, it is the value of the principal quantum number. As an example, the ground state of the caesium atom is denoted $6^2S_{1/2}$.

Table 6.1. *Main spectroscopic features of alkali atoms used to realise atomic time and frequency standards. The value of the hyperfine transition frequency of caesium 133 is fixed by the definition of the second. Uncertainties in the measured values of the hyperfine transition frequencies of the other atoms are indicated.*

Atom (or ion)	Atomic mass	n	I	F	λ_{D_1} (nm)	λ_{D_2} (nm)	ν_0 (Hz)
H	1	1	1/2	0;1	121.5673	121.5668	1 420 405 751.770(3)
Rb	87	5	3/2	1;2	794.8	780.0	6 834 682 610.904 29(9)
Cs	133	6	7/2	3;4	894.3	852.1	9 192 631 770
Hg^+	199	6	1/2	0;1	194.2	165.0	40 507 347 996.841 6(4)

The total angular momentum $\hbar\mathbf{J}$ of the electron is the vector sum of its orbital angular momentum and its spin angular momentum. According to the combination rules for atomic angular momenta, the associated quantum number J can only take values $L + 1/2$ or $|L - 1/2|$. This means that $J = 1/2$ in the ground state and $J = 1/2$ or $3/2$ in the first excited state. Corresponding to the two possible values of J in this excited state are two slightly different energy levels. These are referred to as the *fine structure levels* and denoted $^2P_{1/2}$ and $^2P_{3/2}$ in the spectroscopic notation. The energy difference between them can be expressed in terms of the fine structure constant α which has a value of roughly $1/137$. These levels are relevant to the optical pumping method, soon to be discussed. Two very close spectral lines denoted D_1 and D_2 in the emission or absorption spectra of the alkali atoms arise from transitions between the $^2S_{1/2}$ and $^2P_{1/2}$ for the first and between $^2S_{1/2}$ and $^2P_{3/2}$ for the second. Wavelengths of the fine structure doublets for hydrogen, rubidium, caesium and the Hg^+ ion are given in Table 6.1.

(b) Hyperfine structure

The proton and neutron also possess spin, with value $1/2$. The nuclear angular momentum $\hbar\mathbf{I}$ is found by combining those of the constituent nucleons in a certain way. The value I of the associated quantum number is an odd multiple of $1/2$ for alkali isotopes with odd atomic mass (see Table 6.1).

The total angular momentum $\hbar\mathbf{F}$ of the atom is the vector sum of the angular momenta of the electron cloud and the nucleus. The quantum number F associated with the total angular momentum is such that $|I - J| \leq F \leq I + J$ and F can only vary by integer steps between these bounds. In the ground state

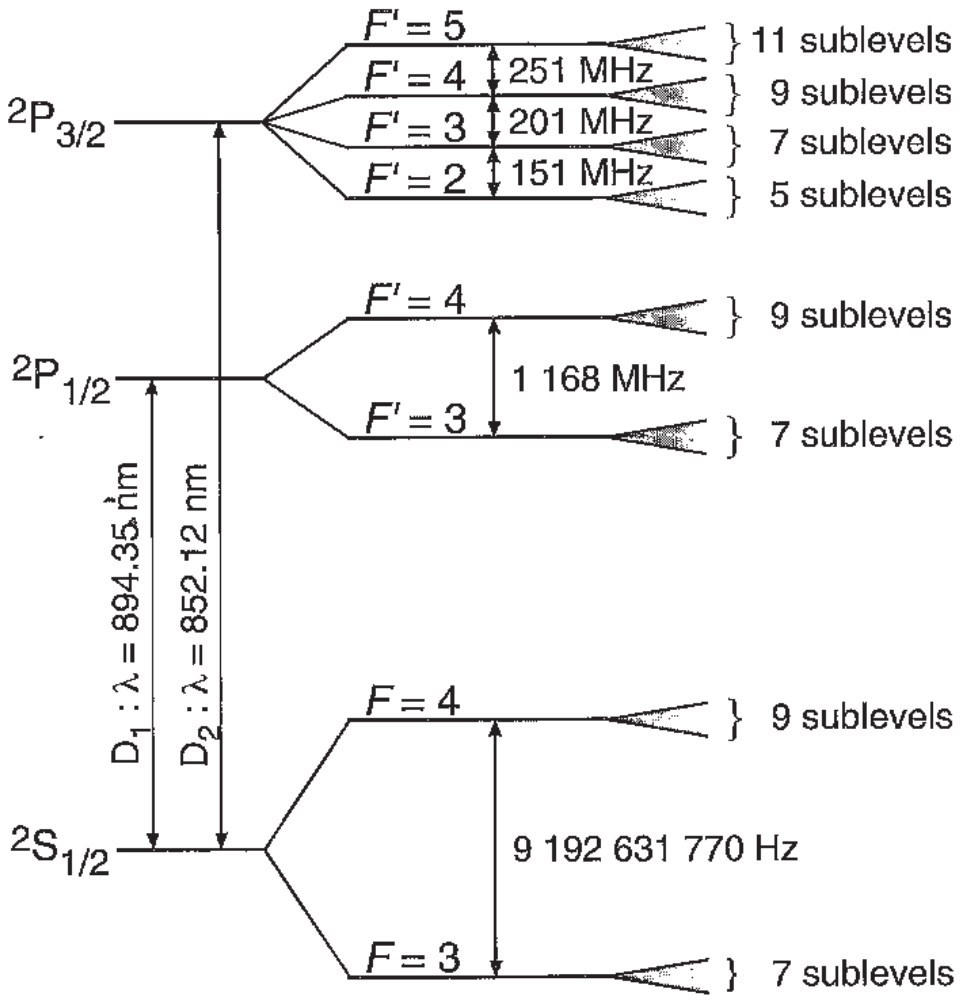


Fig. 6.1. Energy levels of the caesium atom in the ground state and the first excited state.

and also in the $^2P_{1/2}$ state, F is equal to $I - 1/2$ or $I + 1/2$. In the $^2P_{3/2}$ state, the number of possible values of F is greater than 2, if $I > 1/2$.

The projection of the angular momentum $\hbar\mathbf{F}$ along a quantisation axis can only take a discrete range of values, equal to $\hbar m_F$. The whole number m_F is the magnetic quantum number associated with F. Its $2F + 1$ possible values are $-F, -F + 1, \ldots, F - 1, F$.

Figure 6.1 shows the energy levels of the caesium atom in the $^2S_{1/2}$, $^2P_{1/2}$ and $^2P_{3/2}$ states, and also the possible values of F in this case. Throughout the following, F' will denote the value of F in the excited states $^2P_{1/2}$ and $^2P_{3/2}$.

Magnetic moments are associated with the nuclear and electron spins. That of the nucleus is

$$\boldsymbol{\mu}_I = g_I \mu_B \mathbf{I}, \tag{6.1}$$

where μ_B is the Bohr magneton and g_I the nuclear Landé factor. The magnetic

moment of the unpaired electron is given by

$$\boldsymbol{\mu}_J = -g_J \mu_B \mathbf{S} \,, \tag{6.2}$$

where g_J is the Landé factor of the electron bound to the atom. The value of g_J is positive and lies close to 2. The value of g_I is of the order of 1000 times smaller. It is positive for the atoms and the ion considered here. As an example, in the hydrogen atom $g_I = 3.042\,01 \times 10^{-3}$.

These magnetic moments give rise to a mutual magnetic energy. In a zero magnetic field, the magnetic interaction splits each of the energy levels considered above into as many *hyperfine levels* as there are accessible values of F or F'. This splitting is illustrated in Figure 6.1.

Calculating the energy difference between the two hyperfine structure levels in zero external field involves expressing the magnetic excitation created by the outer electron at the nucleus. In the hydrogen atom, the transition frequency ν_H between levels $F = 0$ and $F = 1$ of the ground state is given to a first approximation by

$$\nu_H = \frac{4}{3}\frac{mc^2\alpha^4 g_I}{h}\left(1 + \frac{m}{M_p}\right)^{-3} \,, \tag{6.3}$$

where m and M_p are the electron and proton masses, respectively, c is the speed of light and α the fine structure constant, equal to $e^2/4\pi\varepsilon_0\hbar c$, with e the electron charge and ε_0 the permittivity of the vacuum. This approximate formula gives a value for the hyperfine transition frequency with relative deviation 10^{-3} from its experimentally determined value. In fact, a large number of corrections need to be considered, expressed in terms of the fine structure constant, but we shall not discuss them here. Advanced and complex calculations are required to calculate them. At the present time, the theoretical expression for the hyperfine transition frequency of the hydrogen atom in the ground state is known to the same kind of accuracy as the fundamental physical constants.

Calculation of the hyperfine splitting in the ground states of alkali atoms other than hydrogen is a still more complex exercise. Indeed, even at the lowest order of approximation, the presence of electrons in the inner shells must be taken into account.

As a general rule, the magnetic interaction energy between the outer electron and the nucleus is extremely small. It follows that the hyperfine frequency ν_0 in the ground state lies in the ultrahigh frequency range, as can be seen from Table 6.1.

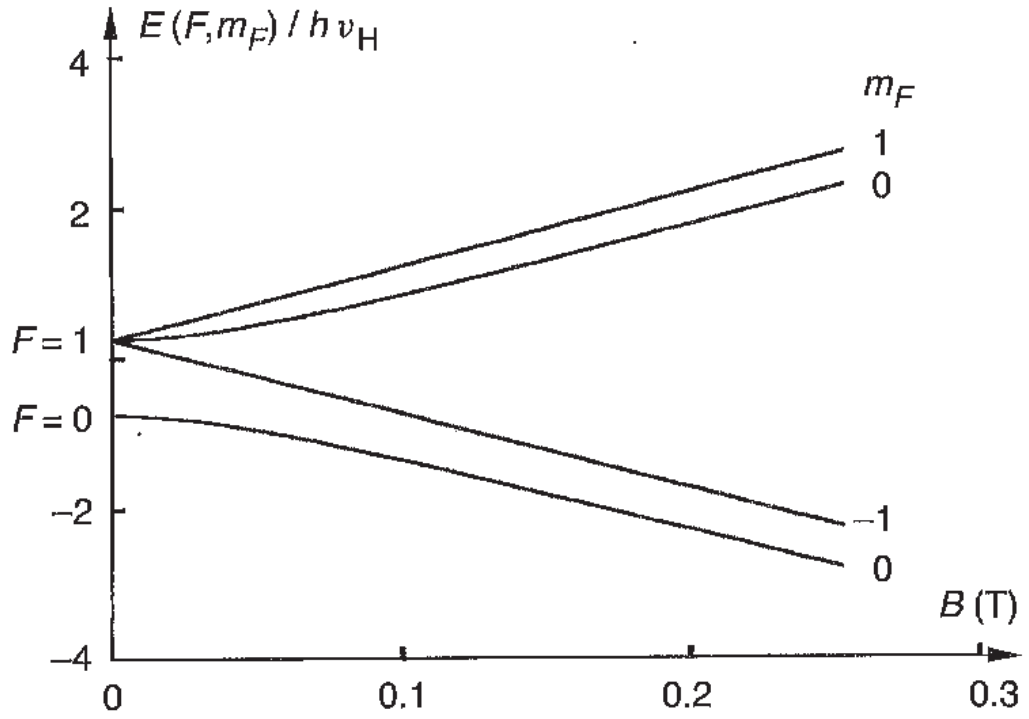


Fig. 6.2. Variation in energy of the hyperfine sublevels of the ground state hydrogen atom as a function of applied magnetic field.

A static magnetic flux density† **B** defines the direction of quantisation and splits the hyperfine levels into as many hyperfine sublevels as there are possible values of m_F, i.e., $2F + 1$. This is the *Zeeman effect*. The variation in energy $E(F, m_F)$ of each of these sublevels can be calculated exactly as a function of the applied field B for the alkali atoms in the ground state. It is given by the Breit–Rabi formula

$$E(F, m_F) = -\frac{h\nu_0}{2(2I+1)} - g_I \mu_B B m_F + \varepsilon \frac{h\nu_0}{2}\left(1 + \frac{4m_F}{2I+1}x + x^2\right)^{1/2}, \tag{6.4}$$

where

$$x = \frac{(g_J + g_I)\mu_B B}{h\nu_0}. \tag{6.5}$$

When $F = I + 1/2$, $\varepsilon = +1$, and when $F = I - 1/2$, $\varepsilon = -1$. This formula can be used to calculate the shift in frequency of a particular transition between two hyperfine sublevels as a function of the applied magnetic field. Figures 6.2 and 6.3 show the variation in $E(F, m_F)$ for the hydrogen atom and the caesium atom, respectively.

Equation (6.4) and Figures 6.2 and 6.3 bring out some remarkable properties.

† Throughout the following, we shall use the more common term 'magnetic field'.

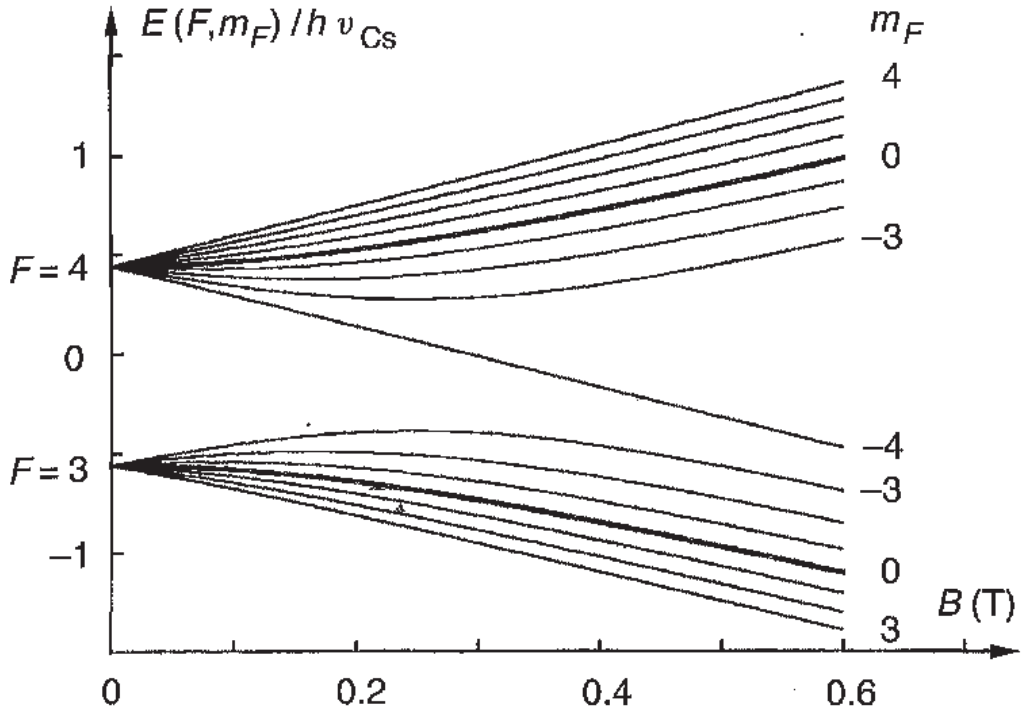


Fig. 6.3. Variation in energy of the hyperfine sublevels of the ground state caesium atom as a function of applied magnetic field.

(i) The energy of the $m_F = 0$ sublevels varies very slowly when the applied magnetic field is small. It is the transition between these levels, sometimes called the *clock transition*, that is exploited in atomic clocks. The change $\Delta\nu_Z$ in the frequency of the transition between the two $m_F = 0$ sublevels varies quadratically with B according to

$$\Delta\nu_Z = K_0 B_0^2 , \tag{6.6}$$

where the factor K_0 is a function of atomic constants alone, and B_0 denotes the value of the magnetic field when it is small. The value of B_0 is typically 10^{-7} T in the hydrogen maser and 6×10^{-6} T in the caesium beam clock. The corresponding values of $\Delta\nu_Z$ are then 2.77×10^{-3} Hz and 1.5388 Hz, respectively.

(ii) For small values of the applied magnetic field, the energies of levels with $m_F \neq 0$ vary either linearly, or almost linearly. Hence, with the values of B_0 just quoted, the frequency of the transition

$$F = 1\,,\ m_F = 0 \quad \longleftrightarrow \quad F = 1,\ m_F = \pm 1$$

is 1.399 kHz for hydrogen, whilst the shift for the transition

$$F = 3\,,\ m_F = 1 \quad \longleftrightarrow \quad F = 4\,,\ m_F = 1$$

is 42.050 kHz for caesium. It is therefore possible, by measuring the

frequency of certain transitions between hyperfine sublevels, to establish the value of the applied magnetic field. The frequency shift for the clock transition can then be deduced using (6.6).

(iii) For large values of the applied magnetic field, energy levels separate into two groups. The physical reason is that the magnetic energy acquired by the electron dominates over its interaction energy with the nucleus. The magnetic behaviour of the atom thus begins to resemble that of the outer electron. The energy of the latter is an increasing function of B_0 when the electron spin is oriented antiparallel to the direction of the magnetic field, and a decreasing function of B_0 when the electron spin lies parallel to the vector $\vec{\mathbf{B}}_0$. The $m_F = 0$ sublevels therefore belong to different groups. Atoms can therefore be assigned to one or other of the levels between which the clock transition is produced. This feature lies at the origin of the *magnetic state selection method* (see Section 6.2.2).

Excited states also possess a hyperfine structure if $I \neq 0$. Figure 6.1 shows the hyperfine structure levels of the $^2P_{1/2}$ and $^2P_{3/2}$ states of the caesium atom, and also the number of sublevels corresponding to the possible values of m_F. Hyperfine energy differences are smaller than in the ground state.

An electric field applied to an atom shifts its energy levels. This is the *Stark effect*. However, the effect is quadratic for the energy levels considered here. In practice, the associated frequency shifts are very small. The very low sensitivity to the electric field provides a way of confining ions, such as mercury ions, in order to observe their clock transition in optimal conditions.

Let us just mention that the quantum number I is zero for most isotopes with even atomic mass. Consequently, they do not exhibit hyperfine structure.

6.1.3 Selection rules

Interaction cannot occur between an atom and electromagnetic radiation unless the laws of energy and angular momentum conservation are obeyed.

If E_1 and E_2 are the energies of two atomic levels ($E_2 > E_1$), a necessary condition for interaction is that the frequency ν_0 of the electromagnetic wave should satisfy Bohr's condition

$$E_2 - E_1 = h\nu_0 \,, \tag{6.7}$$

which expresses energy conservation.

Energy exchange can occur via three different processes.

- *Absorption.* An incident photon of energy $h\nu_0$ is absorbed by an atom in

the state of energy E_1. The atom moves into the state of energy E_2 and the photon disappears.

- *Spontaneous emission.* An atom previously raised to the E_2 energy level spontaneously decays, at a random instant of time and without external cause. Its energy falls to E_1 and a photon of energy $h\nu_0$ is created, called the *fluorescence photon.*
- *Stimulated* or *induced emission.* This type of interaction, predicted by Einstein, is the basis for *laser* and *maser* effects. An excited atom of energy E_2 in the presence of a photon of energy $h\nu_0$ falls to the E_1 energy level under the influence of the incident electromagnetic radiation. A photon is thereby created, but without the incident photon disappearing in the process. A key feature of stimulated emission is that the electromagnetic wave associated with the emitted photon is in phase with the incident wave, so that the latter is amplified.

In order for a transition to take place between two atomic levels, the condition (6.7) is necessary but not sufficient. Selection rules must also be satisfied. These result from properties of the atomic angular momenta. Conservation of angular momentum in the interaction with the radiation requires the latter to have a certain polarisation.

Let us consider transitions between hyperfine sublevels of a given state, e.g., the ground state. They include the clock transition and the transitions used to measure the applied magnetic field B_0. Two types of transition are allowed:

- those with $\Delta F = \pm 1$ and $\Delta m_F = 0$, which require an ultrahigh frequency magnetic field parallel to the static magnetic field;
- those with $\Delta F = 0$ and $\Delta m_F = \pm 1$ or $\Delta F = \pm 1$ and $\Delta m_F = \pm 1$, which require a low frequency ($\Delta F = 0$) or ultrahigh frequency ($\Delta F = \pm 1$) magnetic field perpendicular to the static magnetic field.

Transitions can also occur between sublevels of hyperfine structure belonging to different states such as $^2S_{1/2}$, $^2P_{1/2}$ and $^2P_{3/2}$. These arise in optical pumping of caesium and rubidium atoms, and of mercury ions, as well as in the optical detection of their clock transition. Transitions between fine structure levels $^2P_{1/2}$ and $^2P_{3/2}$ are forbidden. The only possible transitions are those between levels of the state $^2S_{1/2}$ and levels of the state $^2P_{1/2}$ on the one hand, and between levels of the state $^2S_{1/2}$ and levels of the state $^2P_{3/2}$ on the other. Allowed transitions are those for which either $F' - F' = \pm 1$ and $m_F - m_{F'} = 0$, or either $F - F' = 0$ or ± 1 and $m_F - m_{F'} = \pm 1$. To simplify, we shall not specify the corresponding polarisation of the radiation.

The selection rules allow us to choose between the various possible transitions.

6.1.4 Resonance width

In practice, time and frequency standards are designed in such a way that the resonance is not Doppler broadened (see Section 6.2.3). Its width $\Delta\nu_0$ is then related to the observation time Δt by the Heisenberg energy–time uncertainty relation

$$\Delta\nu_0 \sim \frac{1}{\Delta t} . \tag{6.8}$$

The probability of spontaneous emission per unit time varies as the cube of the transition frequency between the relevant level and the ground state. In practice, it turns out that the lifetime τ of the fine structure levels is very small, of the order of a few nanoseconds: 33 ns for the Cs atom and 2 ns for the Hg^+ ion, for example. The natural width $1/2\pi\tau$ of the resonance under consideration, that is, the one resulting from spontaneous emission alone, is thus close to 5 MHz for Cs and 100 MHz for Hg^+.

In contrast, the hyperfine sublevels of the ground state with $F = I + 1/2$ have a much longer lifetime, of the order of a few years. This is due to the small frequencies of hyperfine transitions in the ground state, compared with optical transition frequencies. The lowest energy sublevels, with $F = I - 1/2$, have almost infinite lifetimes. Practically speaking, spontaneous emission does not therefore affect the widths of hyperfine resonances in the ground state.

In atomic frequency standards, typical observation times range between 1 ms in the rubidium clock and certain caesium clocks and a few seconds in the Hg^+ clock. The widths of atomic resonances then lie between about 10^3 and 10^{-1} Hz.

It follows that, for typical values of the applied magnetic field, the various possible transitions are well separated from one another. To a first approximation, everything therefore happens as though, amongst all the hyperfine structure levels, only two of them actually interact with the ultrahigh frequency magnetic field. The evolution of the atom's properties can then be treated within the simplified context of a two-level system.

The quality factor $Q_{at} = \nu_0/\Delta\nu_0$ of the atomic resonance is very high. It lies between 10^7 and 10^{12}. The very high level of this parameter is a decisive factor when considering stabilities of atomic frequency standards. To some extent, it contributes to their accuracy.

6.2 Operations involved in an atomic clock

Similar operations take place in all types of atomic clock. Atoms are observed in extremely dilute conditions. A population difference is achieved between hyperfine levels. The clock transition is brought about in conditions where very high resolution spectroscopy is possible. The transition is detected. A weak magnetic field, protected from outside perturbation, is set up in the zone where the atoms interact with the electromagnetic radiation. Electronic means are used to provide users with practically applicable signals.

6.2.1 Observation in very dilute media

It is clear that the atoms must be as little perturbed as possible by neighbouring atoms. This is why, in atomic clocks, atoms are always diluted, in order to minimise the rate at which interatomic collisions occur. Atoms are observed in an atom beam, a diffuse cloud or a very low pressure cell. In an atom beam (e.g., caesium clock), the probability of collision is lower than in a cell for the same density of atoms. Clouds can be in ballistic motion, as in a fountain of cooled caesium atoms, or confined, as in the mercury ion clock. Cells may be materially constituted (e.g., hydrogen maser, rubidium clock), or produced by means of an electric field (e.g., mercury ion clock). Except in the rubidium clock, where the cell is sealed, efficient pumping of residual gases is essential to maintain a high vacuum.

6.2.2 Preparation and detection of atomic states

Good conditions for observing transitions between ground state hyperfine levels require preparation of the atomic system.

Consider two sublevels of the hyperfine structure, with energies E_1 and E_2 ($E_2 > E_1$). Let n_1 be the population of level E_1 at time t, i.e., the number of atoms with energy E_1 at that time, and likewise, n_2 the population of level E_2 at the same time. Under the effect of the electromagnetic radiation, and during a time lapse $\mathrm{d}t$, absorption causes atoms to transit from level E_1 to level E_2, so that the population of the latter increases by $\mathrm{d}n_2$. Then

$$\mathrm{d}n_2 = p_{12} n_1 \mathrm{d}t \ , \tag{6.9}$$

where p_{12} is the probability of absorption per unit time. Likewise, induced emission causes atoms to transit from level E_2 to level E_1, and the population of level E_2 thus changes by

$$\mathrm{d}n_2 = -p_{21} n_2 \, \mathrm{d}t \ , \tag{6.10}$$

where p_{21} is the probability of induced emission per unit time.

Einstein showed that $p_{12} = p_{21}$. The total change in the population of energy level E_2 is therefore

$$\mathrm{d}n_2 = p_{12}(n_1 - n_2)\,\mathrm{d}t\ . \tag{6.11}$$

Up to a sign, this is also the change in population of level E_1, since population exchange is limited to these two levels alone.

Equation (6.11) shows that the population of a level changes more in the case where the population difference $n_1 - n_2$ is great to begin with. Changes in atomic properties will therefore be correspondingly easier to observe, in the sense that the signal-to-noise ratio will be larger, when the population difference between the two levels is increased.

At thermodynamic equilibrium, the ratio of the populations in two non-degenerate levels, such as those considered here in non-zero external magnetic field, is given by *Boltzmann's law*. Hence,

$$\frac{n_1}{n_2} = \exp\frac{E_2 - E_1}{kT}\ , \tag{6.12}$$

where k is Boltzmann's constant and T the temperature. If we assume that the transition frequency between the two levels is 10 GHz and the temperature 300 K, then

$$\frac{n_1 - n_2}{n_2} = 1.6 \times 10^{-3}\ .$$

The relative value of the population difference for the two levels between which a clock transition takes place is thus very small at thermodynamic equilibrium.

In order to ensure that the hyperfine transition is detected in satisfactory conditions, this equilibrium must be broken using methods which create as large a population difference as possible, namely, *magnetic deflection* [6.6], or *optical pumping* [6.7]. It is here that we must pay for the advantages of using the very low hyperfine transition frequencies.

(a) Magnetic deflection

The potential energy of hydrogen atoms in the state $F = 0$, $m_F = 0$ and caesium atoms in the state $F = 3$, $m_F = 0$, among others, decreases as the applied magnetic field is raised, as can be seen from Figures 6.2 and 6.3. They are therefore attracted by regions of intense magnetic field, since every physical system tends to minimise its potential energy. In contrast, hydrogen atoms in the state $F = 1$, $m_F = 0$ and caesium atoms in the state $F = 4$, $m_F = 0$, among others, are repelled by regions of intense magnetic field. It can be shown

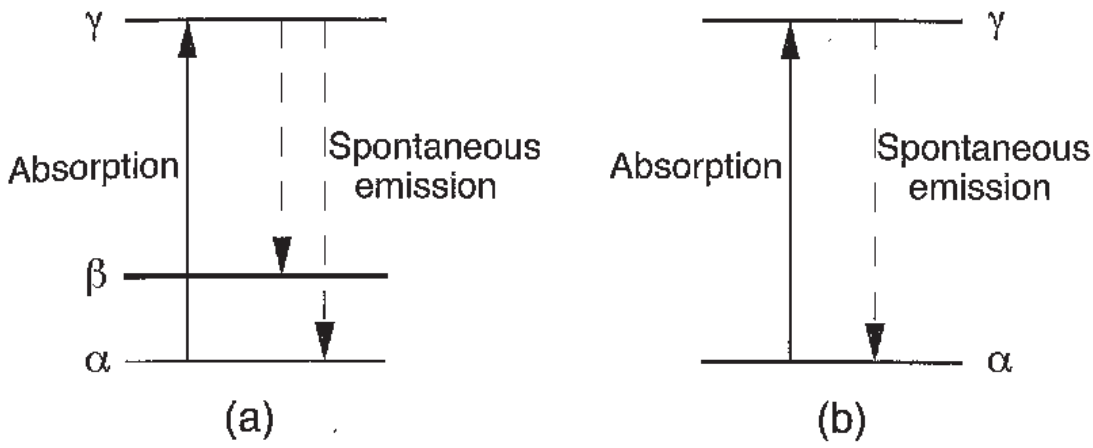


Fig. 6.4. Optical preparation and detection of states. (a) Pumping transition, (b) cycling transition.

that the force **F** acting on the atom is

$$\mathbf{F}(F, m_F) = -\frac{\partial E(F, m_F)}{\partial B} \nabla B \,, \tag{6.13}$$

where $E(F, m_F)$ is given by (6.4) and ∇B is the gradient of the applied magnetic field strength. The absolute value of $-\partial E(F, m_F)/\partial B$ is maximal in intense fields, where it is equal to the Bohr magneton. In practice, a field of around 1 tesla is used, with gradient of the order of a few teslas per centimetre.

In a strong, non-uniform field, the two $m_F = 0$ states are therefore subjected to opposing forces. In an atomic beam, atoms in one of these states can be guided towards a region where they are used to operate a clock, whilst atoms in the other state are evacuated towards a region where they play no useful role.

(b) Optical pumping

Optical preparation of the atomic state is used in optically pumped caesium clocks, rubidium clocks and mercury ion clocks.

Consider the simplified energy diagram in Figure 6.4a. For example, energy levels α and β are hyperfine structure levels $F = 3$ and $F = 4$ of the caesium atom in the ground state, and level γ is the hyperfine structure level $F' = 3$ of the state $^2P_{3/2}$.

At thermodynamic equilibrium, levels α and β are more or less equally populated, whereas γ is practically empty, since for the optical transition $\alpha \leftrightarrow \gamma$, we have

$$\frac{E_\gamma - E_\alpha}{kT} \gg 1 \quad \text{and hence} \quad \frac{n_\gamma}{n_\alpha} \ll 1 \,,$$

by (6.12). Radiation tuned to this transition can thus be absorbed, causing the atom to move from level α to level γ. Then, by spontaneous emission, the

atom falls very quickly back down to the ground state, either to level α or to level β. The probabilities of ending up in the latter two states are virtually the same. After several absorption and spontaneous emission cycles, atoms are thus transferred from level α to level β. A population difference is thereby created between the two levels of the ground state, all the more efficiently as the number of photons exciting the transition $\alpha \rightarrow \gamma$ is large. A laser tuned to the optical transition frequency is thus the most appropriate light source for implementing optical methods.

A transition like $F = 3 \rightarrow F' = 3$ in the caesium atom is called a *pumping transition* because the selection rules (Section 6.1.3) allow the atom to return to a hyperfine level that differs from the starting level, thereby producing a population difference between these levels.

(c) Optical detection

Optical methods can also be used to provide a measure of the population in a given level. There are several possibilities.

(i) When the level α in Figure 6.4a is depopulated, the incident light cannot be absorbed and the atomic medium becomes transparent. It can be made opaque again if atoms in level β are forced to come back to level α under the effect of radiation at a suitable ultrahigh frequency. This possibility is implemented in the rubidium clock.

(ii) Referring once again to Figure 6.4a, if level α is depopulated, the incident light no longer transfers atoms towards level γ. Spontaneous emission therefore ceases and there is no fluorescent light. If the transition $\beta \rightarrow \alpha$ is imposed by a suitable ultrahigh frequency, fluorescence will reappear. With sufficiently intense incident light, the mean number of fluorescence photons emitted each time an atom returns to level α lies between one and a few units. It depends on the values of F and F' associated with levels α and γ. This detection process is used in the optically pumped caesium clock and the mercury ion clock.

(iii) Let us suppose that the incident light is tuned to the transition between level $F = 3$ of the state $^2S_{1/2}$ and level $F' = 2$ of the state $^2P_{3/2}$ in the caesium atom. Then spontaneous emission can only take the atom back down to the starting level $F = 3$ because the transition $F' = 2 \rightarrow F = 4$ is forbidden (Section 6.1.3). We thus have a closed two-level system, as shown in Figure 6.4b, and a *cycling transition* is produced. Since the lifetime of the excited state is extremely short, the atom is very soon available to take part in the next cycle. The number of fluorescence photons is proportional to the intensity of the incident light and the duration of

interaction. When this detection method is used in a caesium clock, the detection yield amounts to several dozen fluorescence photons per atom.

6.2.3 Eliminating the Doppler effect

We begin by showing that, when appropriate steps are not taken, the Doppler effect is an obstacle to very high resolution spectroscopy.

When a source of radiation and a device capable of detecting that radiation are in relative motion, with a non-zero component of velocity along the straight line joining them, the frequency of the signal appears to change. This so-called Doppler effect is commonly observed in the propagation of sound waves.

In the case of a progressive electromagnetic wave, a simplified expression for the Doppler frequency shift implied by (3.6) is

$$\nu - \nu_0 = \nu_0 \frac{v}{c} \cos\theta \ , \tag{6.14}$$

where ν and ν_0 represent the frequency of radiation emitted by the atom in motion at speed v and at rest, respectively, and θ is the angle between the direction of motion of the atom and the direction of propagation of the radiation. The expression for the frequency shift is the same whether the atom emits or absorbs. We are assuming that $v/c \ll 1$, which amounts to considering the Doppler effect as limited to first order in v/c.

If the atom is in a collimated beam, the shift in the transition frequency is given directly by (6.14). Let us consider an atomic hydrogen beam, produced by a source at 300 K. If the angle between the beam and radiation directions differs from orthogonality by 1°, the hyperfine transition frequency is shifted by about 200 Hz, a fractional change of 1.4×10^{-7}. Such a shift is unacceptable for the applications considered here.

In a cell, thermal excitations cause the velocity vectors of the atoms to point in random directions, although isotropically on average. The resonance is not then shifted by the Doppler effect, but broadened. The shape of the resonance curve is determined by the Maxwell distribution of atomic or molecular velocities. At typical temperatures, the Doppler width of the resonance is of the order of 10^{-6} times the value of the resonance frequency. The exact value depends on the mass of the atom and the temperature of the gas. In other words, the quality factor of the resonance cannot be greater than 10^6 in the presence of such a Doppler effect. This value is too low to allow accurate enough measurement of the position of the resonance maximum.

In the atomic frequency standards to be described in the present book, limitations arising from the Doppler effect are overcome by restricting the spatial extent of the region in which atoms and radiation interact.

(a) Caesium clocks

The first order Doppler effect results from the change in phase of a progressive wave as a function of position. Now in the antinodal zone of the magnetic field (or electric field) of a pure stationary electromagnetic wave, the phase no longer depends on the position. Everything then happens as though the atoms located in this zone were subject to an on-the-spot oscillation and the first order Doppler effect does not occur. Such a configuration is used in the caesium clock. Each interaction between the atoms and the ultrahigh frequency magnetic field takes place in some antinodal zone of this field, which belongs to a stationary wave maintained in a resonant cavity.

(b) Other atomic clocks

Solutions adopted in other types of clock, where atomic motions are disordered, spring from a proposal made by Dicke [6.8]. In order to bring out the condition that must be fulfilled, consider an atom which, if it were at rest, would emit radiation of frequency ν_0 and wavelength λ. To simplify, we make the following assumptions:

(i) the atom moves backwards and forwards between two walls a distance d apart, and in a direction perpendicular to those walls;
(ii) the absolute value of the atomic velocity is the same in both directions;
(iii) atomic properties are not affected when atoms are forced to turn back at the walls.

An observer located on the axis of motion receives radiation whose frequency is alternately equal to $\nu_0(1 + v/c)$ and $\nu_0(1 - v/c)$, through Doppler shifting. The emission spectrum received therefore includes one component at frequency ν_0 and lateral lines at frequencies $\nu_0(1 \pm nv/c)$, where n is an integer. If more generally a set of atoms follows similar motions to the one described above, but at different speeds, the spectral components of the radiation received add together when at the frequency ν_0 and stand next to one another when they belong to lateral lines. It can be shown that if Dicke's condition

$$d < \frac{\lambda}{2} \tag{6.15}$$

is satisfied, the emission spectrum comprises a line at frequency ν_0 that is not Doppler broadened, standing above a pedestal formed by the set of lateral lines. The width of the pedestal is of the same order of magnitude as the usual Doppler width. However, it has small amplitude, so that only the central line is visible in practice. It is thus possible to observe the atomic emission

cleared of the Doppler effect. Similar arguments apply when atoms absorb incident radiation. Condition (6.15) for the suppression of Doppler broadening is equally valid for progressive waves and standing waves. The result can also be generalised to confinement in three dimensions.

Several methods can implement this spectroscopy without Doppler broadening.

- *Buffer gas.* Atoms whose hyperfine transition is to be observed are diluted in an inert gas, called a buffer gas. When (6.15) holds, and d is the mean free path between two collisions with the buffer gas, Doppler broadening is eliminated. This method is used in the rubidium clock.
- *Atom confinement.* In the hydrogen maser, atoms are confined in a cell by bouncing off its walls. In the mercury ion clock, an electric field maintains them within a bounded region of space. In both cases, condition (6.15) is designed into the system.

Equation (6.14) is a simplified expression for the Doppler frequency shift, to first order in v/c. In atomic frequency standards, the second order term is significant. It arises from relativistic time dilation for atoms moving relative to the measuring apparatus (Section 3.2.3). There is no way of suppressing or compensating for this effect, apart from selecting slow-moving atoms or implementing an efficient way of slowing them down (Section 6.4.3). The corresponding frequency shift $\Delta\nu_0$ is given by

$$\Delta\nu_D = -\frac{1}{2}\left(\frac{v}{c}\right)^2 \nu_0 \,, \quad \text{or} \quad \Delta\nu_D = -\frac{E_{kin}}{Mc^2}\nu_0 \,, \tag{6.16}$$

where E_{kin} is the kinetic energy of the atom of mass M. For a caesium atom emanating from an oven at a temperature of around 100 °C, $v \approx 200\ \mathrm{m\,s^{-1}}$. This means that $\Delta\nu_D/\nu_0 \approx -2 \times 10^{-13}$. Although very small, it cannot be neglected.

6.2.4 Interaction between atoms and electromagnetic field

Interaction between the atoms and the ultrahigh frequency wave usually occurs in a resonant electromagnetic cavity, in the presence of a weak, uniform, static magnetic field.

The oscillating and static magnetic fields are parallel over the volume occupied by the atoms to allow selective observation of the clock transition, for which $\Delta F = \pm 1$, $\Delta m_F = 0$ (Section 6.1.3).

Since the value of the static magnetic field is much smaller than the Earth's magnetic field, the interaction zone must be enclosed within a magnetic screen.

This plays the double role of significantly attenuating variations in the surrounding magnetic field.

Indeed, (6.6) shows that a small variation δB_0 in the magnetic field applied to the atoms causes a change in the resonance frequency equal to $2K_0B_0\delta B_0$. As an example, in a caesium clock, the relative change in the transition frequency is close to 3×10^{-14} when $\delta B_0/B_0 \approx 10^{-4}$.

Differential attenuation factors between 10^4 and 10^5 are commonly achieved in the hydrogen maser and caesium clock, for example. To reach such an efficiency, the magnetic shield is made from several nested shells of material with very high magnetic permeability. A magnetic field is produced by an electric current circulating in coils placed inside the shield, the aim being to obtain a field as homogeneous and stable as possible.

6.2.5 Associated electronic systems

There are two types of atomic clock: *passive clocks* and *active clocks*.

Passive clocks include the caesium clock, the rubidium clock and the mercury ion clock. The transition is excited by means of an electromagnetic signal and the closer the excitation frequency lies to the resonance frequency, the bigger is the response. The probe frequency must therefore be generated. This is achieved with the help of a quartz oscillator and frequency synthesis methods.† Using feedback loops, the resonance signal is used to servo control the quartz oscillator frequency on the atomic transition frequency.

The hydrogen maser can be used as a passive or an active clock. In the latter case, operating conditions are arranged so that it can generate an oscillation at the hyperfine transition frequency. The electronic system is then designed like a superheterodyne receiver, to detect the signal emitted by the maser. In addition, it servo controls the phase, and hence also the frequency, of a quartz oscillator on the maser oscillation frequency, via circuits which also make use of frequency synthesis.

In both cases, the atomic transition frequency is transposed to a value of say 10 MHz, which can be exploited directly by the user. Frequency dividers can provide pulses, for example, each second, which serve to mark the time scale produced by the clock. Electronic devices process signals in the ultra-high and radio frequency range. There is steady technological progress in this domain. Reliable, high-performance electronic systems are commonly realised

† Electronic circuits known as frequency mixers, multipliers and dividers are used to apply all four arithmetical operations to the quantity frequency. Signals can thus be produced with frequency ratios equal to any rational fraction.

for the purpose of precisely determining and exploiting the atomic resonance frequency.

Other electronic circuits can carry out many and varied operations, such as generating the extremely stable electric current needed to generate the static magnetic field within the coils.

6.3 Realising an approximation to the isolated atom at rest

The ideal atomic frequency standard would be one in which the reference frequency were defined by a transition between two levels well separated from neighbouring levels, in a single, isolated atom or ion at rest.

Such an ideal situation can only be approximated in practice, mainly because of technical constraints. Any deviation from the ideal causes the transition frequency to shift away from its unperturbed value. The shift may have several components [6.4, 6.9], due to the second order Zeeman effect or the second order Doppler effect, for example. However, these frequency shifts are very small when expressed as relative values and the uncertainty in their evaluation is such that the inaccuracy in certain laboratory caesium clocks was of the order of 10^{-15} in 2000.

We shall now summarise the physical origins of some residual frequency shifts, indicating briefly steps that might be taken to reduce them. A detailed discussion of these shifts and a description of work carried out to date to bring us closer to the ideal situation will then be given for each type of frequency standard.

(i) The two hyperfine structure levels are degenerate in zero field, and it might be thought that the transition frequency could be made insensitive to the value of B_0 simply by fulfilling this condition. However, in the case of caesium, for example, B_0 would have to be smaller than 5×10^{-11} T over the whole region in which the hyperfine structure transition was being observed in order to ensure that all possible transitions between sublevels of the states $I + 1/2$ and $I - 1/2$ did not differ by more than 1 Hz. The intensities of these transitions would also have to exhibit symmetry properties about the central transition $(4, 0) \leftrightarrow (3, 0)$, otherwise the observable resonance would be distorted and a practically unacceptable error would result when measuring the resonance frequency. In a device of reasonable size and cost, it is not yet known how to achieve a sufficiently homogeneous magnetic field, sufficiently insensitive to outside perturbation, for such a small value of B_0. It is therefore preferable to

submit atoms to a magnetic field that clearly separates the various possible transitions.

In fact the value of B_0 results from a compromise between two opposing requirements. It must be big enough to ensure that the wings of neighbouring transitions cannot distort the reference resonance to an unacceptable degree. But it must be as small as possible to limit the sensitivity of the clock transition frequency to fluctuations δB_0 in B_0.

It is easier to reduce the practically chosen value of B_0 if neighbouring resonances are very narrow and symmetrically placed about the reference resonance.

(ii) Atoms are subjected to blackbody radiation. This results from thermal motions of charge carriers in the materials surrounding the atoms under observation. The corresponding frequency shift, close to -2×10^{-14} at room temperature, has become significant because of current progress in atomic frequency standards. The only way to overcome this problem would be to cool part or all of the equipment. The technical constraints involved would be virtually insurmountable, at least for caesium beam standards.

(iii) Atoms are in motion. Despite precautions taken to suppress the first order Doppler effect, it cannot be totally eliminated. Moreover, the second order Doppler effect is unavoidable. Speeds of atoms or ions must therefore be reduced. Current and future progress in atomic frequency standards thus hinges upon finding efficient ways of slowing down the atoms and ions.

(iv) Collisions between alkali atoms, or between these and a buffer gas or the walls of a cell, shift energy levels of the hyperfine structure. A favourable configuration is provided by an atom beam or a low density ion or atom cloud. The ideal situation invoked at the end of this chapter is one in which a single ion can be observed.

(v) State selection by optical pumping causes a shift in the hyperfine structure levels called the light shift. This shift is significant in the rubidium clock. The problem can be overcome by spatially or temporally separating the optical and ultrahigh frequency interactions, as in the optically pumped caesium clock, the caesium fountain and the mercury ion clock.

(vi) Atoms couple with the ultrahigh frequency field, generally in a resonant cavity. The apparent frequency of the atomic transition depends slightly on the cavity tuning frequency (cavity pulling). The corresponding frequency offset is intrinsically smaller in a passive clock than in an active clock (unless the oscillation condition is approached as in a passive

hydrogen maser). In addition, it grows smaller as the atomic resonance narrows.

(vii) In passive clocks, determination of the atomic resonance frequency is sensitive to the spectral symmetry of the frequency modulated excitation signal. Great care is required to ensure this symmetry. Furthermore, in caesium and rubidium clocks, certain frequency shifts depend on the amplitude of the oscillatory field, whose value must therefore be stabilised.

6.4 Caesium beam clocks

The first operational caesium beam atomic frequency standard was built in 1955 by Essen and Parry at the National Physical Laboratory in the United Kingdom [6.10]. Since then, the study and construction of caesium clocks have been undertaken with varying degrees of success in many countries, including Canada, China, France, Italy, Japan, Korea, Poland, East and West Germany, Switzerland, the USSR and the USA.

The aim for laboratory devices is to make the best possible physical realisation of the definition of the second. These devices are *primary time and frequency standards*. For example, the primary standards of the Physikalisch-Technische Bundesanstalt (PTB) in Braunschweig in Germany have been producing the second with an inaccuracy of around 10^{-14} for many years now [6.11, 6.12, 6.13, 6.14]. Moreover, caesium clocks are manufactured commercially and several thousand such instruments are now in operation.

Until 1990, all caesium clocks were built in much the same way as Essen and Parry's original instrument, using in particular magnetic deflection of the atoms. Technological refinements led to very reliable, high performance devices. During this period, the most important innovations were introduced by the PTB.

The development of semiconducting lasers made it possible to apply optical pumping methods to the caesium beam clock, replacing magnetic state selection by optical preparation and detection. The first demonstration was carried out at the Institut d'électronique fondamentale in Orsay, France [6.15]. The optically pumped caesium beam clock was subsequently studied in several laboratories, including the then National Bureau of Standards, now renamed as the National Institute of Standards and Technology (NIST) in Boulder, Colorado, and also at the Laboratoire de l'horloge atomique in Orsay, France [6.16, 6.17]. In 2000, two optically pumped caesium beam laboratory frequency standards reached the quality of primary standards. They are located at the Laboratoire primaire du temps et des fréquences (LPTF) in Paris and NIST [6.18, 6.19].

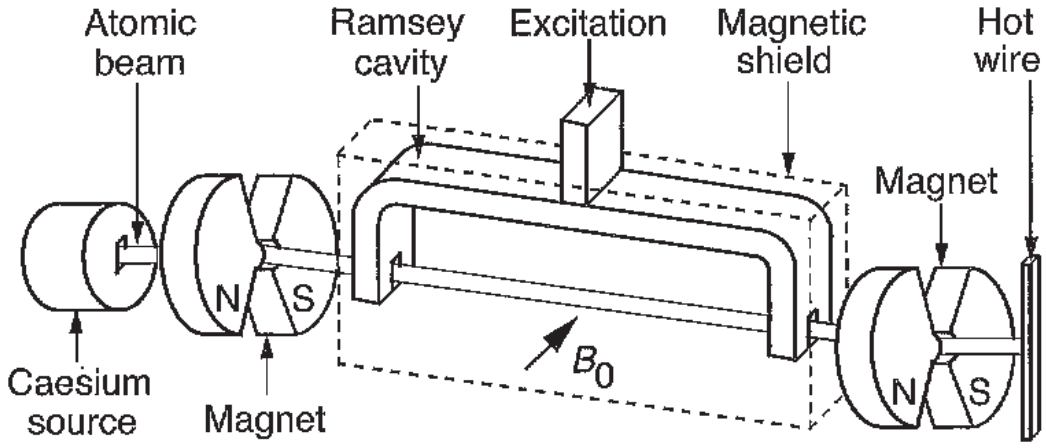


Fig. 6.5. Schematic of a magnetic deflection caesium beam tube. The mass spectrometer, electron multiplier and certain other components are not shown. The deflection angle has been greatly exaggerated.

Over the last ten years or so, laser cooling methods have been studied and developed worldwide in a number of laboratories, including the Ecole normale supérieure (ENS) in Paris [6.20]. Their first application to an experimental primary standard using a fountain of slow caesium atoms, at the LPTF, produced better results than had previously been obtained, with inaccuracy as low as 10^{-15} [6.21, 6.22, 6.23]. It is clear that in the future the best caesium clocks will use laser-cooled atoms.

It is not possible to describe here all variants of the caesium beam frequency standard, nor all the ideas being investigated for its further development. We shall discuss the operating principles and the main features with reference to the commonest model, that is, the commercialised magnetic deflection device. We shall indicate the specific characteristics and performance of laboratory instruments, whose chief objective is to produce a physical realisation of the second that is as accurate as possible. We shall describe optically pumped caesium beam clocks and compare them with magnetic deflection devices. Finally, we shall describe the way atoms can be cooled by laser irradiation, noting that the advantages of using cooled atoms produce quite spectacular performances in atomic frequency standards.

6.4.1 Magnetically deflected caesium beam clock

(a) The caesium beam tube

Figure 6.5 shows schematically the structure of a caesium beam tube with magnetic deflection of atom trajectories. To simplify, we have chosen the common-

est and most basic version, typifying commercialised products presently used for ground and space applications. In laboratory primary standards, magnetic deflection is implemented in a slightly more sophisticated manner.

An *oven* containing several grams of Cs 133, the only stable isotope of this element, is heated to about 100 °C. A system of channels forms a guided beam of caesium atoms in which the sixteen hyperfine sublevels are equally populated. The beam crosses a *first magnet* producing a strong, non-uniform magnetic field. The atoms are subjected to a force perpendicular to the plane of the figure when they pass through this field (see Section 6.2.2). A fraction of the atoms are deflected towards the device axis. This fraction includes one of the $m_F = 0$ states, generally the state $F = 3$, $m_F = 0$. The deflection angle is of the order of 1°. The trajectory of the remaining atoms, including those in the state $F = 4$, $m_F = 0$, is tilted relative to the axis. It has not been shown in Figure 6.5. The corresponding atoms eventually strike walls which absorb caesium.

The beam proceeds to cross the *microwave interaction zone*, which includes a two-branch resonant cavity. Atoms are successively subjected to two ultrahigh frequency magnetic fields, oscillating at 9.192... GHz, separated by a region without variable field. This configuration was devised by Ramsey [6.24]. The distance L between the two oscillatory fields can be freely chosen, and this determines the effective interaction time Δt in (6.8). L is of the order of one metre in laboratory primary standards, whereas it lies closer to 15 cm in commercialised versions. The whole interaction zone is magnetically shielded in order to protect it from the ambient magnetic field and its variations. A homogeneous and stable *magnetic field* is set up there, close to one tenth the strength of the terrestrial magnetic field. (This magnetic field is sometimes called the C-field, whilst the first and second state selecting fields are referred to as magnet A and magnet B, respectively.) Throughout the interaction zone, atoms change hyperfine sublevel, transferring from sublevel $F = 3$, $m_F = 0$ to sublevel $F = 4$, $m_F = 0$ when the frequency of the electromagnetic field injected into the cavity is in resonance with the clock transition.

Figure 6.6 shows in a schematic way the structure of each region of the oscillatory field at the point where the atomic beam crosses the end of one branch of the resonant cavity. The latter is closed by a short circuit, and the ultrahigh frequency magnetic field acting on the atoms is the field at one of the antinodal regions of the standing wave maintained in the cavity. The first order Doppler effect is thereby suppressed (Section 6.2.3), up to a small residual effect (Section 6.4.1e). In the region common to the atomic beam and the cavity arm, the oscillatory magnetic field lies parallel to the static magnetic field, thus satisfying the selection rule for transitions $\Delta F = \pm 1$, $\Delta m_F = 0$. Low-cross section

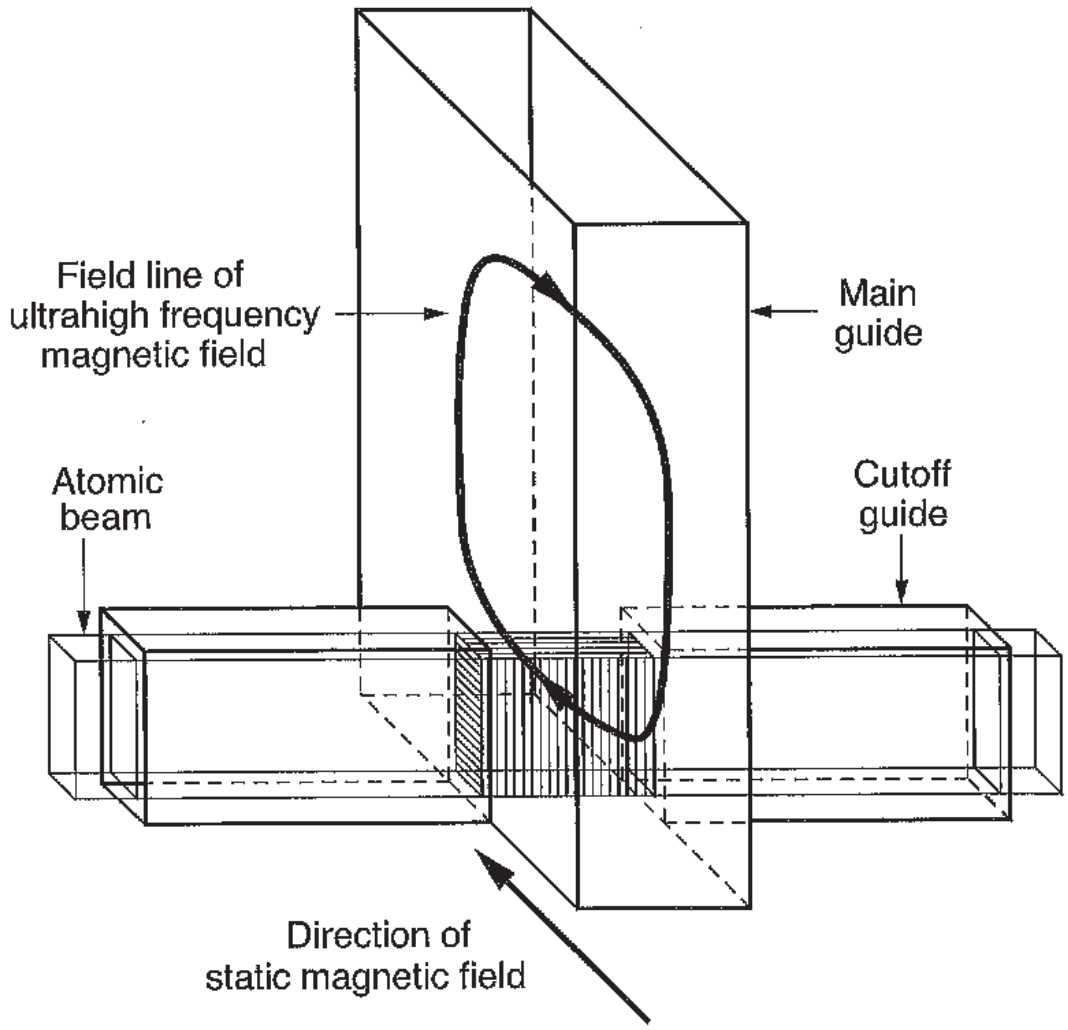


Fig. 6.6. Structure of a region of the oscillatory field. Interaction with atoms takes place in the shaded zone.

wave guides are fixed at the ends of each branch, coaxially with the atomic beam. They prevent the electromagnetic field from radiating out through the openings made in the cavity walls and thereby creating a progressive wave propagating parallel to the atomic beam.

The *second state selection magnet*, similar to the first, deflects the atoms that have made the hyperfine transition towards a *surface ionisation detector*. This consists of a metal wire with work function greater than the ionisation energy of caesium. When the wire is heated to about 1000 °C, incident caesium atoms are evaporated in the form of caesium ions Cs^+. In laboratory caesium clocks, the ionic current is of the order of 1 to 10 pA and can be measured directly by an electrometer. In commercial products, the current is ten times smaller in magnitude. The ion flux is then filtered through a mass spectrometer in order to eliminate unwanted ions emitted by the hot wire. The Cs^+ then reach a dynode

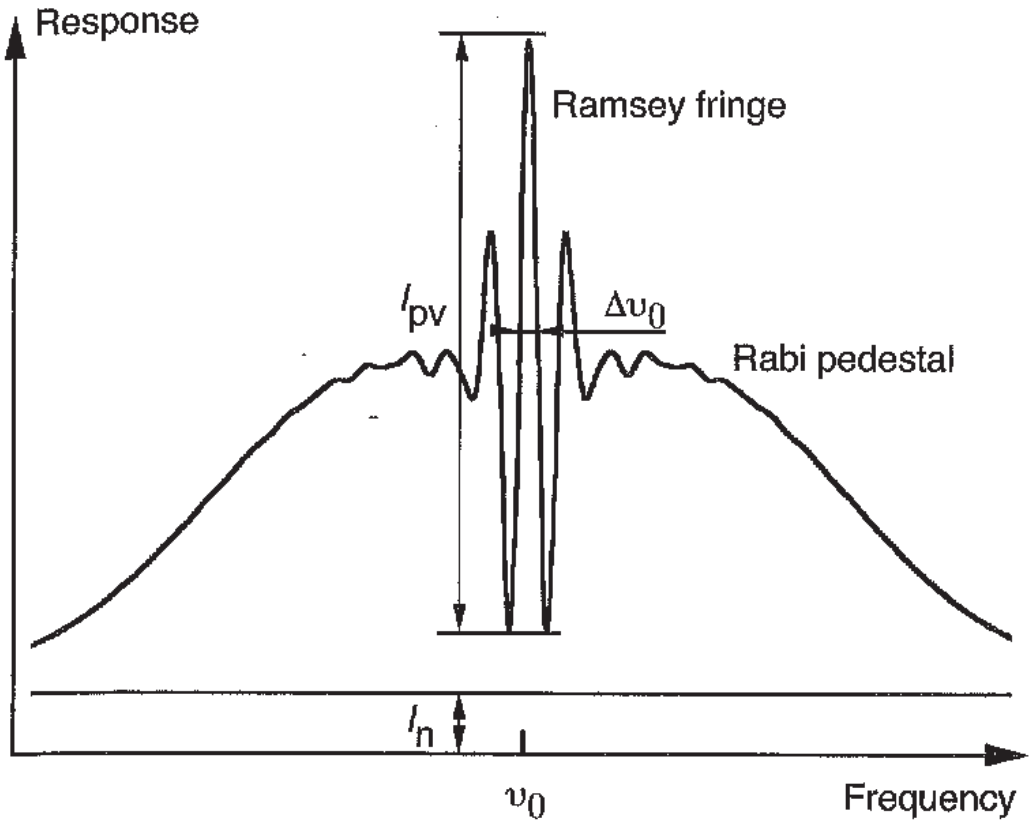


Fig. 6.7. Shape of the resonance curve for a caesium beam tube.

where they cause secondary electrons to be emitted. The flux of these electrons is amplifed by an electron multiplier.

The electric current thus obtained constitutes the output signal of the caesium beam tube. It is proportional to the probability that atoms have undergone the transition $F = 3, m_F = 0 \rightarrow F = 4, m_F = 0$.

(b) Frequency response of a caesium beam tube

The caesium beam tube behaves like a resonator in which the output current depends on the frequency of the applied signal. The general shape of the resonance curve is shown in Figure 6.7. We observe that it has a narrow central part, called the *Ramsey fringe*, situated in the centre of a wider structure, known as the *Rabi pedestal*. (We are considering here what is currently the most common situation, where the two oscillatory fields are in phase. A Ramsey cavity can also be made in which the two oscillatory fields are completely out of phase. In this case, the central fringe is inverted and has a minimum at the resonance. The general appearance of the Rabi pedestal is unchanged.)

The central fringe is the useful part of the resonance curve for servo controlling a quartz oscillator on the atomic transition frequency. Its width is

determined by the time of flight of the atoms between the two oscillatory fields. It is of the order of 500 Hz in commercially manufactured caesium beam tubes. It can be as small as 1 Hz in laboratory instruments using very slow caesium atoms, cooled by laser (Section 6.4.3b). The width at half maximum of the Rabi pedestal depends on the time taken to cross each oscillatory field. It is close to 20 or 25 kHz in the version described above where the atoms effuse from an oven.

Let us begin by giving the simplified algebraic expression for the narrow part of the resonance curve. We shall then justify this simplified expression theoretically and establish the full formula which takes into account both the central part and the wings of the resonance curve. The derivation calls for basic knowledge of quantum mechanics and can be skipped in a first reading.

Formula for the central part of the resonance curve. As we shall show below, the expression for the probability $P(\tau)$ that the clock transition occurs can be simplified if we are only concerned with the immediate vicinity of the resonance frequency. This probability is then given by

$$P(\tau) = \frac{1}{2} \sin^2 b\tau \, [1 + \cos(\omega - \omega_0)T] \, , \tag{6.17}$$

where τ is the time taken by the atom to pass through each oscillatory field and T is the time of flight between these fields. Quantities ω and ω_0 represent the angular frequencies of the oscillatory field and the atomic transition, respectively. The probability $P(\tau)$ depends on the amplitude of the oscillatory field through the parameter b. The latter is related to the ultrahigh frequency magnetic field B' by

$$b = \frac{\mu_B B'}{\hbar} \, , \tag{6.18}$$

where μ_B is the Bohr magneton. The quantity b, often called the Rabi frequency, is an angular frequency. It characterises the rate at which the oscillatory field causes the quantum state of the atom to evolve.

Equation (6.17) shows that the probability of the hyperfine transition occurring in the caesium beam tube has a maximum when $\omega = \omega_0$ and that the width at half maximum of the resonance curve is, for a monokinetic beam,

$$\Delta\nu_0 = \frac{1}{2T} \, , \tag{6.19}$$

in agreement with the uncertainty relation (6.8). Moreover, the term $\sin^2 b\tau$ shows that this probability is optimised when the amplitude of the oscillatory

field satisfies $b\tau = \pi/2$, still assuming that all atoms in the beam have the same velocity.

The simplified expression (6.17) is valid provided that the inequality

$$|\omega - \omega_0| \ll b \tag{6.20}$$

is satisfied. This can also be written

$$\frac{|\nu - \nu_0|}{\Delta\nu_0} \ll \frac{L}{l}\frac{b\tau}{\pi}\,, \tag{6.21}$$

where l and L denote the lengths of each of the oscillatory fields and their separation, respectively. Since, in practice we generally have $b\tau \sim \pi/2$ and $L/l > 10$, the approximation is adequate to describe the central fringe of the resonance curve. Furthermore, it is useful for a simplified analysis of the properties of the caesium beam resonator.

In reality, the atomic velocities are distributed about a mean value. In the case of a magnetically deflected caesium beam tube, it differs from the velocity distribution produced by the oven. This is because the angle of deflection in the two state selection magnets depends sensitively on the velocity. Let $f(\tau)$ be the normed distribution of interaction times. The simplified expression for the average probability that an atom undergoes transition is thus equal to

$$P = \frac{1}{2}\int_0^\infty f(\tau)\,\sin^2 b\tau\,[1 + \cos(\omega - \omega_0)T]\,\mathrm{d}\tau\,. \tag{6.22}$$

The velocity distribution introduces dispersion in the position of the various maxima of the function $\cos(\omega - \omega_0)T$, except when $\omega = \omega_0$. The dispersion is increased as we move away from the resonance frequency, and this reduces the amplitude of the lateral lines. For $|\nu - \nu_0|$ several times greater than the width of the central line, the response of the caesium beam tube exhibits a monotonically varying profile, symmetrical about ν_0. In order to interpret the shape of the Rabi pedestal, we must consider the non-simplified expression for $P(\tau)$, to be established in the next section.

The current I output from the tube is a linear function of P. It may also contain a contribution I_n, due to unwanted caesium atoms, for example. We thus have, near the resonance,

$$I = I_n + \frac{I_0}{2}\int_0^\infty f(\tau)\,\sin^2 b\tau\,[1 + \cos(\omega - \omega_0)T]\,\mathrm{d}\tau\,, \tag{6.23}$$

where I_0 is a constant.

Calculating the response curve of a caesium beam tube. The interaction between atoms and the electromagnetic field can only be rigorously treated

within the framework of quantum mechanics [6.25]. We shall use the basic elements of this theory to establish the main intermediate steps in the calculation.

The quantum state of the atom is characterised by the state vector $|\Psi\rangle$, in the Dirac notation, belonging to the state space of the system. Let $|\Psi_1\rangle$ and $|\Psi_2\rangle$ be state vectors forming an orthonormal basis for this space. They allow us to represent the quantum state of the atom and its time development. Putting

$$|\Psi\rangle = c_1|\Psi_1\rangle + c_2|\Psi_2\rangle \,, \tag{6.24}$$

where c_1 and c_2 are probability amplitudes for states $|\Psi_1\rangle$ and $|\Psi_2\rangle$ to be occupied, respectively. Both depend on time, in general.

The basis vectors $|\Psi_1\rangle$ and $|\Psi_2\rangle$ are chosen to be eigenvectors of the unperturbed Hamiltonian operator. The energies associated with eigenvectors $|\Psi_1\rangle$ and $|\Psi_2\rangle$ are E_1 and E_2, respectively, with $E_2 > E_1$. These energy values are the eigenvalues of the unperturbed Hamiltonian operator under consideration. The matrix $\mathcal{H}_0$ representing this operator is therefore diagonalised in this basis, with matrix elements E_1 and E_2 on the diagonal (and other entries zero). Since the energy is only defined up to an additive constant, we add the quantity $-(E_1 + E_2)/2$, so that

$$\mathcal{H}_0 = \frac{\hbar}{2}\begin{pmatrix} -\omega_0 & 0 \\ 0 & \omega_0 \end{pmatrix} , \tag{6.25}$$

where $\omega_0 = (E_2 - E_1)/\hbar$ is the angular transition frequency.

It can be shown that, for the hyperfine transition under consideration, the matrix $\mathcal{H}_1$ representing the perturbation to the Hamiltonian operator is given by

$$\mathcal{H}_1 = \frac{\hbar}{2}\begin{pmatrix} 0 & 2b\cos\omega t \\ 2b\cos\omega t & 0 \end{pmatrix} , \tag{6.26}$$

where the parameter b is defined by (6.18). It is assumed constant along atomic trajectories, in each arm of the resonant cavity, and hence throughout the period τ of interaction with each of the two oscillatory fields.

The matrix $\mathcal{H}$ representing the total Hamiltonian is found immediately by adding together (6.25) and (6.26). The Schrödinger equation

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = \mathcal{H}|\Psi\rangle \tag{6.27}$$

then leads to

$$\left.\begin{aligned} \mathrm{i}\frac{\mathrm{d}c_1}{\mathrm{d}t} &= -\frac{\omega_0}{2}c_1 + \frac{b}{2}\left[\exp(\mathrm{i}\omega t) + \exp(-\mathrm{i}\omega t)\right]c_2 \\ \mathrm{i}\frac{\mathrm{d}c_2}{\mathrm{d}t} &= \frac{b}{2}\left[\exp(-\mathrm{i}\omega t) + \exp(\mathrm{i}\omega t)\right]c_1 + \frac{\omega_0}{2}c_2 \end{aligned}\right\} . \tag{6.28}$$

The atomic state depends on the time of observation t and the time θ elapsed since the perturbation was first applied. Hence, the total derivative with respect to time is

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial}{\partial t} + \frac{\partial}{\partial \theta}$$

and we seek solutions of the following form, in which effects depending on the two variables are separated:

$$\left.\begin{aligned} c_1(t,\ \theta) &= \gamma_1(\theta) \exp \frac{\mathrm{i}\omega t}{2} \\ c_2(t,\ \theta) &= \gamma_2(\theta) \exp -\frac{\mathrm{i}\omega t}{2} \end{aligned}\right\} . \tag{6.29}$$

This leads to

$$\left.\begin{aligned} \mathrm{i}\frac{\partial \gamma_1}{\partial \theta} &= \frac{\Omega_0}{2}\gamma_1 + \frac{b}{2}\gamma_2 \\ \mathrm{i}\frac{\partial \gamma_2}{\partial \theta} &= \frac{b}{2}\gamma_1 - \frac{\Omega_0}{2}\gamma_2 \end{aligned}\right\} , \tag{6.30}$$

where

$$\Omega_0 = \omega - \omega_0 . \tag{6.31}$$

In (6.30) we have dropped terms at angular frequency 2ω, applying a standard approximation. Indeed, it can be shown that they introduce very rapidly varying components with negligible amplitude into the solution. Equations (6.30) therefore form a linear system with constant coefficients. During transit through each region we have to consider, the change in atomic state can thus be expressed in matrix form,

$$\begin{pmatrix} \gamma_1(\text{final}) \\ \gamma_2(\text{final}) \end{pmatrix} = \mathcal{M}^{(k)} \begin{pmatrix} \gamma_1(\text{initial}) \\ \gamma_2(\text{initial}) \end{pmatrix} , \tag{6.32}$$

where k numbers the regions in the order they are crossed. The first and third regions are the first and second oscillatory fields crossed by the atoms, respectively. The second region is the intermediate space without oscillatory field.

We consider the relation between the values of c_2 and γ_2, for example, on leaving the third region. If t_0 is the time of entry into the first region, the atom leaves the third region at time $t_0 + 2\tau + T$. Let $\gamma_2(\tau,\ T,\ \tau)$ be the value of γ_2 obtained after successive transit through the three regions. It is straightforward to show from (6.29) and (6.32) that

$$c_2(t_0 + 2\tau + T,\ 2\tau + T) = \gamma_2(\tau,\ T,\ \tau) \exp -\frac{\mathrm{i}\omega(t_0 + 2\tau + T)}{2} . \tag{6.33}$$

The population of the energy level E_2 is the square of the modulus of c_2. On exiting the third region, it is therefore equal to $|\gamma_2(\tau, T, \tau)|^2$. Consequently, we only need to know the values of γ_2. Likewise for γ_1.

Let us bring out the interference phenomenon that leads to the periodic dependence of the response of a monokinetic caesium beam tube on $\omega - \omega_0$. The matrix $\mathcal{M}^{(2)}$ obtained by putting $b = 0$ into (6.30) is particularly simple. In fact,

$$\mathcal{M}^{(2)} = \begin{pmatrix} \exp -\dfrac{i\Omega_0 T}{2} & 0 \\ 0 & \exp \dfrac{i\Omega_0 T}{2} \end{pmatrix} . \tag{6.34}$$

Regions 1 and 3 are identical and the corresponding matrices equal. For the moment, we are interested in their general form. For $k = 1, 3$,

$$\mathcal{M}^{(k)} = \begin{pmatrix} \Gamma_{11}^{(k)} & \Gamma_{12}^{(k)} \\ \Gamma_{21}^{(k)} & \Gamma_{22}^{(k)} \end{pmatrix} . \tag{6.35}$$

Let us suppose that the atom is in energy state E_1 when it enters the first oscillatory field. We then have $\gamma_1(0) = 1$ and $\gamma_2(0) = 0$. The quantum state of the atom on leaving the third region is given by

$$\begin{pmatrix} \gamma_1(\tau, T, \tau) \\ \gamma_2(\tau, T, \tau) \end{pmatrix} = \mathcal{M}^{(3)} \mathcal{M}^{(2)} \mathcal{M}^{(1)} \begin{pmatrix} 1 \\ 0 \end{pmatrix} . \tag{6.36}$$

In this case the modulus squared of $\gamma_2(\tau, T, \tau)$ is equal to the probability that the transition takes place. Since the time of flight T is proportional to τ, we therefore set

$$P(\tau) = |\gamma_2(\tau, T, \tau)|^2 , \tag{6.37}$$

which yields

$$P(\tau) = \left| \Gamma_{11}^{(1)} \Gamma_{21}^{(3)} \exp -\frac{i\Omega_0 T}{2} + \Gamma_{22}^{(3)} \Gamma_{21}^{(1)} \exp \frac{i\Omega_0 T}{2} \right|^2 . \tag{6.38}$$

In the first term on the right hand side, $\Gamma_{11}^{(1)}$ and $\exp(-i\Omega_0 T/2)$ determine the variation in the probability amplitude for occupation of energy level E_1 in regions 1 and 2, respectively. The factor $\Gamma_{21}^{(3)}$ represents the probability amplitude for a transition from level E_1 to level E_2 in region 3. The first term thus describes the change in quantum properties of an atom starting in level E_1, remaining at this level through regions 1 and 2, then making the transition to level E_2 in region 3. Likewise, the second term represents the change in quantum properties of an atom starting in level E_1, making the transition from level E_1 to level E_2 in region 1 and then remaining at this level in regions 2 and 3.

These comments show that there are two ways of moving from the initial to the final state. The probability amplitude for occupying the final state is the sum of the partial probability amplitudes associated with these two 'paths'. Since the phases of the two probability amplitudes vary differently as functions of Ω_0, along the two 'paths', interference occurs between them. This is represented by the cross term when we calculate the squared modulus of the sum of the two partial amplitudes. The periodic variation of $P(\tau)$ in Ω_0 arises here.

Let us now write down the elements of matrices $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(3)}$. When the assumption (6.20) holds, we may set $\Omega_0 = 0$ in (6.30) to give

$$\mathcal{M}^{(1)} = \mathcal{M}^{(3)} = \begin{pmatrix} \cos\dfrac{b\tau}{2} & -\mathrm{i}\sin\dfrac{b\tau}{2} \\ -\mathrm{i}\sin\dfrac{b\tau}{2} & \cos\dfrac{b\tau}{2} \end{pmatrix} . \tag{6.39}$$

The entries of the above matrix show that the quantity $b\tau$, with units of an angle, characterises the change in probability amplitude for occupation of the levels in each of the oscillatory fields.

The simplified form for the probability that a transition should occur in a monokinetic caesium beam tube follows from (6.37–6.39). The result is

$$P(\tau) = \frac{1}{2}\sin^2 b\tau\,(1 + \cos\Omega_0 T) . \tag{6.40}$$

Finally, for the general case of arbitrary Ω_0 with respect to b, the system of differential equations (6.30) solves to give

$$\mathcal{M}^{(1)} = \mathcal{M}^{(3)} = \begin{pmatrix} \cos\dfrac{\Omega\tau}{2} - \mathrm{i}\dfrac{\Omega_0}{\Omega}\sin\dfrac{\Omega\tau}{2} & -\dfrac{\mathrm{i}b}{\Omega}\sin\dfrac{\Omega\tau}{2} \\ -\dfrac{\mathrm{i}b}{\Omega}\sin\dfrac{\Omega\tau}{2} & \cos\dfrac{\Omega\tau}{2} + \mathrm{i}\dfrac{\Omega_0}{\Omega}\sin\dfrac{\Omega\tau}{2} \end{pmatrix} , \tag{6.41}$$

where

$$\Omega^2 = \Omega_0^2 + b^2 . \tag{6.42}$$

The general expression for the probability $P(\tau)$ is then [6.6]

$$P(\tau) = 4\frac{b^2}{\Omega^2}\sin^2\frac{\Omega\tau}{2}\left(\cos\frac{\Omega\tau}{2}\cos\frac{\Omega_0 T}{2} - \frac{\Omega_0}{\Omega}\sin\frac{\Omega\tau}{2}\sin\frac{\Omega_0 T}{2}\right)^2 . \tag{6.43}$$

For a given atomic velocity distribution, the shape of the resonance curve is found by multiplying $P(\tau)$ by the interaction time distribution function $f(\tau)$ and then integrating over all possible values of τ.

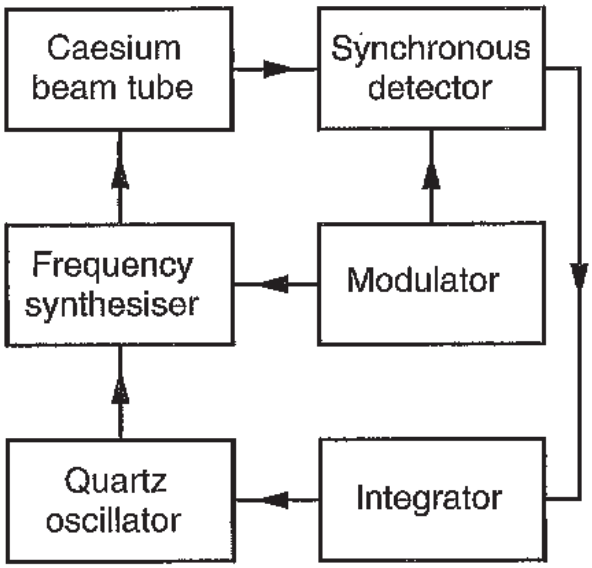


Fig. 6.8. Block diagram for the electronic system servo controlling a quartz oscillator on the frequency of an atomic resonance.

(c) Controlling the frequency of a quartz oscillator by means of an atomic resonance

Since the caesium beam tube is a resonator, the caesium beam clock is a passive atomic frequency standard. The signal exciting the atomic resonance must therefore be supplied. In addition, it should be noted that, because the resonance curve has an extremum at the resonance frequency, it is difficult to point precisely at that frequency. We must obtain a signal that is proportional to the offset between the applied frequency and the reference frequency. It is this signal that makes it possible to lock the frequency of a quartz oscillator onto that of an atomic transition.

The ultrahigh frequency signal, at 9.192...GHz, supplying the resonant cavity, is generated by frequency synthesis from a quartz oscillator whose frequency is often 10 MHz. The synthesiser can modulate the frequency produced about the value $\nu_0 = \nu_{Cs} + \Delta\nu_Z$, where ν_{Cs} is the hyperfine transition frequency of the ground state caesium atom, i.e., 9 192 631 770 Hz, and $\Delta\nu_Z$ is the frequency shift due to the second order Zeeman effect, close to 1.5 Hz. The value of $\Delta\nu_Z$ is related by (6.6) to the magnetic field applied to the atoms. In laboratory primary standards, the frequency of the interrogation signal is modulated around the value $\nu_0 = \nu_{Cs} + \Delta\nu_Z + \Delta\nu'$, where $\Delta\nu'$ is the algebraic sum of frequency shifts other than the shift due to the second order Zeeman effect (Section 6.4.1e).

The resonance frequency is sought via a quite standard method, as shown schematically in Figure 6.8. The frequency of the ultrahigh frequency signal is

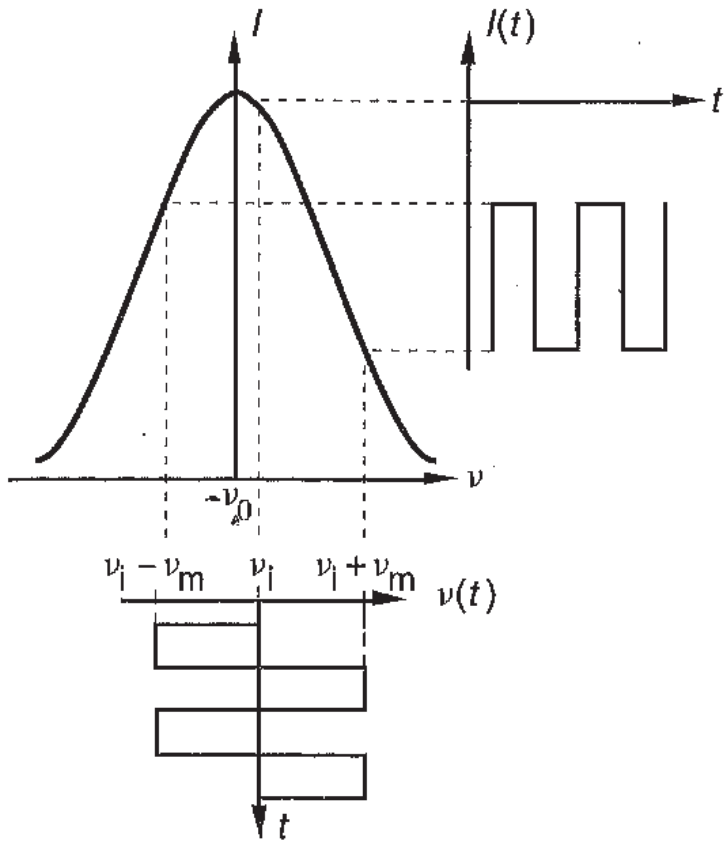


Fig. 6.9. Resonator response to an excitation whose frequency is modulated by a square wave. It is assumed that $\nu_i \neq \nu_0$.

periodically modulated according to

$$\nu(t) = \nu_i + \nu_m g(t) \,, \tag{6.44}$$

where the modulation depth is represented by ν_m, whilst the mean value of the applied frequency is ν_i. The function $g(t)$ is periodic with frequency ν_M and mean value zero representing the shape of the modulation wave. We shall take $g(t)$ to be the square wave function with value $+1$ over one half of the modulation period, and -1 over the other half.

Transient phenomena occurring when the applied frequency changes value will not be taken into account. They are short lived with fundamental frequency equal to $2\nu_M$ and their effects can be eliminated when the caesium beam tube response is processed. Figure 6.9 illustrates the resonator response to this kind of frequency modulation, when $\nu_i \neq \nu_0$.

The variable component in the response of the caesium beam tube to the modulation is proportional to $\nu_i - \nu_0$ when this detuning is small compared with the resonance width. It vanishes when $\nu_i - \nu_0$ is zero and its phase changes by π when the detuning changes sign. This component is amplified and demodulated synchronously with the modulation signal. To this end, it is multiplied by the function $g(t)$. The mean value of the signal thereby obtained constitutes

the *error signal* E. Using the simplified expression (6.23) for the shape of the resonance curve and assuming that $\nu_i - \nu_0$ is much smaller than the resonance width, we obtain

$$E = K I_0(\nu_i - \nu_0) \int_0^\infty T f(\tau) \sin^2 b\tau \, \sin(2\pi \nu_m T) \, d\tau \, , \tag{6.45}$$

where K is a constant. The error signal is zero when $\nu_i = \nu_0$, and it is proportional to the offset between the excitation frequency ν_i and the resonance frequency ν_0, if this difference is small enough. The slope of the error signal is proportional to the derivative $\partial P/\partial\omega$ of the probability P defined by (6.22), for $\omega - \omega_0 = \pm\omega_m$. It depends on the transit time T. It increases as the width of the resonance falls, in other words, as the quality factor increases. The slope also depends on the distribution of interaction times $f(\tau)$, the amplitude of the ultrahigh frequency signal represented by b, and the depth ν_m of the frequency modulation. Optimal values for these parameters are such that $b\tau \sim \pi/2$ and $\nu_m \sim 1/4T$. The latter condition implies that the modulation depth must be chosen close to the half-width of the resonance curve. The absolute value of the slope $\partial P/\partial\omega$ of the resonance curve reaches its maximum here.

A frequency discriminator is therefore set up by modulating the excitation frequency and carrying out synchronous detection of the response from the caesium beam tube. It is then a straightforward matter to servo control the frequency of the quartz oscillator in such a way as to impose $\nu_i = \nu_0$. Its frequency is related to that of the atomic resonance by a constant of proportionality. The latter, denoted $1/K_s$, is some rational fraction depending only on the architecture of the frequency synthesiser included in the servo loop.

The response of the atomic resonator is processed using digital electronic methods. However, the main features of the frequency locking can be described with the help of the equivalent analogue functional diagram shown in Figure 6.10. The notation ΔV or $\Delta V(p)$, for example, denotes the Laplace transform of the deviation of the quantity $v(t)$ from its equilibrium value. The error signal delivered by the synchronous detector can be written

$$\Delta E(p) = K_d \left[\Delta\Omega_r(p) - K_s \Delta\Omega_{sc}(p)\right] \, , \tag{6.46}$$

where K_d is a constant, $\Delta\Omega_r(p)$ represents changes in the angular frequency of the reference standard, and $\Delta\Omega_{sc}(p)$ represents the same for the quartz oscillator when it is servo controlled. If $G(p)$ is the gain of the operational filter, variations in its output voltage are given by

$$\Delta V(p) = G(p) \Delta E(p) \, . \tag{6.47}$$

This voltage is applied to the varactor which serves to adjust the quartz oscil-

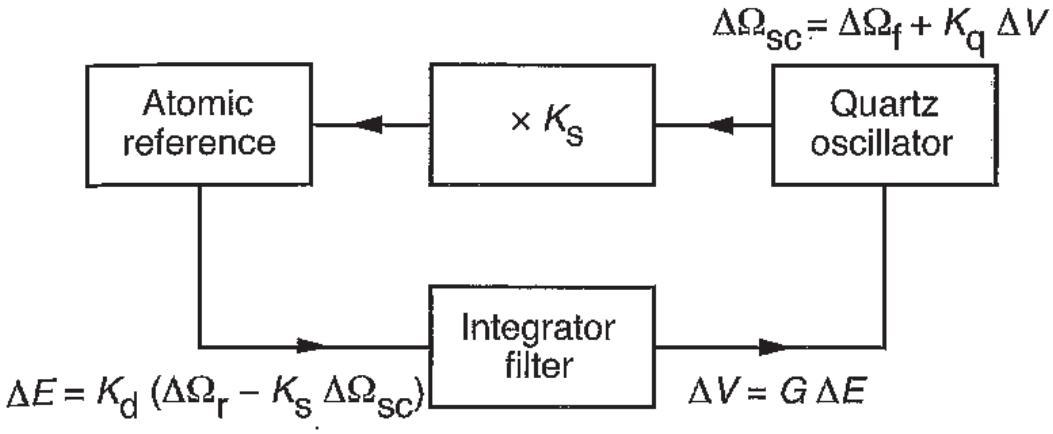


Fig. 6.10. Servo controlling the quartz oscillator frequency by means of a passive atomic reference: functional diagram.

lator frequency and we have

$$\Delta\Omega_{sc}(p) = \Delta\Omega_f(p) + K_q \Delta V(p)\ , \tag{6.48}$$

where K_q is a constant and $\Delta\Omega_f$ represents the frequency variation of the quartz oscillator when it is free, i.e., when the servo loop is broken.

Restricting to the simple case where the operational filter reduces to an amplifier followed by a pure integrator, we have

$$G(p) = \frac{A}{RCp}\ , \tag{6.49}$$

where A is a constant and RC is the time constant of the integrator circuit. We then obtain, from the above equations,

$$\Delta\Omega_{sc}(p) = \frac{\mathcal{T}p}{1+\mathcal{T}p}\Delta\Omega_f(p) + \frac{1}{1+\mathcal{T}p}\frac{\Delta\Omega_r(p)}{K_s}\ , \tag{6.50}$$

where $\mathcal{T}$ is the time constant of the response of the servo loop to a perturbation. It is equal to

$$\mathcal{T} = \frac{RC}{AK_dK_sK_q}\ . \tag{6.51}$$

In practice, it has a value between one and a few seconds.

If we now consider random frequency fluctuations, the spectral density S_y of their relative fluctuations (Section 5.2.4) is given by

$$S_{y,\,sc}(f) = \frac{(2\pi f\mathcal{T})^2}{1+(2\pi f\mathcal{T})^2}S_{y,\,f}(f) + \frac{1}{1+(2\pi f\mathcal{T})^2}S_{y,\,r}(f)\ , \tag{6.52}$$

where f is the Fourier frequency and subscripts sc and f mean the same as before.

The transfer functions appearing on the right hand side of (6.50) are those of

a high-pass filter for the first and a low-pass filter for the second. The cutoff frequency of these filters is equal to $1/2\pi\mathcal{T}$. The nature of the transfer functions is conserved when the structure of the operational filter is more elaborate.

Variations in the frequency of the servo controlled oscillator involve two contributions, as can be seen from (6.50) and (6.52): the high frequency components of the frequency variations of the free quartz oscillator and the low frequency components of the frequency variations of the reference. This implies that:

- in the short term, for observation times τ shorter than $\mathcal{T}$, the servo controlled oscillator exhibits the same frequency fluctuations that it would if it were free;
- in the medium and long term, for $\tau > \mathcal{T}$, the servo controlled quartz oscillator reproduces the frequency variations of the reference, up to the constant K_s.

It follows from these observations that two conditions must be satisfied if the caesium clock is to have good frequency stability. Firstly, the quartz oscillator must be of the very best quality. (There are other reasons for requiring this, which we shall not go into here.) Secondly, variations in the atomic frequency must clearly be as small as possible.

(d) Medium term frequency instability

In the following, we shall be particularly concerned with frequency variations related to the properties of the caesium beam tube. In this section, we consider noise associated with detection of the atomic beam. This determines the frequency instability of the standard for sampling periods τ of a few seconds to several days.

The atoms in a beam and the charge carriers in an electric current move through cross sections of the beam or the conductor, respectively, at randomly distributed times. The relevant distribution is a Poisson distribution. It follows that the current output from the caesium beam tube fluctuates about its mean value. This fluctuation is called *shot noise*. For a constant value ν_i of the incident frequency, defined by (6.44), the servo loop interprets this fluctuation as being due to a variation in the resonance frequency of the caesium beam tube, and hence in the reference frequency.

In order to identify the key parameters influencing this frequency variation, let us further simplify expression (6.23) for the output current. In practice, the shape of the central fringe is close to sinusoidal and we may write, to a

sufficient approximation for our present purposes,

$$I = I_n + \frac{1}{2} I_{pv} \left[1 + \frac{\cos \pi (\nu - \nu_0)}{\Delta \nu_0} \right] , \tag{6.53}$$

where I_{pv} and $\Delta\nu_0$ represent the peak-to-valley height of the resonance curve and its width at half maximum, as shown in Figure 6.7.

Consider a half period of modulation, during which the frequency applied to the atoms is $\nu = \nu_0 + \nu_m$. We would obtain the same final result by considering the other half period of modulation. The current fluctuation δI resulting from small fluctuations δI_n and δI_{pv} in I_n and I_{pv} is given by

$$\delta I = \delta I_n + \frac{1}{2} \delta I_{pv} \left[1 + \cos \frac{\pi \nu_m}{\Delta \nu_0} \right] . \tag{6.54}$$

If the currents I_n and I_{pv} were constant, but the resonance frequency were variable, equal to $\nu_0 + \delta\nu_0$ with $\delta\nu_0 \ll \Delta\nu_0$, we would have

$$\delta I = \frac{1}{2} I_{pv} \left[\cos \frac{\pi \nu_m}{\Delta \nu_0} + \frac{\pi \delta \nu_0}{\Delta \nu_0} \sin \frac{\pi \nu_m}{\Delta \nu_0} \right] . \tag{6.55}$$

Equating the two expressions for δI and assuming that the modulation depth is equal to the half-width of the line, i.e., that $\nu_m = \Delta\nu_0/2$, we obtain the following formula for the apparent relative variation in the resonance frequency, induced by fluctuations in I_n and I_{pv}:

$$\frac{\delta \nu_0}{\nu_0} = \frac{1}{\pi} \frac{2\delta I_n + \delta I_{pv}}{I_{pv} Q_{at}} , \tag{6.56}$$

where $Q_{at} = \nu_0/\Delta\nu_0$ is the quality factor of the atomic resonance.

When a flux of particles, with average value λ, satisfies a Poisson distribution, the standard deviation $\sigma'_{\delta\lambda}$ of the fluctuation $\delta\lambda$ in this flux is such that

$$\sigma'_{\delta\lambda} = \left(\frac{\lambda}{\tau} \right)^{1/2} , \tag{6.57}$$

where τ is the time over which the fluctuation is observed. The standard deviation, in the ordinary sense, of the relative frequency fluctuations $\delta\nu_0/\nu_0$ is thus equal to

$$\sigma'_{\delta\nu_0/\nu_0} = \frac{1}{\pi} \frac{(2I_n + I_{pv})^{1/2}}{I_{pv} Q_{at}} \tau^{-1/2} . \tag{6.58}$$

Since the relative frequency fluctuations are usually characterised by the Allan deviation, for reasons explained in Sections 5.2.5 and 5.2.6, it is useful to begin by establishing an expression for their power spectral density. Shot noise falls into the category of white noise, with power spectral density equal to

twice the mean value of the particle flux. Since the two sources of shot noise are independent, the power spectral density of the relative frequency fluctuations $y = \delta\nu_0/\nu_0$ is equal to

$$S_y = \frac{2}{\pi^2}\frac{2I_n + I_{pv}}{I_{pv}^2 Q_{at}^2} . \tag{6.59}$$

The Allan deviation of the relative frequency fluctuations of a caesium beam clock follows immediately (Section 5.2.6):

$$\sigma_y(\tau) = \frac{1}{\pi}\frac{(2I_n + I_{pv})^{1/2}}{I_{pv} Q_{at}}\tau^{-1/2} . \tag{6.60}$$

In the special case of white frequency noise, this standard deviation turns out to be equal to the ordinary standard deviation given by (6.58). Note that the numerical factor $1/\pi$ has no real significance given the approximations made along the way.

Equation (6.60) quantifies the harmful effect of the unwanted atomic flux I_n. Every effort is made to reduce it to a negligible value when designing this kind of clock.

Note that, if I_n is negligible compared with I_{pv} and for given τ, the frequency instability $\sigma_y(\tau)$ varies as the quantity

$$q = I_{pv}^{-1/2} Q_{at}^{-1} .$$

The flux of atoms reaching the detector, and hence I_{pv}, varies as D^{-2}, where D is the distance between the oven and the detector, whereas the quality factor of the atomic resonance varies in proportion to the separation L between the two oscillatory fields. The quantity q thus varies as D/L. Now the value of this ratio, greater than 1 but generally less than 2, is approximately the same in all caesium beam tube designs, whether they be laboratory instruments or compact commercialised devices. It follows that the length of a caesium beam tube does not play a key role in the frequency instability of a device when this instability is determined by shot noise in the flux.

Equations (6.59) and (6.60) can be written in more general form. Indeed, since $I_n + I_{pv}/2$ is the output current in the presence of frequency modulation, the quantity $2I_n + I_{pv}$ represents the shot noise power P_N in a noise pass band of 1 Hz. The power of the useful signal is $P_S = I_{pv}^2/4$. The signal-to-noise power ratio for the specified pass band measured at the fringe flank, i.e., at frequency $\nu_0 \pm \nu_m$, is thus equal to

$$\left.\frac{P_S}{P_N}\right|_{1\text{ Hz}} = \frac{I_{pv}^2}{4(2I_n + I_{pv})} . \tag{6.61}$$

We may then write

$$S_y = \frac{1}{2\pi^2} \frac{1}{Q_{at}^2 (P_S/P_N)_{1\,Hz}} . \tag{6.62}$$

The Allan deviation becomes

$$\sigma_y(\tau) = \frac{1}{2\pi} \frac{1}{Q_{at} (P_S/P_N)_{1\,Hz}^{1/2}} \tau^{-1/2} . \tag{6.63}$$

The last two relations show that the relative frequency instability depends on the relative accuracy with which the signal-to-noise ratio allows us to aim at the atomic resonance. For a given τ, the frequency instability decreases as the quality factor and signal-to-noise ratio increase. This conclusion is valid for all passive clocks.

In commercial designs, the width of the atomic resonance is close to 500 Hz and the signal-to-noise power ratio defined above is of the order of 10^5 to 10^6. We thus have $\sigma_y(\tau) \sim 10^{-11}\tau^{-1/2}$, the exact value depending on the model considered. The commercial model with best performance has a frequency instability in the neighbourhood of $5 \times 10^{-12}\tau^{-1/2}$, quite comparable with the best caesium beam laboratory primary standard based on magnetic beam deflection. (The inaccuracy of the latter is about a hundred times smaller.)

Figure 6.11 shows typical frequency instabilities of caesium clocks, among others.

In the long term, for τ greater than a few days, the value of $\sigma_y(\tau)$ no longer decreases. The frequency instability is then no longer limited by shot noise, but by variations in the frequency shifts of the atomic resonance.

(e) Residual shifts in clock frequency

A certain number of physical effects and instrumental imperfections produce a slight change in the hyperfine transition frequency of the ground state caesium atom, as it is observed in the caesium beam tube [6.4]. We analyse the origins of the most significant of these frequency shifts, indicating the order of magnitude of their relative value. We restrict the discussion to frequency shifts specific to the frequency standard under consideration. The effects of gravitational potential and instrumental motion are specified in Chapter 3.

Any variation in these frequency shifts will cause long term frequency instability. In addition, uncertainty in their values determines the inaccuracy of the clock.

In order to understand the aims behind attempts to make these frequency shifts as small as possible, it is essential to have a clear idea of their origins.

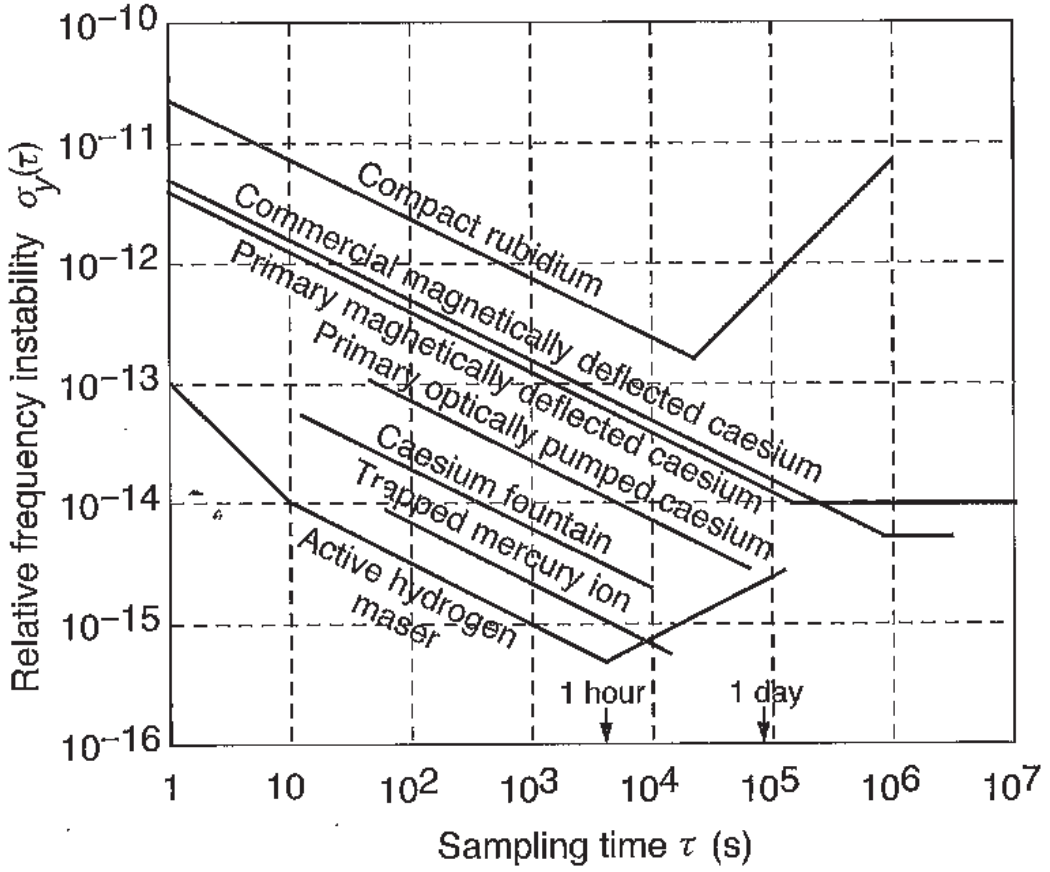


Fig. 6.11. Relative frequency instability characterised by the Allan deviation $\sigma_y(\tau)$ for various frequency standards. Curves give asymptotic values. The figure shows frequency instabilities as observed in 1997. By 2000, the frequency instabilities of two of the frequency standards considered here had been reduced by a factor of about 2.5 in the case of the optically pumped caesium beam primary standard and about 3.5 for the caesium fountain.

(1) Frequency shifts of physical origin

Second order Zeeman effect. Seven transitions $\Delta F = \pm 1$, $\Delta m_F = 0$ are observed when the ultrahigh frequency magnetic field lies parallel to the static magnetic field B_0, as can be seen from the measurements recorded in Figure 6.12 [6.26]. The central resonance, due to the transition

$$F = 3,\ m_F = 0 \quad \longrightarrow \quad F = 4,\ m_F = 0\ ,$$

undergoes a quadratic frequency shift $\Delta\nu_Z$ given by (6.6). For the typical value $B_0 = 6 \times 10^{-6}$ T, we have $\Delta\nu_Z \simeq 1.5388$ Hz, or a relative value of about 1.7×10^{-10}. This is the largest of the frequency shifts affecting the reference transition. The final accuracy of the standard is critically dependent on the way this effect is determined.

In practice, we use the atomic properties themselves and the value of B_0 is deduced from the frequency shift in one or several of the six transitions

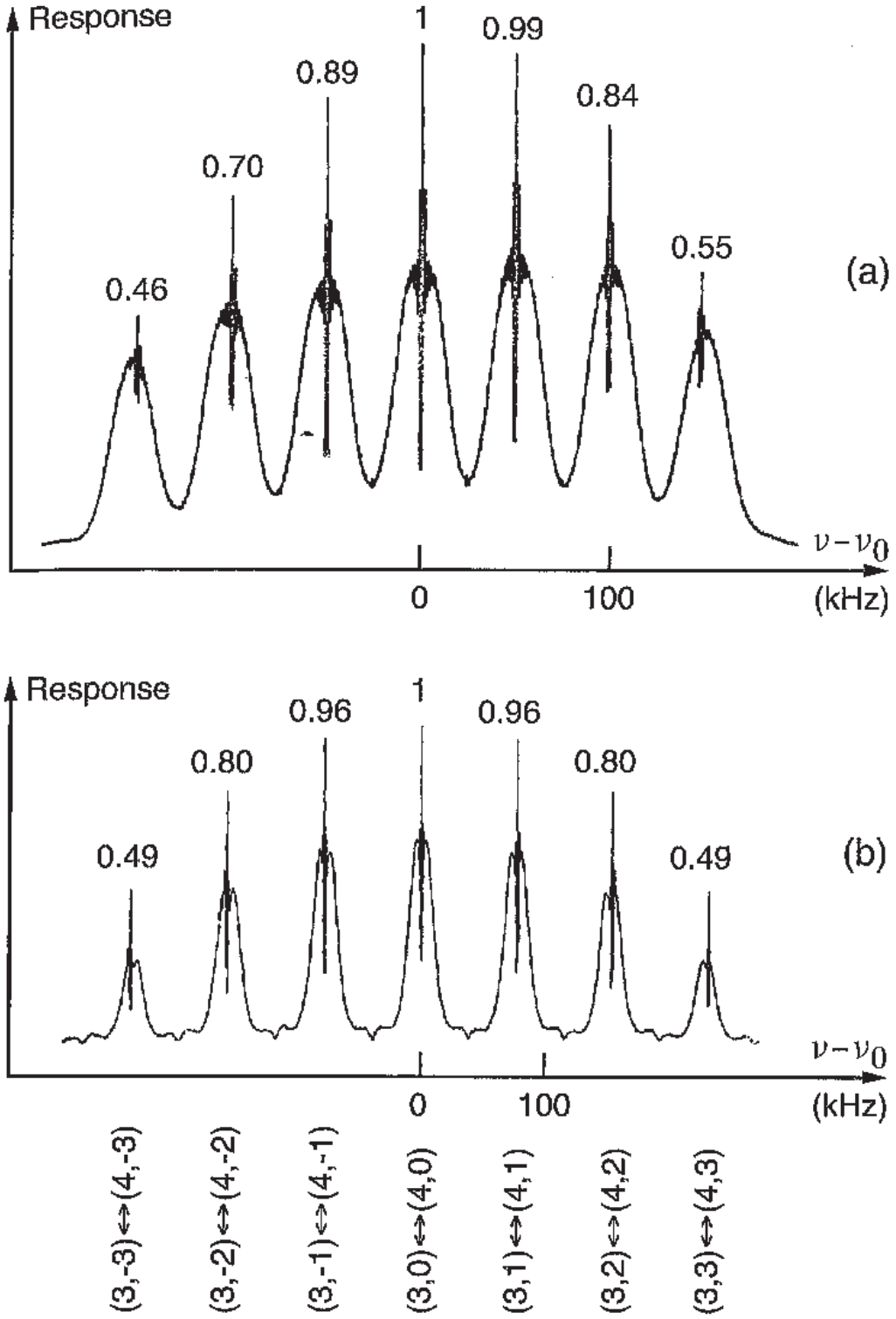


Fig. 6.12. Microwave resonances in a caesium beam tube. (a) Magnetic deflection tube, (b) optically pumped tube. Heights of lateral lines are given relative to the height of the central line. The two cavities have different structures and the values of B_0 are different. (Taken from [6.26].)

$(3, m_F) \leftrightarrow (4, m_F)$ for which $\Delta m_F = 0$. The Breit–Rabi equation (6.4) can be used to express this change as a function of B_0 and the atomic constants. Expanding out the expression, the first term is proportional to B_0. In commercially produced clocks, it is enough to know this term for the present

levels of accuracy. In caesium beam primary standards, the quadratic term in the expansion is taken into account.

In modern commercial devices, the value of B_0 is actively controlled. To this end, the mean value of the applied frequency ν_i defined by (6.44) is switched for a fraction of a second from the value $\nu_{Cs} + \Delta\nu_Z$ to another predetermined value $\nu_{Cs} + \Delta\nu'_Z$, where ν_{Cs} is the hyperfine transition frequency unaffected by the Zeeman effect. B_0 is adjusted in such a way that the transition

$$F = 3,\ m_F = 1 \quad \longleftrightarrow \quad F = 4,\ m_F = 1$$

resonates at the frequency $\nu_{Cs} + \Delta\nu'_Z$. We then have, to a sufficient approximation, $B_0 = \Delta\nu'_Z/K$, where the factor K depends only on atomic constants. It has a value close to 7×10^9 Hz T^{-1}. This adjustment is made automatically, once or several times a minute. The value of the applied magnetic field is thus servo controlled at a constant value [6.27]. In this way, the effects of varying ambient magnetic field are eliminated.

When the applied magnetic field is inhomogeneous, the measurement gives the mean value $\langle B_0 \rangle$ along the beam, whilst the frequency shift due to the second order Zeeman effect is proportional to $\langle B_0^2 \rangle$, the mean value of B_0^2. In general, $\langle B_0 \rangle^2 \neq \langle B_0^2 \rangle$. Here then is a source of error that must be minimised by producing a magnetic field as uniform as possible within the magnetic shielding.

Second order Stark effect. This shift in the hyperfine transition frequency of the ground state alkalis is a quadratic function of the electric field. In a caesium beam tube, this field is certainly extremely weak because the magnetic shielding also functions as a Faraday cylinder. A value of 1 V cm^{-1}, which is an overestimate, would cause a relative frequency shift of 2.5×10^{-16}. This is negligible.

Effect of blackbody radiation. Inside a closed container, the electromagnetic field in equilibrium with a blackbody at the ambient temperature exhibits a very broad spectrum, with maximum at around 30 THz ($\lambda = 10$ μm). At 300 K, the root mean squared values of the electric and magnetic fields are 832 V m^{-1} and 2.77×10^{-6} T, respectively.

An alternating electric field has a similar effect to a direct electric field. For the caesium atom, the relative value of the corresponding frequency shift $\Delta\nu_c$ is given by

$$\frac{\Delta\nu_c}{\nu_{Cs}} = -1.69 \times 10^{-14} \left(\frac{T}{300}\right)^4 , \tag{6.64}$$

where T is the thermodynamic temperature. This frequency shift must be taken into account in laboratory caesium beam standards.

The alternating magnetic field also causes a shift in the transition frequency of the ground state hyperfine structure. It can be shown that it has the same value for all the atoms considered here. This frequency shift is close to 10^{-17} in relative value and is currently negligible.

Effect of spin-exchange collisions. When two alkali atoms collide, the outermost electrons may swap over. Such a collision is called a spin-exchange collision. It induces a shift in the ground state hyperfine transition frequency, proportional to the atomic density, the average relative velocity of the atoms and the effective cross section of the spin-exchange frequency shift. In 2000, there was still no theoretical estimate for this cross section in a caesium beam where atomic speeds have the same order of magnitude as thermal excitation speeds around room temperature. The calculation is complicated by the great number of hyperfine sublevels. At the present time, no experimental investigation has revealed evidence for such a frequency shift in a caesium beam clock with beam generated in an oven at around 100 °C. However, in caesium clocks using cooled atoms, it is measurable and must therefore be taken into account (Section 6.4.3b).

Second order Doppler effect. The atoms are moving relative to the device probing their resonance frequency. They therefore manifest relativistic time dilation (Chapter 3), and this leads to a shift in the transition frequency due to the second order Doppler effect. When the atoms all have the same speed v, this frequency shift is given by (6.16).

The shape of the resonance curve is then described by the following equation, deduced from (6.23):

$$I = \frac{1}{2} I_0 \int_0^\infty f(\tau) \sin^2 b\tau \left\{ 1 + \cos \left[\omega - \omega_0 \left(1 - \frac{v^2}{2c^2} \right) \right] T \right\} \mathrm{d}\tau \,, \quad (6.65)$$

where the term I_n, assumed independent of the frequency, has been dropped.

The observed resonance curve is found by superposing the elementary resonance curves associated with given values of the speed v. As the speed increases, curves are shifted to lower frequencies. Their weighting depends on the speed via the factor $f(\tau) \sin^2 b\tau$, since τ is inversely proportional to the speed. It follows that the total resonance curve exhibits a very slight asymmetry.

In the presence of a square wave frequency modulation of depth ν_m, it can be shown that the apparent resonance frequency, i.e., the frequency at which the error signal goes to zero (Section 6.4.1c), is given by

$$\frac{\Delta\nu_D}{\nu_0} = -\frac{\int_0^\infty v f(\tau) \sin^2 b\tau \ \sin(2\pi\nu_m T)\, d\tau}{2c^2 \int_0^\infty \frac{1}{v} f(\tau) \sin^2 b\tau \ \sin^2(2\pi\nu_m T)\, d\tau} . \qquad (6.66)$$

This equation shows that the frequency shift depends on:

- ν_m, as a consequence of the asymmetry in the resonance curve;
- b, and hence the amplitude of the ultrahigh frequency excitation.

The value of ν_m is fixed by the design of the associated electronics, and it is insensitive to environmental conditions. In contrast, the value of b may vary when these conditions change. In modern caesium clocks, the value of b is held constant by means of an appropriate feedback system, to be described shortly.

The frequency shift due to the second order Doppler effect is of the order of 10^{-13} in the caesium clocks we are considering here. In general, its precise value depends on the distribution of interaction times τ, among other things, and hence on the atomic velocity distribution, as well as the value of b. However, when the velocity distribution is narrow, (6.66) tends to $-v^2/2c^2$ and the frequency shift becomes insensitive to the values of b and ν_m. It is therefore preferable to use atoms with as small a range of velocities as possible, and furthermore, with velocities as small as possible. We shall return to this point in Section 6.4.3.

(2) Frequency shifts of instrumental origin

Effect of phase difference between oscillatory fields. In a pure stationary electromagnetic wave, the phase of the magnetic field varies suddenly by π across each of the nodes, whilst remaining constant between them. However, there are losses in the resonant cavity because of the finite conductivity of its walls. There is therefore a progressive wave carrying the energy lost by the Joule effect. This means that the stationary wave is not pure. The phase of the magnetic field still changes rapidly by a value close to π in the vicinity of the nodes, but it varies in an almost linear way in the interval between them.

Figure 6.13 shows schematically how the phase varies in the resonant cavity of a caesium beam clock, between the point where the two arms join and their short-circuited ends. It has been assumed that the two arms have slightly

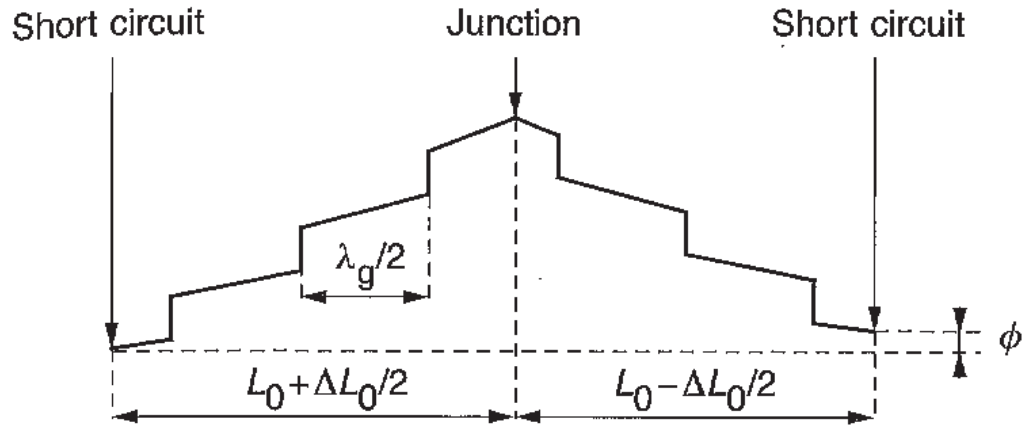


Fig. 6.13. Asymptotic representation of phase variation in the two arms of the resonant cavity. The antinode of the oscillatory field occurs near the short circuits. The phase jumps by π across each node. The difference in length of the two arms and the gradual change in phase are both greatly exaggerated. λ_g is the wavelength in the wave guide.

different lengths. Since the phase is continuous at the join, we observe a phase difference ϕ between the two oscillatory magnetic fields. This phase difference is extremely small, of the order of about ten microradians. It results from the finite conductivity of the materials used and tolerance limits in the construction of the cavity.

The frequency response of a caesium beam tube (Section 6.4.1b) can be recalculated assuming that the phase of the second oscillatory field leads the phase of the first by ϕ. The simplified expression (6.22) for the probability that a hyperfine transition should occur then becomes

$$P(\tau) = \frac{1}{2}\sin^2 b\tau \{1 + \cos[(\omega - \omega_0)T + \phi]\} , \tag{6.67}$$

for a beam in which all atoms have the same speed.

The resonance frequency is shifted by $\Delta\nu_\phi$, where

$$\frac{\Delta\nu_\phi}{\nu_0} = -\frac{\phi}{2\pi\nu_0 T} = -\frac{v}{2\pi\nu_0 L}\phi , \tag{6.68}$$

where T is the time of flight between the two oscillatory fields separated by distance L. If this frequency shift is written in terms of the imbalance ΔL_0 in the electrical lengths of the two cavity arms, we obtain

$$\frac{\Delta\nu_\phi}{\nu_0} = K_\phi \frac{v}{c} , \tag{6.69}$$

with c the speed of light and the quantity K_ϕ given by

$$K_\phi = -\frac{\lambda}{\lambda_g}\frac{L_0}{L}\alpha_g \Delta L_0 , \tag{6.70}$$

where λ and λ_g are the wavelengths of radiation applied to atoms in the vacuum and in the guide, respectively, α_g is the attenuation constant of the ultrahigh frequency magnetic field along the guide, and L_0 is the mean length of the arms. Typical values here are $\lambda/\lambda_g = 0.7$, $L_0/L = 0.5$ and $\alpha_g = 2.6 \times 10^{-2}\ \mathrm{m}^{-1}$. If we assume that $v = 150\ \mathrm{m\,s}^{-1}$ and $\Delta L_0 = 10^{-4}$ m, this yields $|\Delta\nu_\phi/\nu_0| \sim 5 \times 10^{-13}$. Any length difference between the two arms of the resonant cavity must therefore be reduced as far as possible.

Equation (6.69) shows that the effect of the phase difference between the two oscillatory fields can be interpreted as a first order residual Doppler effect, related to the presence of progressive waves in the resonant cavity. The attenuation constant K_ϕ depends on how carefully the cavity has been made. With the above numerical values, K_ϕ is of the order of 10^{-6}.

Since the frequency shift depends on atomic velocities, the existence of a velocity distribution distorts the resonance line, in the same way as the second order Doppler effect.

When the frequency is modulated, the frequency shift is given by

$$\frac{\Delta\nu_\phi}{\nu_0} = -\frac{\phi}{2\pi\nu_0 L}\,\frac{\displaystyle\int_0^\infty f(\tau)\sin^2 b\tau\ \sin(2\pi\nu_m T)\,\mathrm{d}\tau}{\displaystyle\int_0^\infty \frac{1}{v} f(\tau)\sin^2 b\tau\ \sin^2(2\pi\nu_m T)\,\mathrm{d}\tau}\,. \tag{6.71}$$

The general comments made with regard to the second order Doppler effect also apply. When the atomic velocities are distributed over a range of values, the frequency shift depends on the modulation depth ν_m, and also the amplitude of the ultrahigh frequency excitation, represented by b. Furthermore, it is clear that the frequency shift decreases as the atomic speeds decrease.

The sign of the frequency shift changes when the direction of the atomic beam is reversed. This implies that the shift can in principle be measured. Such a process is implemented in most laboratory caesium beam primary frequency standards.

In each interaction region, the phase of the ultrahigh frequency magnetic field is position dependent. This spatial variation is also due to the progressive wave transporting energy lost in the cavity walls. Although it can be minimised by a specific configuration in the end part of each cavity arm, it still exists when the phase difference between the centres of the two interaction regions is zero. The phase difference ϕ considered above must therefore be interpreted as the total phase difference between the two oscillatory fields, i.e., the sum of the phase difference related to the differing lengths of the two arms and the distributed phase difference, the latter being averaged over the atomic paths in each interaction zone. In magnetically deflected caesium clocks, this averaging

can only be estimated rather approximately. Indeed, the deflection of the atoms in the state selecting magnets depends on their velocity. It follows that atomic velocity and path distribution are coupled in a cross section of the beam, and this proves difficult to model precisely.

In the reversed beam method, it is thus important to reproduce the path distribution in the two interaction regions as closely as possible for the two directions of the atomic beam.

Effect of cavity detuning. If the resonance frequency ν_{cav} of the cavity does not exactly equal that of the atomic resonance, ν_0, the amplitude of the ultrahigh frequency magnetic field varies asymmetrically when the frequency is modulated about ν_0. Hence, for a square wave frequency modulation, the value of b is not the same over two successive half-periods. The mean excitation frequency must then be shifted so that, despite this imbalance, the response of the caesium tube assumes two equal values over consecutive half-periods of the frequency modulation. The error signal used to servo control the quartz oscillator frequency is zero under this condition.

Let us suppose that the detuning $\nu_c - \nu_0$ of the resonant cavity is small compared with the width of its own resonance. The apparent shift $\Delta\nu_{cav}$ in the atomic resonance frequency is then given by

$$\frac{\Delta\nu_{cav}}{\nu_0} = \frac{\nu_c - \nu_0}{\nu_0} \frac{2T_c^2 \int_0^\infty b\tau f(\tau) \sin b\tau \cos b\tau (1 + \cos \omega_m T)\, d\tau}{\int_0^\infty T^2 f(\tau) \sin^2 b\tau \frac{\sin \omega_m T}{\omega_m T}\, d\tau}, \quad (6.72)$$

where $\omega_m = 2\pi\nu_m$ and ν_m is the depth of frequency modulation. By the above hypothesis, the amplitude of the ultrahigh frequency magnetic field varies only slightly during modulation, and b represents one of its values. T_c is the time constant of the cavity response, related to its quality factor by

$$T_c = \frac{Q_c}{\pi\nu_c}. \quad (6.73)$$

Concerning orders of magnitude, (6.72) and (6.73) together with the relation $Q_{at} \sim 2T\nu_0$ imply

$$\frac{\Delta\nu_{cav}}{\nu_0} \sim \frac{\nu_c - \nu_0}{\nu_0} K_{cav} \frac{Q_c^2}{Q_{at}^2}, \quad (6.74)$$

where K_{cav} is a constant depending on ν_m, b and the velocity distribution. This equation tells us that, for a given cavity detuning, the apparent shift of the atomic resonance frequency is more or less proportional to the squared

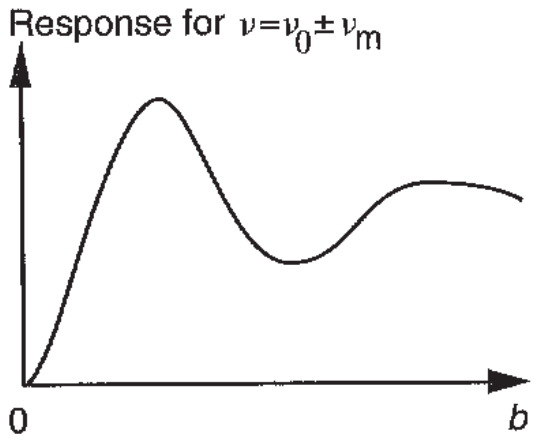


Fig. 6.14. Varying response of the caesium beam tube for different amplitudes of the oscillatory field when its frequency is either $\nu_0 + \nu_m$ or $\nu_0 - \nu_m$.

ratio of the quality factors for the cavity and the atomic resonance. It is clearly desirable to reduce the value of this ratio as much as possible.

When the applied frequency is modulated as indicated, i.e., when we have $\omega - \omega_0 = \pm\omega_m$ in (6.23), the response of the caesium beam tube is given by

$$I(b, \omega_m) = I_n + \frac{I_0}{2}\int_0^\infty f(\tau)\sin^2 b\tau\,(1 + \cos\omega_m T)\,d\tau\,. \tag{6.75}$$

Note that the numerator of the right hand side of (6.72) is proportional to the derivative with respect to b of the tube response, viz., $\partial I(b, \omega_m)/\partial b$. Hence, if the resonant cavity is detuned, the frequency shift disappears when

$$\frac{\partial I(b, \omega_m)}{\partial b} = 0\,. \tag{6.76}$$

This occurs for those values of the amplitude of the ultrahigh frequency excitation where the response of the caesium beam tube is at an extremum. This feature is specific to square wave frequency modulation.

Figure 6.14 shows the kind of variation observed in $I(b, \omega_m)$ for changing amplitude of the ultrahigh frequency excitation. The operating point chosen corresponds to the first maximum of the tube response.

In modern caesium beam frequency standards, an electronic servo control seeks the amplitude of the ultrahigh frequency excitation at which the resonator response has a maximum in the presence of square wave frequency modulation, i.e., when the applied frequency is either $\nu_0 + \nu_m$ or $\nu_0 - \nu_m$. It maintains the corresponding value of this amplitude, and hence, the corresponding value of b [6.28, 6.29]. In this way, using the properties of the caesium beam tube:

- the frequency shift associated with detuning of the resonant cavity is eliminated;

- the long term value of b is stabilised, together with all the frequency shifts arising from it.

Effect of neighbouring transitions. The hyperfine sublevels closest to the sublevels $F = 3$, $m_F = 0$ and $F = 4$, $m_F = 0$ (Figure 6.2) may perturb the clock transition. This perturbation arises from two sources, leading to frequency shifts known as *Rabi pulling* and *Ramsey pulling*.

- *Rabi pulling.* The wings of the Rabi pedestal for transitions

$$F = 3,\ m_F = 1 \quad \longleftrightarrow \quad F = 4,\ m_F = 1$$

and

$$F = 3,\ m_F = -1 \quad \longleftrightarrow \quad F = 4,\ m_F = -1$$

contribute to the response of the caesium beam tube, although only very slightly, when the excitation frequency is equal to the clock frequency ν_0. Lateral lines are symmetrically placed about ν_0. However, their amplitudes may differ, as can be seen from the spectrum in Figure 6.12a, which typifies magnetically deflected caesium beam tubes. In such resonators, the magnetic fields attainable with state selection magnets are not sufficient to ensure that $\partial E(F, m_F)/\partial b$ in (6.13) will have the same value for all hyperfine sublevels. It follows that the deflection angle depends on the hyperfine sublevel under consideration, and hence that the amplitudes of lateral resonances are not symmetric on either side of the central resonance.

When such an asymmetry exists, a frequency-dependent contribution is added to the tube response described by (6.23). This slightly distorts the reference resonance line and there is a shift in the resonance frequency. It is mainly dependent on the power of the microwave signal, since the amplitude of the wings of the Rabi pedestal is proportional to this power, and also to the value of the static magnetic field B_0 which determines the separation of the spectral components in Figure 6.12a [6.30].

- *Ramsey pulling.* Up to now, we have been assuming that the ultrahigh frequency magnetic field lies parallel to the static magnetic field. However, there is also a component lying orthogonal to the static field, and this for two reasons. Firstly, the lines of the ultrahigh frequency magnetic field must close within the region occupied by each resonance mode in the cavity. This means that these lines have to curve in the region crossed by the atoms, as shown in Figure 6.6. In addition, the static magnetic field may not be perfectly aligned. Transitions $\Delta F = \pm 1$, $\Delta m_F = \pm 1$ can thus be induced, although with much smaller probabilities than transitions $\Delta F = \pm 1$, $\Delta m_F = 0$.

The four transitions

$$(3,\ 0) \quad \longleftrightarrow \quad (4,\ \pm 1) \quad \text{and} \quad (3,\ \pm 1) \quad \longleftrightarrow \quad (4,\ 0)$$

share one level with the clock transition

$$(3,\ 0) \quad \longleftrightarrow \quad (4,\ 0)\ .$$

It follows that the wave function for sublevels (3, 0) and (4, 0) is perturbed when transitions $\Delta F = \pm 1$, $\Delta m_F = \pm 1$ occur.

It has been shown that this effect adds a term to the clock transition probability and that it also distorts the resonance line when the ultrahigh frequency spectrum of the caesium atom is not symmetric, like the one in Figure 6.12a. As above, the associated frequency shift depends on the power of the ultrahigh frequency wave and static field B_0 [6.31].

In commercially designed caesium beam tubes using magnetic deflection, the relative frequency shift resulting from the presence of neighbouring transitions can reach values of 10^{-12}. It can be reduced by:

(i) designing the tube so that the heights of the resonances $(3,\ 1 \leftrightarrow 4,\ 1)$ and $(3,\ -1 \leftrightarrow 4,\ -1)$ are as close as possible;

(ii) limiting the spatial extent of the interaction with the oscillatory field;

(iii) aligning the directions of the static and ultrahigh frequency magnetic fields as accurately as possible.

All other things being equal, the frequency shift decreases as the atomic resonance narrows. It is generally small in primary frequency standards.

Effect of Majorana transitions. If the magnetic field varies along atomic trajectories, as it does near the magnetic state selectors, atoms are subject to a time dependent perturbation due to their motion. This perturbation may have spectral components corresponding to the separation between sublevels of states $F = 3$ and $F = 4$. The resulting transitions $\Delta F = 0$, $\Delta m_F = \pm 1$ are called Majorana transitions. If they couple with the transition

$$(3,\ 0) \quad \longleftrightarrow \quad (4,\ 0)\ ,$$

the clock transition frequency is shifted. When building the atomic clock, every effort is made to produce sufficiently gradual magnetic field variations to ensure that these transitions do not occur.

Effect of unwanted spectral components. The spectrum of the interrogation signal includes unwanted components. Some of these, produced in the supply circuitry, occur at the frequency of the electricity grid and its multiples. Others arise from unwanted frequencies generated when synthesising the excitation signal.

The presence of such components perturbs the properties of the quantum system made from the two states $F = 3$, $m_F = 0$ and $F = 4$, $m_F = 0$. The energy difference between these levels is thereby altered, and this leads to a change in the transition frequency. The effect vanishes if the interfering spectral components are symmetrical in position and amplitude about the atomic transition frequency. Otherwise, the frequency shift is the most sensitive to the presence of interference lines when these lines lie in a spectral interval centred on the frequency ν_0 and extending to several times the width of the Rabi pedestal. Over this range, every effort is made to hold their power down to a factor of 10^{-7} of the useful ultrahigh frequency signal.

Effect of imperfect electronic systems. If we assume that the atomic resonance width is 500 Hz, the electronics associated with the caesium beam tube must measure the resonance frequency with an error that cannot exceed one millionth of the line width, if we require the relative error in the determination of the resonance frequency to remain below 5×10^{-14}.

The electronic system associated with the caesium beam tube must therefore be designed and built with the greatest possible care. The required specifications are most easily achieved using digital signal processing methods.

The effect of such imperfections is most effectively reduced by observing a narrow resonance.

(3) Other frequency shifts. It is always possible that other, quite unsuspected frequency shifts are taking place. It was thus that Ramsey pulling was only identified and modelled quite recently, towards the end of the 1980s.

The search for such frequency shifts is a permanent concern for physicists designing and exploiting atomic frequency standards, and in particular, caesium beam primary standards. From this standpoint, it is of great importance to build independent caesium beam primary standards in several laboratories, using as many different designs as possible, and then to check that the frequencies delivered by these instruments are mutually compatible, up to inherent measurement uncertainties.

(f) Long term frequency instability, non-reproducibility and inaccuracy

Frequency shifts in the atomic resonance are caused by the macroscopic environment of the atoms. Any change in this environment can therefore modify the observed atomic transition frequency. This is the origin of long term frequency instability and non-reproducibility. Although there are many frequency shifts, they are very small in relative value compared with the atomic frequency, the largest being something like 10^{-10}. Since the relative variation in these frequency shifts is itself very small, long term frequency instability and non-reproducibility are tiny, as we shall see. Likewise, the relative uncertainty in the value of the frequency shifts is extremely small and inaccuracies in caesium time and frequency standards are also minuscule.

Long term frequency instability. In a caesium beam frequency standard, frequency shifts depend mainly upon the value B_0 of the magnetic field, the atomic velocity distribution, the positions of atomic trajectories relative to the ultrahigh frequency magnetic field in the resonant cavity, the tuning frequency of the cavity, and the amplitude and spectral quality of the ultrahigh frequency excitation. The value of one or other of these parameters can change under the effect of the ambient magnetic field, mechanical stresses, temperature, atmospheric pressure and even humidity. It follows that the frequency of the standard may vary in the long term in response to changes in environmental conditions or ageing effects in the materials and electronic components.

Caesium beam primary standards are located in a carefully controlled environment. In addition, the values of significant parameters are regularly measured. Long term frequency instability levels out at a value no greater than 10^{-14} for $\tau > 10$ days in these standards.

In modern commercial clocks, the value of the static magnetic field is actively controlled by properties of the atoms themselves. Another servo loop based on the properties of the caesium resonator holds the amplitude of the ultrahigh frequency excitation at a constant value, thereby suppressing the effects of resonant cavity detuning. As a result, the system is rendered far less sensitive to environmental perturbations. Long term frequency instability is also very small, in fact, less than 10^{-14} for $\tau > 10$ days.

Non-reproducibility. This concept applies in particular to commercial models. It describes the aptitude of an atomic frequency standard of given design to reproduce the nominal value of its output frequency, without calibration, when environmental conditions lie within some specified range of values. Non-reproducibility characterises the relative deviation of the output frequency, either inside a set of instruments, or for a given piece of equipment, from one

run to the next. It is usually expressed in terms of the standard deviation of the relative values of the observed frequency differences.

Non-reproducibility decreases with improved design and better production control during manufacture. It lies between 5×10^{-13} and 10^{-11}, depending on the model.

Inaccuracy. The accuracy of a caesium beam frequency standard characterises its ability to deliver a frequency agreeing with the definition of the second, and this without calibration. To achieve this, isolated caesium atoms should be observed at rest. However, in a real device, atoms move and are subject to various perturbations. The inaccuracy is equal to the overall uncertainty found by combining all uncertainties arising from the different frequency shifts, divided by the unperturbed frequency (Section 5.5).

These uncertainties are of two types. As an example, consider what happens when the effects of phase differences between the oscillatory fields are measured. Half the difference between the frequencies measured for each direction of the beam is equal, in principle, to the frequency shift resulting from a phase difference between the two oscillatory fields. During each measurement, the frequency delivered by the standard fluctuates randomly under the effect of shot noise in the beam, for example. A *random standard deviation* is therefore associated with the measured value of the frequency shift. We must also take into account the fact that the spatial dependence of the phase of the oscillatory field and the path distribution in a cross section of the atomic beam are not perfectly known. Frequency shifts associated with the spatially dependent phase difference are unlikely to have the same absolute value for the two beam directions. A systematic error is therefore produced in the measured frequency shift. Its estimated value is characterised by a standard deviation, sometimes called the *systematic deviation*. For want of a better method, it is treated as arising randomly [6.32]. The variance in the measured effect due to the phase difference between the two oscillatory fields is thus equal to the sum of the squares of the two standard deviations.

Note that in order to measure the (experimentally accessible) frequency shifts affecting a primary standard, an auxiliary frequency standard is necessary. This standard need not be particularly accurate. It must be very stable over the time interval in which the operating parameters of the test standard are varied. A hydrogen maser is generally used as auxiliary frequency standard.

In commercial caesium clocks, the output frequency is corrected for the shift due to the second order Zeeman effect. For the user, everything happens as though the hyperfine transition were observed in zero magnetic field. In contrast, the output frequency is affected by all the other frequency shifts, such as

the second order Doppler effect and the effects of a phase difference between the two oscillatory fields. The estimated values of all these frequency shifts are considered as uncertainties in the measured value of the atomic transition frequency. They contribute to the inaccuracy of these frequency standards. The specified value of the inaccuracy lies between 10^{-12} and 3×10^{-11}, depending on the design.

In laboratory caesium beam standards, the resonance width is an order of magnitude narrower than in commercial systems, the aim being to reduce those frequency shifts that depend on this width. This effect may be achieved by making longer machines, with $L \sim 1$ m. In some cases, magnetic deflection is used to select beam atoms whose speed is close to 90 m s^{-1}, whereas the most probable speed in the beam is around 250 m s^{-1} [6.11, 6.12, 6.13, 6.14]. The various frequency shifts are measured or evaluated as precisely as possible. Inaccuracies in magnetically selected laboratory caesium beam frequency standards range between 7×10^{-15} and 10^{-13}, depending on the design [6.14, 6.33].

Table 6.3 in Section 6.8 compares the inaccuracies of caesium clocks with those for other types of atomic frequency standard.

Note that the inaccuracy of a given frequency standard (e.g., 3×10^{-12}) is necessarily an upper limit for its non-reproducibility (e.g., 5×10^{-13}) and also for its long term frequency instability (e.g., 2×10^{-12} for the lifetime of the instrument).

6.4.2 Optically pumped caesium beam clock

Among the semiconductor lasers developed to meet the needs of optical telecommunications, gallium arsenide lasers emit at a wavelength close to 0.85 μm, which can easily be made to coincide with the caesium D_2 line (Figure 6.1). Transitions between the $^2S_{1/2}$ and $^2P_{3/2}$ levels can thus be used to select and detect states optically. However, semiconductor lasers emitting at 0.89 μm are hard to come by. Although experiments have demonstrated that the D_1 line could be extremely useful, because of the smaller number of levels in the $^2P_{1/2}$ state, it is still not commonplace to benefit from this advantage [6.17].

(a) Description

The optically pumped caesium beam tube includes a source in which the atomic beam is produced and a resonant cavity similar to those in the magnetic deflection device. A uniform static magnetic field of strength B_0 is also applied to the atomic beam in a direction parallel to the oscillatory fields in

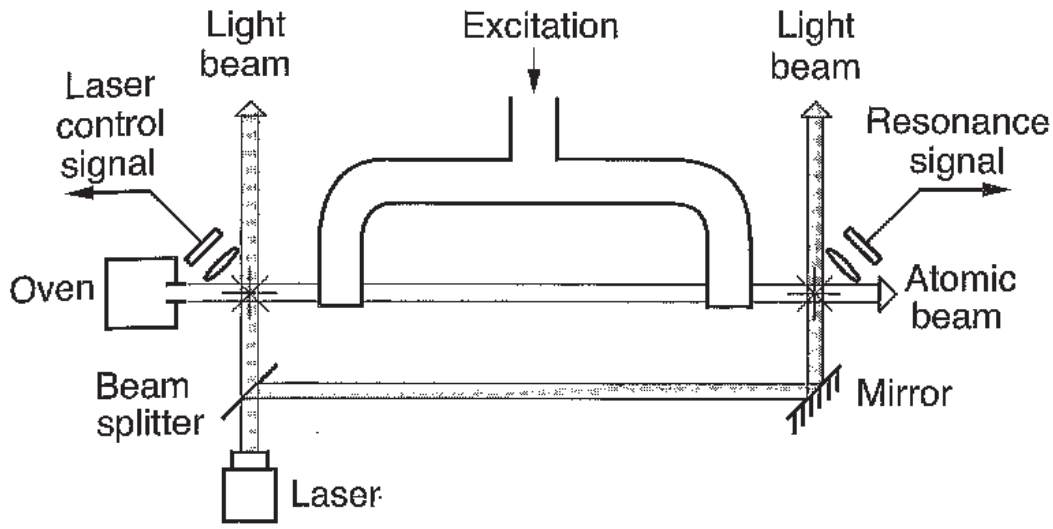


Fig. 6.15. Schematic diagram of a caesium beam tube with optically prepared and detected atomic states. We assume that a single laser has been used. Magnetic shielding is not shown.

the region enclosing the two cavity arms and the region separating them. The magnets are replaced by zones in which atoms interact with *laser light*. Figure 6.15 shows the simplest arrangement, in which a single laser provides the light needed to prepare and detect the states [6.17].

The suitably polarised light beam from the laser has a power of a few mW. It is separated into two parts using a semi-transparent beam splitter. The two beams thus obtained cross the atomic beam at right angles on either side of the resonant cavity. Two interaction zones are thereby created between atoms and light. A magnetic field is produced in each. It lies parallel to the magnetic field applied in the region where the ultrahigh frequency transition occurs and its value depends on the choice of optical transition. Some are more efficient for magnetic fields of the order of 5×10^{-5} T, about ten times greater than the usual values of B_0 which lie closer to 6×10^{-6} T [6.34].

State selection occurs in the first optical interaction region. By optical pumping (Section 6.2.2), atoms are transferred from one level of the ground state hyperfine structure, i.e., $F = 3$ or $F = 4$, into the other. Several pumping schemes exist. Consider σ light, for example, i.e., light polarised linearly in a direction perpendicular to the static magnetic field. If it is tuned to the transition between levels $F = 3$ of the ground state and $F' = 3$ of the $^2P_{3/2}$ state, such light is then said to excite the (3–3)σ transition. It places about 13% of the atoms in the $F = 4$, $m_F = 0$ hyperfine sublevel of the ground state. Sublevels $F = 4$, $m_F \neq 0$ are slightly less populated, but to the same level for equal values of $|m_F|$, as can be seen from the symmetry of the diagram in Figure 6.16. Sublevels of the state $F = 3$ are depopulated [6.34].

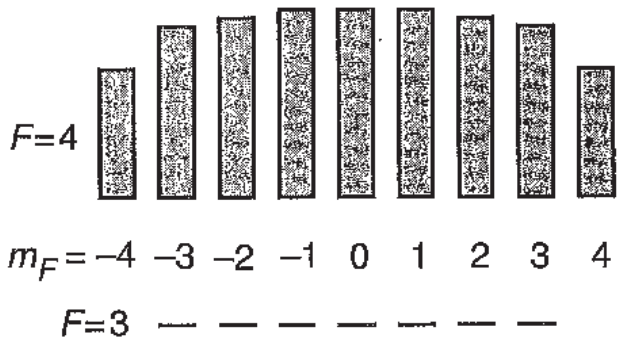


Fig. 6.16. Bar diagram showing the population distribution in levels $F = 3$ and $F = 4$ of the ground state for (3–3)σ pumping.

The population difference between levels $F = 4$, $m_F = 0$ and $F = 3$, $m_F = 0$ is thus equal to 13% of the number of atoms supplied by the source. This value should be compared with that obtained by magnetic state selection. In this case, it is at most equal to the initial population of atoms belonging to the sublevel $F = 3, m_F = 0$ which is deflected towards the axis of the caesium beam tube, viz., 1/16 or about 6% of the initial number of atoms. In reality, the fraction of useful atoms is well below this value because magnetic deflection is velocity selective. Atoms moving too quickly or too slowly are eliminated. However, in the case of optical pumping, the angle of deflection of the atomic beam, resulting here from momentum exchange with laser photons, is at least a hundred times smaller than it is for magnetic state selectors. A negligible number of atoms are thereby sacrificed. We thus observe that state preparation by optical pumping can produce a strong resonance signal.

The flux of fluorescence photons emitted as atoms spontaneously decay from the $^2P_{3/2}$ to the $^2S_{1/2}$ level reaches a maximum when the emission frequency of the laser coincides with the optical transition frequency. The fluorescence photons are collected and used to control the laser frequency by a method similar in principle to the one described in Section 6.4.1c. To this end, the laser frequency is modulated and the fluorescence signal detected synchronously with the modulation. An error signal is thereby obtained. In this case, it acts in two different ways. It influences the temperature of the laser diode using a Peltier effect device, thus producing a rough adjustment of the laser frequency. Fine control is achieved by adding the error signal to the current that supplies the laser and determines the laser emission frequency.

The clock transition is detected in the second optical interaction region. After the atomic beam has passed through the microwave interaction region, the population of the $F = 3, m_F = 3$ level is proportional to the probability that the clock transition has occurred. The number of atoms that the incident laser light transfers from the $F = 3$ level of the ground state to the $F' = 3$ level

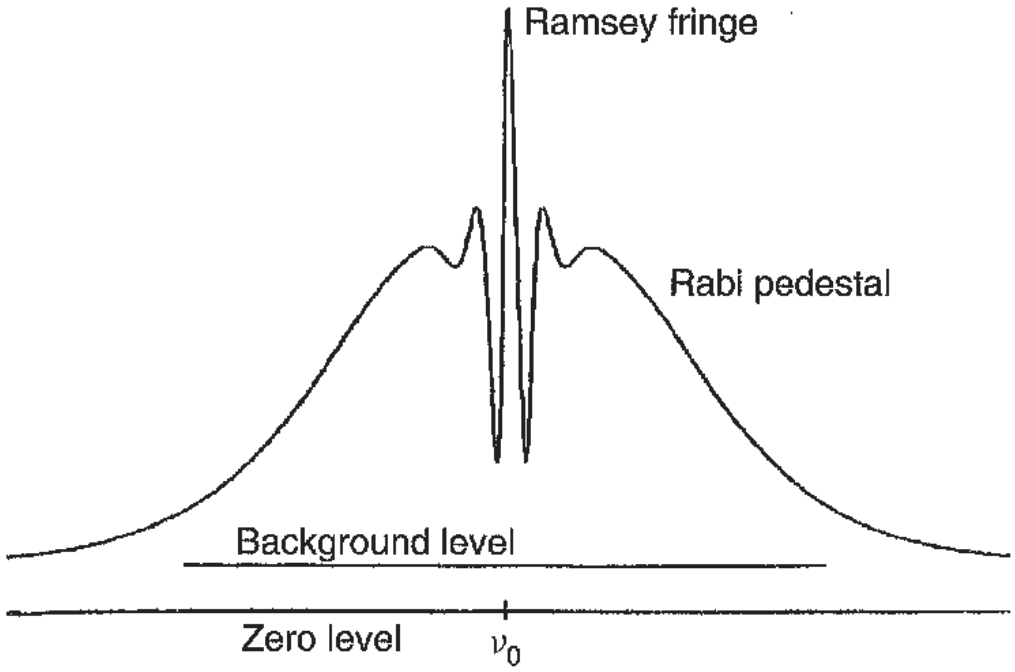


Fig. 6.17. Experimental resonance curve in an optically pumped caesium beam tube. The width of the central fringe is 500 Hz. Taken from [6.17].

of the excited state is proportional to this probability. The same goes for the flux of fluorescence photons. The mean number of photons emitted when an atom in the $F = 3$, $m_F = 0$ state is detected is equal to about four. These are collected with the greatest possible efficiency, in a suitable optical system, in order to ensure that the useful signal is as large as possible compared with stochastic fluctuations. The latter result from thermal excitations in the photon detector, among other things. Unwanted light sources are also minimised as far as possible in the vicinity of the detector. They arise from reflection or scattering of the light beam introduced into the caesium beam tube.

In the described arrangement, state selection and detection are carried out using a *pumping transition*. In practice, optical pumping is completed when the atoms leave each optical interaction region, a few millimetres long. It follows that:

- atoms are used as efficiently as possible;
- the velocity distribution of detected atoms is, in practice, precisely the Maxwellian distribution of atoms in a beam;
- the detected signal is insensitive to the amplitude and frequency noise of the laser.

Figure 6.17 shows an example of the resonance signal obtained in an optically pumped caesium beam tube similar to the one just described, when the microwave excitation frequency sweeps across the clock transition frequency

[6.17]. It can be observed that the lateral Ramsey fringes for this resonance curve are extremely attenuated compared with those in Figure 6.7. This is explained by the broader velocity distribution in optically pumped caesium beam tubes.

The single-laser caesium beam tube described above is well suited to compact systems. Other configurations using different optical transitions have also been studied experimentally. We shall give two examples. The first relates to optically pumped caesium beam laboratory standards, whilst the second aims to improve efficiency in state preparation.

In laboratory devices, detection is improved using a *cycling transition* (Section 6.2.2c). For example, the following arrangement is used [6.35]. The atomic beam is prepared by exciting the transition between level $F = 4$ of state $^2S_{1/2}$ and level $F' = 4$ of state $^2P_{3/2}$. This populates the $F = 3$, $m_F = 0$ sublevel and depopulates the $F = 4$, $m_F = 0$ level. The microwave interaction then repopulates the $F = 4$, $m_F = 0$ sublevel. The latter is detected by means of a second laser source tuned to the cycling transition $F = 4 \leftrightarrow F' = 5$. The spectral purity of this laser is of the utmost importance. We thus obtain several dozen fluorescence photons for every atom that has undergone the clock transition. This increased photon flux masks the intrinsic noise of the detector, thereby improving the signal-to-noise ratio of detection. The latter tends towards the limit imposed by the shot noise of the detected photons. Furthermore, the number of fluorescence photons emitted by each atom is proportional to its transit time through the light beam, so that detection via a cycling transition favours slow atoms. For a given separation between the two oscillatory fields, we thus obtain a narrower resonance than when a single laser is used.

State preparation efficiency can be improved by pumping with two suitably polarised lasers. Several pumping schemes are possible. For example, one laser is tuned to the $F = 3 \leftrightarrow F' = 4$ transition. It pumps atoms in the $F = 4$ level of the ground state. A second laser, tuned to the $F = 4 \leftrightarrow F' = 4$ transition, depopulates the $F = 4$ level, with the exception of the $F = 4$, $m_F = 0$ sublevel, because the transition with $F - F' = 0$, $m_F - m_{F'} = 0$ is forbidden. Finally, all atoms tend to gather in the (4, 0) sublevel. Theory and experiment show that at least 90% of atoms released by the source can thus be transferred into this sublevel. Figure 6.18 compares the experimental microwave spectrum of the seven caesium atom transitions satisfying $\Delta F = 1$, $\Delta m_F = 0$ depending on whether states have been prepared with one or two lasers [6.36]. In the latter case, the intensity of the clock transition is greatly increased, whilst the intensities of the other $\Delta F = 1$, $\Delta m_F = 0$ transitions are significantly attenuated, since the $F = 4$, $m_F \neq 0$ levels are depopulated. However, it was observed that, using the lasers available at the time this experiment was

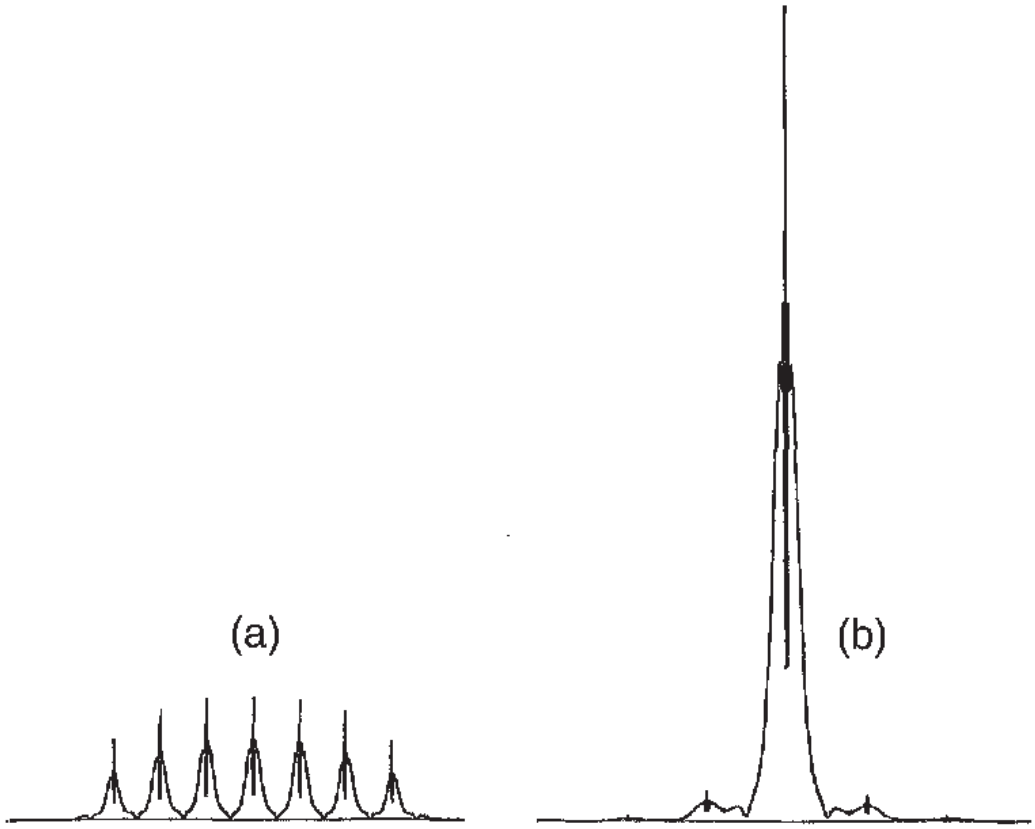


Fig. 6.18. Microwave spectrum of the caesium atom. (a) Single-laser pumping, (b) two-laser pumping. Vertical scales are the same, but horizontal scales are different. Taken from [6.36].

carried out, increased state preparation efficiency was accompanied by a reduced signal-to-noise ratio for observation of the hyperfine resonance. More recently, it has become possible to take full advantage of this improved state preparation using lasers with very narrow emission spectra [6.37].

(b) Medium term frequency instability

As there is no loss of atoms through selective deflection of velocities, and because atoms are more efficiently prepared, the resonance signal is stronger in an optically pumped caesium beam tube. The improvement factor, as measured by I_{pv}, is of the order of 100.

This advantage must be set off against the fact that the resonance is broadened. Indeed, in standard magnetic deflection tubes, the atomic velocity distribution can be displaced towards lower speeds. Typically, the most probable speed of detected atoms lies close to 140 $m\,s^{-1}$ in commercial magnetic deflection tubes, and between 200 and 250 $m\,s^{-1}$ in optically pumped tubes.

In single-laser compact models, noise perturbing observation of the resonance is dominated by noise from the photon detector used in the second optical interaction region. When detection is carried out by means of a cycling transition, a more favourable situation can be achieved in which observed noise arises mainly from shot noise among collected photons.

The advantage gained by a stronger resonance wins through, however. The medium term frequency stability of optically pumped caesium clocks (Figure 6.11) is better than that of instruments based on magnetic state selection. This has been checked both in compact devices [6.38] and in long ones [6.19]. A frequency stability of $3.8 \times 10^{-13}\tau^{-1/2}$ has been obtained using two laser radiations for the state preparation, by optical pumping, and a third for state detection, by a cycling transition [6.37].

(c) Residual frequency shifts

By optically pumping caesium beam tubes, we obtain a better control and a better understanding of the values of frequency shifts affecting hyperfine structure transitions. A further frequency shift occurs, called the light shift. It is due to perturbation of hyperfine energy levels by the light itself, but it is small and does not produce harmful consequences.

Neighbouring transitions and second order Zeeman effect. One important advantage of optical pumping is that it allows us to obtain a symmetrical microwave spectrum, as can be seen in Figure 6.12b. The conditions leading to this are easily satisfied. Pumping light must be polarised perpendicularly to the magnetic field B_0 and the laser frequency must be centred on the optical transition frequency, in such a way that hyperfine sublevels characterised by the same value of $|m_F|$ are equally populated. Rabi and Ramsey pulling then disappear, because nearby transitions perturb the clock transition symmetrically.

The fact that there are no state selecting magnets makes it easier to produce a more homogeneous magnetic field in the microwave interaction regions, since such magnets produce strong magnetic fields outside the core. This in turn reduces uncertainty concerning the value of the frequency shift due to the second order Zeeman effect. In addition, Majorana transitions are also easier to avoid.

Second order Doppler effect and effect of phase difference between the two oscillatory fields. Atomic velocities are greater in an optically pumped resonator than in a magnetic deflection device. However, the velocity distribution is better known. The same is true of the atomic trajectories. Moreover, identical velocity distributions are associated with each possible trajectory.

The second order Doppler effect and existence of a phase difference between the two oscillatory fields therefore lead to larger frequency shifts, but with smaller uncertainties. As the velocity distribution is broader than for magnetic deflection, these frequency shifts are more strongly dependent on the value of the parameter b defined in (6.18), proportional to the amplitude of the oscillatory field. This has no serious consequences when the value of b is servo controlled in such a way as to satisfy (6.76).

Effect of cavity detuning. For given cavity detuning and separation L between the oscillatory fields, the frequency shift is bigger in an optically pumped tube because the greater speed of the atoms broadens the atomic resonance. However, this frequency shift vanishes when the value of b is servo controlled, as already explained.

Light shift. The energies of the two ground state hyperfine structure levels are modified by the laser excitation. This produces a shift in the clock frequency [6.4], known as the light shift.

In an optically pumped caesium beam tube, the frequency shift is due to light perturbing atoms in the microwave interaction region. This light arises from fluorescing atoms in the region common to the atomic beam and light beam, and possibly also from unwanted reflection and scattering of the incident light.

Theoretical analysis and experimental investigations have been carried out on this shift [6.18, 6.39, 6.40]. It is estimated to have a relative value of around 10^{-14} in long laboratory tubes and around 10^{-13} in compact tubes. It can be measured accurately enough not to pose a problem at the present stage in the development of optically pumped caesium beam frequency standards [6.18].

Other frequency shifts. Frequency shifts that have not been mentioned above (e.g., effects of blackbody radiation, collisions, unwanted spectral components, imperfect electronics) do not differ significantly in the two types of caesium beam tube.

(d) Long term frequency instability, non-reproducibility and inaccuracy

In optically pumped caesium beam frequency standards, it is easier to control undesirable frequency shifts and their uncertainties are potentially smaller than in magnetic deflection instruments. This favours a reduction in long term frequency instability, non-reproducibility and inaccuracy.

Relatively recent assessments of optically pumped caesium beam frequency standards have confirmed their potential. The inaccuracy of a laboratory primary standard has thus been estimated at 6.3×10^{-15} and that of a compact instrument at 4×10^{-13} [6.18, 6.38].

It seems likely that optically pumped frequency standards will be able to attain long term frequency instabilities and inaccuracies approaching 10^{-15}.

(c) Conclusions

Optical methods clearly bring with them several advantages. The caesium beam tube is easier to make because atomic trajectories are barely deflected. The more difficult components from a manufacturing point of view, such as the hot wire detector and especially the electron multiplier (in commercial devices) are no longer needed. The fact that frequency stability is improved, an advantage in itself, means also that residual frequency shifts are easier to measure. Optically pumped primary frequency standards deliver frequencies agreeing with their magnetically deflected counterparts. This is highly significant, showing that instruments realising the time and frequency unit in different ways exhibit no systematic bias due to some unknown cause.

6.4.3 Laser-cooled caesium clock

It has long been recognised that the characteristic features of caesium clocks are heavily dependent on the speeds of the atoms whose transitions they exploit. Indeed, as we have shown above, the frequency instability decreases as the resonance line narrows, that is, as the atoms move more slowly. In addition, many frequency shifts are increasing functions of the speed.

The development of methods for cooling atoms has opened up very promising prospects for reducing frequency instability, residual frequency shifts and inaccuracy of caesium atomic frequency standards. We shall present two of these cooling methods. We shall then examine a primary frequency standard using a fountain of cold caesium atoms and describe its performances. This will be followed by a description of projects for cold caesium atom frequency standards designed for use in space. Finally, we shall outline the possible advantages of using rubidium instead of caesium atoms.

(a) Cooling atoms by laser irradiation

A low temperature oven cannot produce an intense beam since vapour pressure is a decreasing function of temperature. Furthermore, the atomic velocity distribution in the vapour phase, in thermodynamic equilibrium with the container walls, tails off rapidly for speeds significantly smaller than the most likely

speed. For example, in a caesium beam produced by an oven at 100 °C, the probability of observing atoms with a speed equal to 5 $\mathrm{m\,s^{-1}}$ is 3×10^4 times lower than the probability of observing atoms whose speed equals the most likely speed in the beam, viz., 260 $\mathrm{m\,s^{-1}}$. In addition, slow-moving atoms are more widely deflected by collisions. It follows that an atomic beam tends to contain very few cold atoms. This is why Zacharias was unable to produce an atomic fountain in the 1950s [6.41]. His idea was to direct an atomic beam vertically upwards. Only the slowest atoms would not reach the top of the device and so would fall back down. In the meantime, they would have been subjected to an oscillatory field during both the rise and the fall, in an interaction zone located close to the atom source. The device was several metres tall to ensure that the time of flight T between the two interaction zones would be of the order of one second.

It was not therefore possible to build atomic frequency standards based on very cold atoms until novel cooling processes had been discovered. The latter arose from various proposals formulated between 1970 and 1975 for cooling atoms in the vapour phase by subjecting them to a continuous photon flux. Theoretical estimates then predicted that it should be possible to cool and even completely immobilise the atoms via momentum and energy exchange with the photons [6.42, 6.43, 6.44]. The development of semiconducting lasers with very high spectral purity emitting at 0.85 μm led to applications to caesium atoms from the middle of the 1980s. Since then, many laser-cooling processes have been devised, analysed and investigated experimentally [6.45, 6.46, 6.47, 6.48, 6.49].

We shall describe the conceptually simplest method for decelerating atoms, used in the first stage of the cooling process. We shall also consider the more subtle method then implemented to perfect the cooling operation, which brings temperatures down as low as 2 μK.

Doppler cooling. Consider the experimental setup illustrated in Figure 6.19a. An atom is subject to two monochromatic progressive light waves, of the same intensity and the same frequency f_L, propagating in opposite directions. Figure 6.19b shows the absorption spectrum of the atom, centred on the resonance frequency f_0. This spectrum is associated with a transition between a ground state level and a level of the excited state $^2\mathrm{P}_{3/2}$ of the caesium atom. The frequency of the light waves is offset so that

$$f_\mathrm{L} < f_0 \,. \tag{6.77}$$

If the atom absorbs photons propagating towards the right, it receives extra momentum in this direction. A few tens of nanoseconds after absorbing each

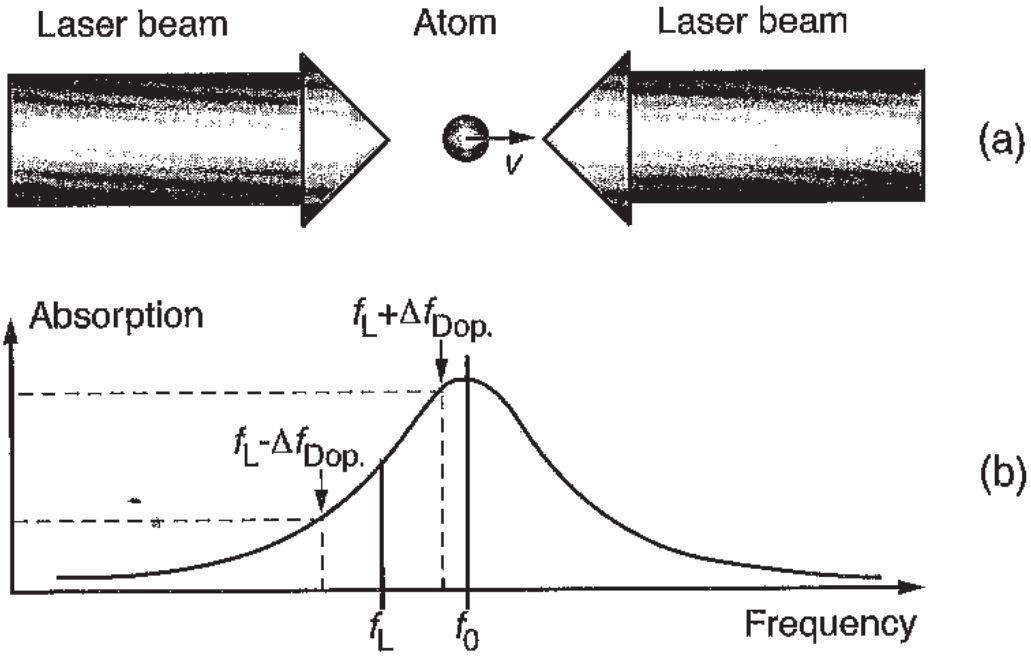


Fig. 6.19. Doppler cooling. (a) An atom, resonating at frequency f_0, is subjected to two light waves of the same amplitude and the same frequency $f_L < f_0$, propagating in opposite directions. (b) The atom absorbs more photons from the laser beam that opposes its displacement.

photon, the atom decays by emitting a fluorescence photon. This emission is isotropic and the recoil momentum associated with fluorescence averages to zero. The overall effect is that the atom gets pushed towards the right. Likewise, the light beam directed towards the left tends to push the atom in this direction.

Let us suppose the atom begins at rest. It absorbs on average the same number of photons from each wave. The mean force felt by the atom is therefore zero. If now the atom is moving at some speed v along the light beam axis, the Doppler effect steps in. In the reference frame of the atom, everything happens as if the frequency of the light waves were shifted by the quantity $\Delta f_{Dop} = f_L v/c$ in absolute value. The frequency of the wave towards which the atom is moving seems to approach its resonance frequency, whereas the frequency of the other wave seems to move away from that frequency, as shown in Figure 6.19b. The atom therefore absorbs more photons opposing its motion than photons that would favour that motion, and consequently it suffers a braking effect. This can be described by

$$\mathbf{F} = -\alpha \mathbf{v} \,, \tag{6.78}$$

where α is a coefficient of viscous friction. The latter is maximal when the frequency of the two light waves is less than the atomic resonance frequency by a quantity equal to the half-width of the resonance curve, which amounts

to about 2.5 MHz for caesium. The coefficient of friction is proportional to the intensity of the light waves.

It is easy to show that, for each absorption–emission cycle, the speed of the atom changes by

$$\Delta v = \frac{h}{M\lambda}\,, \tag{6.79}$$

where M is the mass of the atom and λ the wavelength of the radiation. For caesium, $\Delta v = 3.5 \times 10^{-3}\ \mathrm{m\,s^{-1}}$. The atom must therefore go through several tens of thousands of absorption–emission cycles before it acquires a velocity sufficiently close to zero. The laser light must therefore be fairly intense if cooling is to last for only a fraction of a second.

This viscous deceleration by the Doppler effect cannot produce arbitrarily small speeds. Indeed, the atom is subject to a residual excitation resulting from the random nature of the process itself. Each fluorescence photon is emitted in a random direction and so the atom recoils in an equally random manner. Likewise, the number of fluorescence cycles occurring in a given lapse of time is random and consequently the change in the atom's momentum during this time is also random. A temperature can be associated with this residual excitation. It can be shown that Doppler cooling cannot reduce the temperature below T_D given by

$$kT_\mathrm{D} = \frac{\hbar}{2\tau}\,, \tag{6.80}$$

where τ is the lifetime of the excited state. The value of T_D is close to 120 μK for caesium.

The arrangement shown in Figure 6.19a achieves deceleration in one dimension. Using three pairs of laser beams in mutually orthogonal directions, atoms can be decelerated in all three dimensions. In each pair, the two light waves have the same frequency and amplitude and they propagate in opposite directions. This can be obtained by reflecting the laser beam back along its path.

At the intersection of the three pairs of beams, atoms undergo a viscous force that slows them down whilst conferring a very small scattering velocity upon them. This is why the region at the intersection of the three laser beams is called an *optical molasses*.

Equation (6.78) is valid if atomic speeds are sufficiently small, so that the Doppler frequency shift does not greatly exceed the width of the atomic resonance. If the braking effect is to be effective, the initial speeds of the atoms must therefore be less than a capture speed v_c such that

$$v_\mathrm{c} \sim \frac{\lambda}{2\pi\tau}\,, \tag{6.81}$$

which is independent of the laser intensity, where λ is the wavelength of the light radiation and τ the lifetime of the excited state. In the case of caesium, v_c is of the order of a few metres per second.

In practice, the cooling method considered here is applied to atoms contained in a cell at very low pressure. Their speeds are distributed over a wide range and, although it is small, the probability of finding atoms with speeds below a few metres per second is non-zero. When these enter the optical molasses, they are trapped and cooled. The velocity distribution of the atoms in the cell regenerates through collisions with gaseous phase atoms or with the cell walls, since these collisions redistribute the velocities randomly. In the end, a great many atoms can be trapped. Typically, for laser beams measuring 2 cm in diameter and having intensities of the order of 10 mW cm^{-2}, a ball of cold atoms is obtained, about 1 cm in diameter and containing around 10^7 atoms. Cooling takes about 100 ms.

Properties of the optical molasses can be improved by superposing a non-uniform magnetic field. We may thereby arrange for a restoring force proportional to the distance from the centre of the molasses. Atoms are then trapped in a potential well called a *magneto-optical trap*. Higher atomic densities are then obtained. However, the presence of a magnetic field and the fact that it must be switched off when the atoms are extracted creates practical constraints. As a result, it is usually preferred to cool atoms in a simple optical molasses when they are intended for use in a caesium clock.

Cooling by the Sisyphus effect. It came as a great surprise when it was observed that temperatures below the Doppler limit in (6.80) could be attained.

In fact, more subtle and even more effective cooling mechanisms exist that do not appeal to the Doppler effect. One of these is based on spatial variations in the polarisation of light in a stationary wave and the presence of a large number of energy sublevels in the ground state caesium atom. Cooling occurs through the combined action of energy variations (light shifts) in the ground state sublevels, under the influence of the light, and optical pumping [6.45, 6.46].

Consider a stationary wave formed by superposing two equal amplitude light beams, with orthogonal linear polarisations, propagating in opposite directions. In the stationary wave, the polarisation of the light varies continuously between the linear, right circular, linear, left circular and linear polarisation states, over a distance equal to $\lambda/2$.

We shall simplify the discussion of the cooling mechanism by supposing that the ground state possesses only two sublevels. Their energy is perturbed by the light radiation and depends on its polarisation. The energy of the sub-

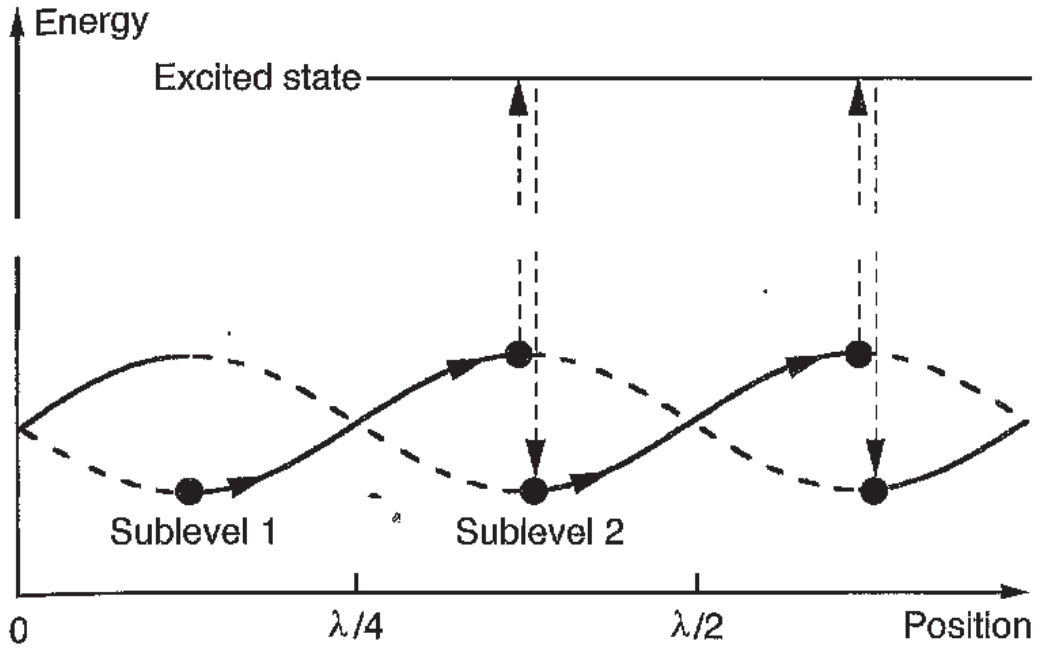


Fig. 6.20. Cooling by the Sisyphus effect. Through its motion, the atom climbs towards a potential peak. Optical pumping causes it to fall back into a trough.

levels therefore varies spatially, with a periodicity of $\lambda/2$. This is shown in Figure 6.20 in a one-dimensional representation. The energy of sublevel 1 is minimal when that of sublevel 2 is maximal, and vice versa.

Let us assume that, at some initial time, an atom occupies sublevel 1 and that it is located in a trough of the potential energy. If it moves to the right, for example, its potential increases and its kinetic energy therefore decreases, i.e., the atom is slowed down. It so happens that it is at the peak of the potential that the light polarisation maximises the probability of the atom being pumped into the excited state. When it reemits a fluorescence photon, it almost immediately finds itself, with high probability, in sublevel 2. But this level has a trough at precisely this location. Continuing its motion, the atom climbs towards the next peak, once again losing kinetic energy. In the end, it will be trapped at the bottom of a valley. The result would have been the same if the atom had moved towards the left, or if it had started in level 2.

This cooling phenomenon has been called the Sisyphus effect in a reference to the hero of Greek mythology who was condemned to push a rock up a mountain, whereupon it immediately proceeded to roll back down into the valley.

In this way, an optical molasses is produced via a mechanism that does not appeal to the Doppler effect. Its features depend differently on the operating parameters. The coefficient of friction is independent of the light intensity I_L and it is proportional to the offset δ between the frequency of the light and

the atomic resonance frequency. The capture speed is given by an equation similar to (6.81), but in which the lifetime of the excited state is replaced by a time constant characteristic of the optical pumping. The latter is inversely proportional to I_L, in such a way that the capture speed goes as I_L. In practice, this speed is smaller than for Doppler cooling. It follows that cooling by the Sisyphus effect must be preceded by Doppler cooling.

The temperature reached is proportional to I_L and inversely proportional to the frequency offset δ, when the latter is large compared with the natural resonance width. When cooling occurs by the Sisyphus effect, the lower temperature limit is determined by the recoil energy of the atom when it emits a fluorescence photon. This limiting temperature is such that

$$kT_R \sim \frac{h^2}{M\lambda^2} . \tag{6.82}$$

For caesium, it is close to 0.1 μK.

In practice, immediately after Doppler cooling has been applied, the intensity I_L of the light beams is reduced by a factor of 10, and the frequency offset δ increased to something like ten times the natural width of the resonance, which is 5 MHz for caesium. A temperature of around 2 μK can then be obtained within a few milliseconds. This corresponds to a root mean square velocity of 1.5 cm s^{-1}.

Launching the atoms. The ball of atoms can be launched at a controlled initial velocity by making use of the Doppler effect.

Consider once again the one-dimensional experimental setup shown in Figure 6.19a, but assuming that the light beam propagating towards the right has frequency $f_L + \Delta f_L$, whilst the beam propagating towards the left has frequency $f_L - \Delta f_L$. If the atom moves to the right at speed v_0 such that

$$v_0 = \frac{c\Delta f_L}{f_L} = \Delta f_L \lambda_L , \tag{6.83}$$

then in a frame of reference in uniform translational motion with the atom, everything happens as if the frequency of the light beam moving to the left is increased by $f_L v_0/c = \Delta f_L$, via the Doppler effect. In this frame, where the atom is at rest, the frequency of the laser beam thus appears to be f_L. A similar argument shows that the same is true for the light beam moving to the right.

The atom is therefore dragged along by a moving optical molasses at the speed $v_0 = \Delta f_L \lambda_L$. Its launch velocity is determined by the value of Δf_L. For caesium, we have $v_0 = 4.25$ m s^{-1} if $\Delta f_L = 5$ MHz.

(b) Clock based on a laser-cooled caesium atom fountain

We have seen how it is possible to launch a ball of cold atoms at a speed of a few metres per second. In principle, it is therefore possible to increase the interaction time Δt in (6.8) up to something like one second, in an instrument of reasonable dimensions. However, in the laboratory, atoms will fall under the effect of the Earth's gravitational field and this must be taken into account. If they were launched horizontally, they would fall by about 5 m after only one second's flight! This would lead to considerable practical difficulties. It is better to take advantage of the Earth's attraction by launching the atoms vertically upwards, so that they rise up several decimetres and then fall back down. This produces an atomic fountain, following the idea put forward by Zacharias.

They are launched with initial speed equal to v_0. The path culminates at a height H above the launch point, given by

$$H = \frac{v_0^2}{2g} , \tag{6.84}$$

where g is the acceleration due to gravity. We find $H = 0.6$ m when $v_0 = 3.4\ \mathrm{m\,s^{-1}}$. The time taken by the atoms to follow their path and return to the launch zone is T_{path} given by

$$T_{\mathrm{path}} = \frac{2v_0}{g} = 2\left(\frac{2H}{g}\right)^{1/2} . \tag{6.85}$$

The value of T_{path} is equal to 0.7 s when the above figures are fed in.

Description of the atomic resonator. The first cold caesium atom fountain created for the study of its properties as a primary frequency standard was made at the LPTF (Laboratoire primaire du temps et des fréquences) in Paris [6.21, 6.22, 6.23]. The experimental setup is shown schematically in Figure 6.21. We shall give a simplified description of the way it works.

The region in which atoms are cooled and launched, and where the atomic states are prepared, is located at the intersection of laser beams marked M. These have diameters of about 2 cm. In this zone, the partial vapour pressure of caesium, determined by the temperature of a cold point, is of the order of 10^{-6} Pa. Cooling and launching are carried out via a cycling transition between the hyperfine levels $F = 4$ of the $^2S_{1/2}$ state and $F' = 5$ of the $^2P_{3/2}$ state. To begin with, the laser power and frequency detuning are adjusted in such a way as to capture and cool atoms through the Doppler effect. This happens for a few hundred milliseconds. Then the atoms are set into upward

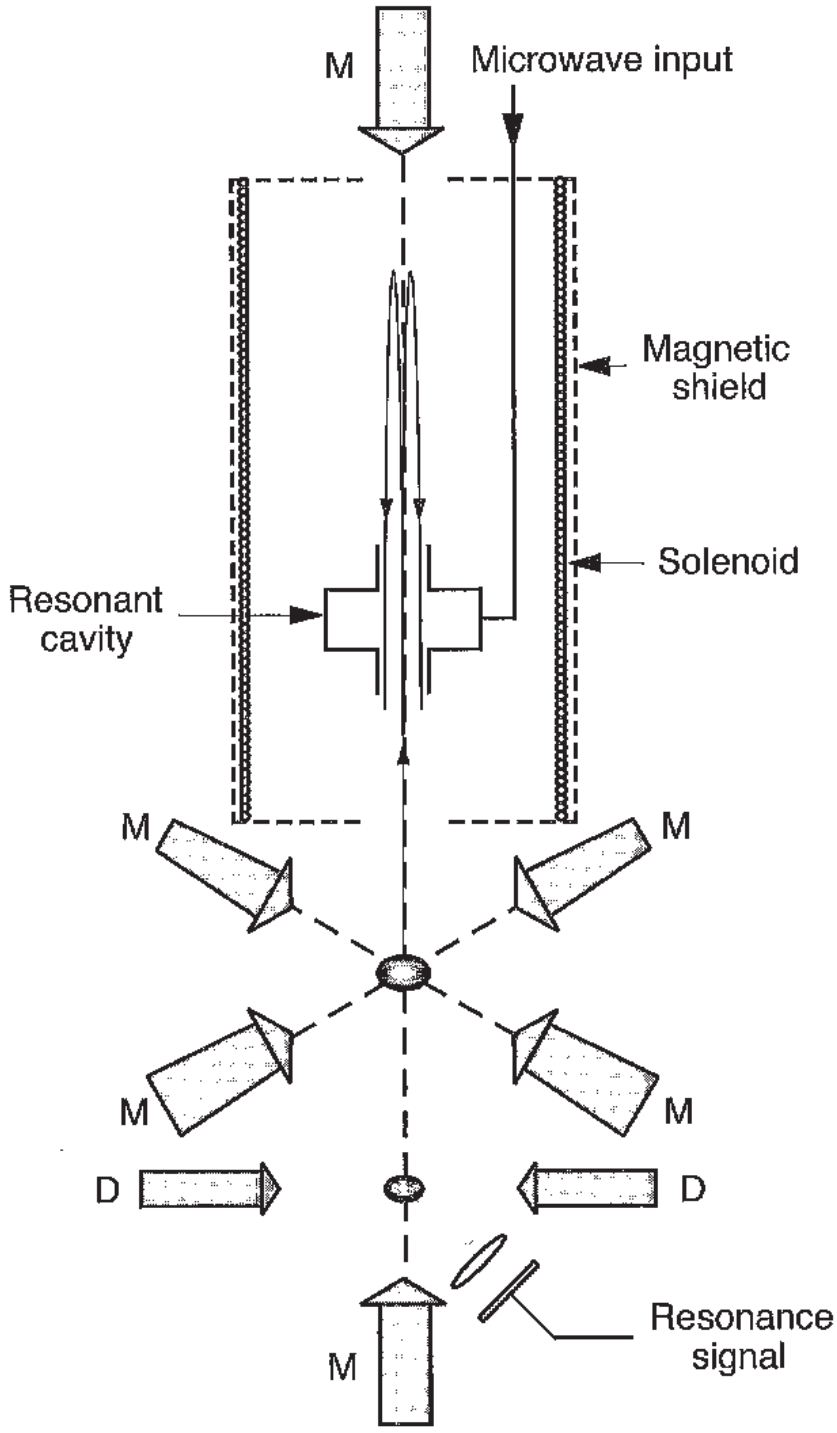


Fig. 6.21. Schematic diagram of a cold caesium atom fountain. M is a light beam forming the optical molasses and D a light beam detecting the hyperfine resonance.

motion by shifting the frequencies of the vertical laser beams for a few milliseconds. The frequency of the upward beam is shifted by $\Delta f_L > 0$ and the frequency of the downward beam by $-\Delta f_L$. The launch speed is given by (6.83). The power of all the laser beams is then reduced and their frequencies diminished by the same amount, close to 50 MHz. Atoms moving upwards are thereby cooled by the Sisyphus effect. In this way, a ball containing about

10^7 atoms is obtained in a region of about 1 cm^3. The ball moves with an initial speed of a few metres per second. The temperature of the atoms is at this point equal to a few microkelvins. Velocity fluctuations are thus of the order of 1 $cm\,s^{-1}$. Atoms in the ball therefore move with more or less the same velocities.

Atomic states are prepared whilst the ball is still crossing the horizontal light beams. The latter are then tuned to the pumping transition $F = 4 \rightarrow F' = 4$ for a few milliseconds and the atoms placed in the $F = 3$ hyperfine level of the ground state. All light beams are then turned off to avoid light shifting the clock transition.

The atoms enter a region where the partial caesium pressure is extremely low, of the order of 10^{-8} Pa, in order to limit the effect of collisions with the residual caesium vapour. They cross the cylindrical resonant cavity for the first time. The cavity is continuously supplied with a signal at 9.192... GHz. They reach the top of their parabolic path, about 30 cm above the cavity, then fall back, crossing the oscillatory field for the second time. The time of flight T between the two crossings is about 0.5 s. In the axial region of the resonant cavity crossed by the atoms, the ultrahigh frequency magnetic field is oriented vertically. The whole region enclosing the resonant cavity and the zone in which the atoms turn back from their upward motion is protected by magnetic shielding made from several layers of material with very high magnetic permeability. A solenoid creates a vertical static magnetic field, which is both uniform and stable. The value of B_0 is small, being close to 10^{-7} T.

The atoms fall until they reach the detection zone, located below the launch point. As they have a non-zero transverse velocity component, they do not all pass into the detection area. A typical number would be around 10^5 atoms. Those that have made the clock transition are now in the $F = 4$, $m_F = 0$ hyperfine sublevel of the ground state. They are detected in a stationary wave formed by the two beams labelled D, tuned to the cycling transition $F = 4 \leftrightarrow F' = 5$. The number of photons collected is proportional to the probability that the clock transition has occurred. A special procedure is used to reduce variations in the detection signal due to fluctuations, from one launch to the next, in the number of atoms contained in the ball.

The atomic resonance is detected sequentially: information concerning the resonator response is collected roughly every second. Figure 6.22 shows a resonance curve obtained by varying the excitation frequency by 0.1 Hz from one launch to the next [6.21]. In this experiment, the time of flight T between the two oscillatory fields was equal to 0.5 s. The width at half maximum of the central fringe is 1 Hz, in agreement with (6.19). The width of the Rabi pedestal, which depends on how long it takes the atoms to cross the oscillatory

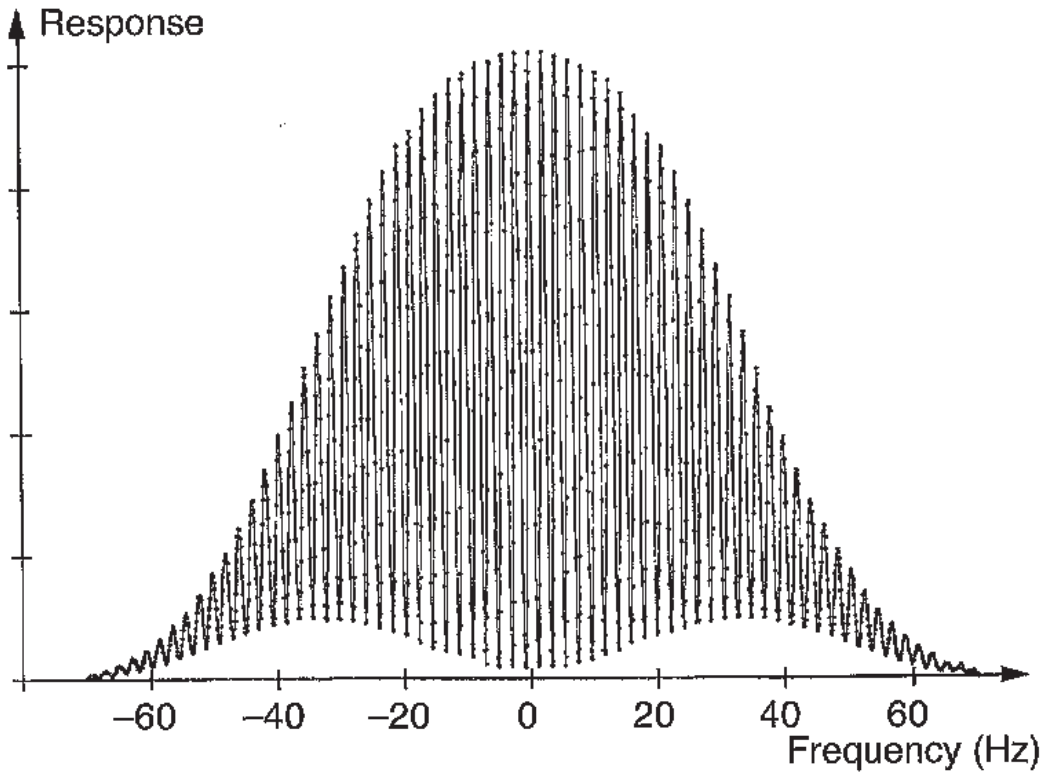


Fig. 6.22. Response of a cold caesium atom fountain. The origin on the frequency scale is placed at the ground state hyperfine transition frequency. From [6.21].

field, is 60 Hz. Ramsey fringes are visible over the whole frequency range explored, because the atoms all have nearly the same velocity. The central fringe is described by (6.17).

Servo controlling an oscillator on the atomic resonance frequency. Since the response of the atomic resonator is only available at a series of discrete times, the method used to servo control an oscillator on the atomic reference frequency must be adapted in consequence. The excitation frequency at 9.192... GHz is synthesised from the signal produced by the oscillator to be controlled and modulated by a square wave. During one measurement cycle, whilst one ball of atoms is prepared and follows its ballistic trajectory, the excitation frequency is $\nu_i + \nu_m$, where ν_m is the modulation depth, equal to the half-width of the resonance at half-maximum. At the end of this cycle, the number of detected photons is N^+. Over the next cycle, the applied frequency is $\nu_i - \nu_m$ and the response is N^- detected photons. The difference $N^+ - N^-$ is zero on average, if ν_i equals the hyperfine transition frequency of the ground state caesium atom. Otherwise, this difference constitutes an error signal which is then fed into a digital integrator. The output acts upon the oscillator frequency and the latter is thus servo controlled by the atomic resonance frequency.

(c) Frequency instability

The medium term frequency instability for this type of atomic frequency standard is determined by fluctuations in the number of detected photons. It can be shown that the Allan deviation for the relative frequency fluctuations, defined in Section 5.2.6, is given by

$$\sigma_y(\tau) = \frac{\sqrt{2}}{\pi Q_{at}} \frac{\sigma_{\nabla N}}{N} T_c^{1/2} \tau^{-1/2} . \tag{6.86}$$

The parameter Q_{at} is the quality factor of the atomic resonance, $\sigma_{\nabla N}$ is the standard deviation of the difference $\delta N^+ - \delta N^-$ between the fluctuations in N^+ and N^-, N is the mean value of N^+ and N^-, and T_c is the duration of one cycle.

The quality factor has a very high value, around 10^{10}, because slow atoms are used, and this clearly implies extremely small frequency instabilities. In 1997, the value measured in the prototype instrument was $1.5 \times 10^{-13}\tau^{-1/2}$ [6.22]. Figure 6.11 shows that this instability is smaller than in other types of caesium clock. In fact, in this measurement, the frequency instability was determined by the spectral quality of the oscillator whose frequency is servo controlled by the atomic resonance. The origin of this limiting effect has been understood. It is related to non-linear effects (aliasing) occurring in the control loop. However, explanation of this effect goes beyond the scope of the present book.

This frequency instability has been reduced and it has attained the value predicted by (6.86), replacing the quartz oscillator by an oscillator with enhanced spectral purity [6.23]. The resonator of the oscillator in question is a sapphire ring cooled to the temperature of liquid helium [6.50]. The frequency instability is then $4 \times 10^{-14}\tau^{-1/2}$. This value is the smallest ever obtained by a primary frequency standard. It represents a frequency instability close to 10^{-16} for $\tau = 1$ day.

(d) Frequency shifts and inaccuracy

Frequency shifts depending on atomic velocities. Because the atoms are moving so slowly, the frequency shift due to the second order Doppler effect is reduced to roughly 10^{-16}. The uncertainty in its value is extremely small since the speeds of the atoms are very precisely known over their trajectories.

The phase difference that may exist between the oscillatory fields in the two branches of a resonant cavity no longer presents any problem in fountain clocks. Indeed, we now have a single oscillatory field which is crossed twice by the atoms. The only remaining effect here is the spatial variation of the phase of the oscillatory field in the cavity. Because atomic paths diverge, atoms are no

longer subject to the same phase during their two transits. Given the low speeds of the atoms, the residual first order Doppler effect is estimated at 2×10^{-16}.

Frequency shifts depending on the quality factor of the atomic resonance. In the cold caesium atom fountain, the quality factor Q_{at} of the atomic resonance is about a hundred times greater than in laboratory standards based on thermal beams. Frequency shifts depending on the value of Q_{at} are therefore strongly suppressed. This is so in particular for the effects of cavity detuning, which vary as Q_{at}^{-2}. These effects can be reduced to a level of 10^{-16}. The uncertainty in these frequency shifts becomes negligible.

Frequency shifts depending on nearby transitions and the static magnetic field. As the width of the Rabi pedestal is considerably reduced and optical pumping guarantees a high level of symmetry in the caesium microwave spectrum, neighbouring transitions can be brought closer to the clock transition by reducing the value of the magnetic field B_0 to about 10^{-7} T. In addition, the uniformity of the magnetic field can be probed by launching atoms up to different heights. The uncertainty in the frequency shift arising from the second order Zeeman effect and nearby transitions can be reduced to a few multiples of 10^{-16}.

Frequency shifts due to collisions. In the ball of cold atoms, the atomic density differs only slightly from that obtained in a standard caesium beam. However, the cross section for the spin-exchange frequency shift increases when the temperature falls [6.51], and the shift in the clock transition frequency becomes measurable by varying the atomic density [6.21, 6.22, 6.52]. It has a value of about 10^{-14} in typical conditions. Extrapolation to zero density eliminates the collision effect. The relative uncertainty in the measurement result is currently smaller than 10%. The frequency shift due to collisions can be reduced. By an improved cooling method, more sophisticated than those discussed in Section 6.4.3a, the transverse component of the atomic velocities can be reduced in such a way that, over the whole trajectory, the atomic density is never much greater than its residual value in the detection zone. The latter is the value determining long term frequency instability. It is also possible to perfect the preparation of the atoms by suppressing $m_F \neq 0$ sublevels which only contribute to collisions.

Other frequency shifts. Frequency shifts that do not depend on the structure of the instrument, such as the shift resulting from blackbody radiation, can be

evaluated with an uncertainty below 10^{-15}. Frequency shifts with instrumental origins, such as those due to Majorana transitions and unwanted spectral components in the ultrahigh frequency signal applied to the atoms, can be sufficiently well assessed to ensure that their value remains below 10^{-15}.

Inaccuracy. By producing fountains of laser-cooled caesium atoms, a spectacular leap has been made towards reaching the ideal of an isolated atom at rest. An inaccuracy of 10^{-15} has been achieved [6.23]. Further improvements are likely to bring a value close to 10^{-16} within our grasp. Many metrology laboratories have undertaken the construction of cold caesium atom fountains with a view to achieving this objective.

(e) Laser-cooled caesium clock in space

In order to increase the quality factor Q_{at} of the atomic resonance and thereby reduce the frequency instability whilst at the same time holding down variations in the frequency shifts that depend on this factor, the time of flight T between the two oscillatory fields must be increased. In the presence of gravity, the fountain height must be increased by a factor of 4 in order to double the value of Q_{at}, as can be seen from (6.85). Limitations may then arise from non-uniformity in the magnetic field, for instance. On the other hand, in the absence of gravity and for a given separation L between the two oscillatory fields, T depends only on the initial speed v_0 in the beam. We may therefore obtain large values of T for reasonable values of L (e.g., $L = 0.5$ m) by reducing the initial speed (e.g., $v_0 = 10\ \mathrm{cm\,s^{-1}}$). Such a situation can be realised on board a satellite. Both the atoms and the instrument containing them fall in the same way so that, after launching, the atoms appear to move with constant, straight line motion in the frame of reference of the experimental device. The Pharao project (Project d'Horloge Atomique par Refroidissement d'Atomes en Orbite) aims to study such a caesium clock in microgravity conditions. Several French research centres are involved, under the auspices of the CNES (Centre national d'études spatiales).

A prototype has been built and successfully tested in the reduced gravity conditions occurring during parabolic flights of a jet plane [6.53]. In space, it is expected that the frequency instability for $\tau = 1$ day and the inaccuracy will both be about 10^{-16}.

This clock and a hydrogen maser [6.54] will probably be installed on the International Space Station in 2004 during the European Space Agency's ACES mission (Atomic Clock Ensemble in Space) [6.55]. Microwave and optical links will provide the means for time and frequency comparisons with ground stations.

The objectives of ACES include improved measurement of the gravitational frequency shift, better tests of the isotropy of the speed of light, and the search for possible variations in the fine structure constant α. With its time transfer equipment, ACES will make it possible to compare time scales produced by remote ground laboratories with an uncertainty as low as 30 ps per day, much smaller than the uncertainty provided by the GPS system (Section 5.6.3).

Similar projects are under development in the USA [6.56, 6.57].

6.4.4 Laser-cooled rubidium clock

The rubidium 87 atom can be optically pumped just as well as the caesium 133 atom, since semiconductor lasers exist at the appropriate wavelength (see Table 6.1). The interest in building fountains of laser-cooled rubidium atoms was clear from the moment it was predicted that the collisional frequency shift of the rubidium 87 hyperfine transition should be significantly smaller than in caesium 133 at equal densities [6.58]. This very interesting property has been confirmed experimentally [6.59, 6.60]. It provides an opportunity to diminish the most bothersome frequency shift, and at the same time to reduce frequency instability by somewhat increasing the atom density. A comparison between the metrological properties of caesium and rubidium frequency standards based on fountains of laser-cooled atoms is now underway.

As a result of the first experiments, a measurement of the hyperfine transition frequency of the rubidium 87 atom has been made with an uncertainty of 1.3×10^{-14} (see Table 6.1) [6.61].

6.5 Hydrogen masers

The hydrogen maser was invented on the basis of research carried out by Ramsey and his team at Harvard, USA, at the end of the 1950s [6.24]. The original aim was to narrow the hyperfine resonance line observed in a caesium beam by passing it through two separated oscillatory fields. To this end, the atoms were held in a cell situated on their path, unable to escape until they had collided several times with the cell walls. The latter were coated with an extremely inert material, such as Teflon or paraffin wax. The experiment was successful. The time interval T between interactions with the two oscillatory fields was increased. However, this success was counterbalanced by a broadening and attenuation of the hyperfine resonance line, caused by the collisions themselves. In addition, the frequency of the resonance was shifted.

It was then realised that hydrogen atoms would be less perturbed by such collisions, because they have smaller mass and less pronounced electrical polarisability. However, the ionisation energy of the hydrogen atom is very high

and it cannot be detected by the hot wire method, in contrast to the caesium atom. Optical methods are also very difficult to implement, because the relevant wavelengths for transitions between the ground state and excited states lie in the extreme ultraviolet (see Table 6.1). It was therefore decided to detect the hydrogen atom using its stimulated emission, as had previously been done in the ammonia maser [6.62, 6.63]. The first hydrogen maser was then successfully built in 1959–60 [6.64].

The hydrogen maser soon became widely used in the USA as a spectrometer and highly stable frequency standard [6.64, 6.65]. Many research centres began to study and develop the idea in Australia, Canada, China, France, Japan, West Germany, Romania, the United Kingdom, Switzerland and the USSR.

At the present time, several hundred hydrogen masers are in use around the world. The majority were built in the USSR [6.66]. In its active (auto-oscillating) form, the hydrogen maser is the time and frequency standard with the smallest frequency instability for sampling periods between 1 s and a few hours, as can be seen from Figure 6.11. (In this figure, it is assumed that the resonant cavity of the hydrogen maser is not controlled by an automatic tuning device. With such an addition, the long term frequency instability is diminished.) In its much more compact passive form (Table 6.2, Section 6.9), the medium term frequency instability of the hydrogen maser, typically close to $10^{-12}\tau^{-1/2}$, is smaller than that achieved in commercialised caesium beam standards. The inaccuracy of (active and passive) hydrogen masers is not as small as it is in laboratory caesium beam standards (Table 6.3). This is due to collisions with the walls of the cell enclosing the atoms. It is nevertheless similar to values for commercial caesium standards. Today the hydrogen maser is mainly used in time and frequency metrology, radioastronomy, geophysics and experimental checks of relativity theory.

The most significant prospect concerns the cryogenic hydrogen maser. The aim is to take advantage of the interesting properties of hydrogen atom collisions on a helium film at a temperature below 1 K.

6.5.1 Basic principles of the hydrogen maser

The hyperfine transition frequency of the ground state hydrogen atom is close to 1420 MHz (Table 6.1). The corresponding value of the wavelength in the vacuum is thus close to 21 cm. The transition used in the hydrogen maser is the same as the one observed by radio astronomers studying interstellar clouds of atomic hydrogen. Such observations lead, among other things, to measurements of the radial velocity component of these clouds, because the transition frequency is Doppler shifted.

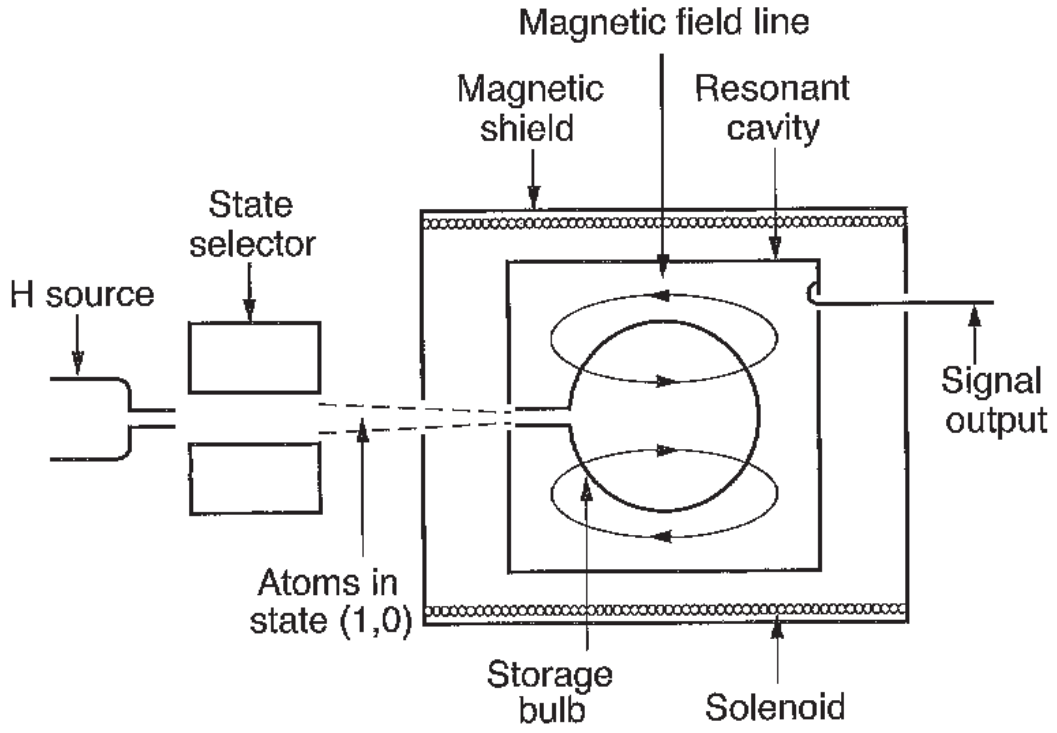


Fig. 6.23. Schematic diagram of a hydrogen maser. The atomic beam is usually vertical.

The original construction of the hydrogen maser is shown schematically in Figure 6.23.

A radiofrequency discharge lamp dissociates a flow of molecular hydrogen. This *atomic hydrogen source* is connected via one or more capillary tubes to a high vacuum region. A collimated atomic hydrogen beam with circular cross section is thus formed. The four hyperfine levels (Figure 6.3) are equally populated at the source output.

States are separated by magnetic deflection (Section 6.2.2). The atomic beam crosses a *multipole magnet*. This comprises four or six poles, magnetised alternately north and south, positioned at regular angular intervals around the beam axis. The poles are located roughly 1 mm from the axis. The magnetic field is contained in a cross section. To a first approximation, its modulus is independent of the polar angle and grows steadily from zero on the axis to a value close to 1 T near the poles. This field deflects atoms in states $F = 1$, $m_F = 0$ and $F = 1$, $m_F = 1$ towards the axis, whilst tending to push atoms in states $F = 1$, $m_F = -1$ and $F = 0$, $m_F = 0$ away from the axis. The tube into the *storage bulb* serves as a diaphragm and it may be considered that only atoms in states $F = 1$, $m_F = 0$ and $F = 1$, $m_F = 1$ are selected. Among these, it is atoms in the state (1, 0) that play the main role. The flux entering the confinement zone is of the order of 10^{12} to 10^{13} atoms per second. The corresponding partial pressure is of the order of 10^{-5} Pa.

The storage bulb is made of quartz, which ensures low dielectric losses. It generally has volume between 1 and 2 dm^3. The inner wall is coated with a fluorine-containing polymer like Teflon. The hydrogen atoms can thus go through 10^4 to 10^5 collisions with the walls before their wave function is significantly perturbed. Practically speaking, this means that confinement times close to 1 s can be obtained.

The confining cell is located at the centre of a cylindrical *resonant cavity* whose axis lies along the axis of the atomic beam. This cavity is tuned to the frequency of the atomic transition. For the design shown in Figure 6.23, the diameter and length of the cavity are around 27 cm.

The general shape of the ultrahigh frequency magnetic field lines is shown by thin lines. Atoms are confined near the centre of the cavity resonance mode, in a region where the phase of the magnetic field is constant. Hence, the transition frequency is no longer affected by the first order Doppler effect (Section 6.2.3) and the width of the atomic resonance is effectively of the order of 1 Hz.

A solenoid, with axis parallel to the cavity axis, produces a constant and uniform magnetic field of around 10^{-7} T. As the ultrahigh frequency magnetic field is roughly parallel to the static field over the region available to the atoms, the clock transition

$$F = 1\,,\ m_F = 0 \quad \longleftrightarrow \quad F = 0\,,\ m_F = 0$$

is allowed to take place by the selection rules.

The hydrogen maser also has magnetic shielding, made from four or five layers of material with very high magnetic permeability. It encloses the resonant cavity and attenuates surrounding field variations by a factor of 10^4 or 10^5.

In most instruments made today, the axis, parallel to the atomic beam, is aligned vertically.

6.5.2 Oscillation condition

Upon entering the storage region, the flux I of atoms in the state $F = 1$, $m_F = 0$ greatly exceeds the flux of atoms in the state $F = 0$, $m_F = 0$. Because the energy of the level (1, 0) is greater than that of level (0, 0) (Figure 6.2), the atomic beam carries energy into the resonant cavity. The atoms can thus amplify the ultrahigh frequency field already present in the cavity, provided that its frequency lies very close to that of the hyperfine transition. They can also sustain an auto-oscillation if the energy carried by the atomic beam is sufficient to compensate for losses from the resonant cavity.

Let us compare the energy flux delivered to the cavity by the atomic beam with the power dissipated in the cavity and hence deduce the oscillation condition.

(a) Power dissipated in the cavity

By a standard result in the theory of electromagnetism, the power P_d dissipated in any structure resonating at the angular frequency ω_c is related to the stored energy W by

$$P_d = \omega_c \frac{W}{Q_c} , \tag{6.87}$$

where Q_c is the quality factor of the resonant cavity. The stored energy is equal to twice the magnetic energy in the cavity. Since the magnetic field is sinusoidal, the time average of the stored energy is equal to

$$W = \frac{1}{2\mu_0} \int_V B^2 \, dv , \tag{6.88}$$

where the integration is carried out over the volume of the cavity. The quantity μ_0 is the magnetic permeability of the vacuum and B is the amplitude of the magnetic field. The latter is a function of the spatial variables, but we may write the result of the integration in the compact form

$$W = V_c \frac{\langle B^2 \rangle_c}{2\mu_0} , \tag{6.89}$$

where V_c is the volume of the cavity and $\langle B^2 \rangle_c$ is the mean squared value of the magnetic field amplitude over the volume of the cavity. This therefore implies

$$P_d = \frac{\omega_c V_c \langle B^2 \rangle_c}{2\mu_0 Q_c} . \tag{6.90}$$

(b) Power delivered by the atomic beam

Let us assume that the electromagnetic field is established in the resonant cavity. Because of their random motions in the storage area, the atoms average out the spatial dependence of the ultrahigh frequency magnetic field component. Everything happens as though they were subjected to a constant amplitude magnetic field of strength $\langle B_z \rangle_b$, where B_z is the component of B lying parallel to the static magnetic field. This is the component that can induce the clock transition $\Delta F = 1$, $\Delta m_F = 0$. The average is carried out over the volume of the storage bulb.

Consequently, the Rabi frequency of the atoms, already considered in Section 6.4.1b, is defined here by

$$b = \mu_B \frac{\langle B_z \rangle_b}{\hbar} . \tag{6.91}$$

In the following, we will also have to consider the *filling factor* η defined by

$$\eta = \frac{\langle B_z \rangle_b^2}{\langle B^2 \rangle_c} . \tag{6.92}$$

This factor characterises the degree to which the atomic cloud and the electromagnetic field in the cavity are coupled.

Approximate expression for the probability that transition occurs. If we assume for the moment that, in the hydrogen maser, the quantum state of the atoms can be represented by the state vector $|\Psi\rangle$ introduced in Section 6.4.1b, the results of that section can be applied.

Following an interaction of duration θ, the probability that the transition from energy state E_2 to energy state $E_1 < E_2$ has occurred is given by

$$P(\theta) = |\gamma_2(\theta)|^2 . \tag{6.93}$$

We have

$$\begin{pmatrix} \gamma_1(\theta) \\ \gamma_2(\theta) \end{pmatrix} = \mathcal{M}^{(1)} \begin{pmatrix} 1 \\ 0 \end{pmatrix} , \tag{6.94}$$

where, at resonance, $\mathcal{M}^{(1)}$ is given by (6.39). We obtain

$$P(\theta) = \sin^2 \frac{b\theta}{2} . \tag{6.95}$$

The probability that an atom stops radiating, either by escaping from the storage cell or by colliding with the walls or another atom, is represented phenomenologically by introducing an exponential distribution for the interaction time [6.64]. Setting

$$f(\theta) = \frac{1}{T_0} \exp -\frac{\theta}{T_0} \tag{6.96}$$

where T_0 is a time constant, the probability averaged over the interaction time is given by

$$P = \int_0^\infty f(\theta) P(\theta) \, d\theta . \tag{6.97}$$

We then obtain

$$P = \frac{1}{2} \frac{T_0^2 b^2}{1 + T_0^2 b^2} \,. \tag{6.98}$$

This result is only approximate. The factor T_0^2 must be replaced by the product of two time constants T_1 and T_2, as we shall see below. The appearance of these two time constants is a standard result when interpreting magnetic resonance experiments in confined gases, liquids or solids.

Exact expression for the probability that transition occurs. The exact properties of hydrogen atoms confined in a storage cell are not as simple to deduce as those of caesium atoms in the caesium beam resonator. Indeed, collisions of the hydrogen atoms with the walls and with one another significantly perturb their state vector, to such an extent that a state vector description of their behaviour is no longer adequate. We must appeal to the notion of density matrix [6.25], which goes somewhat beyond the scope of this book. We shall merely quote the main results [6.4, 6.65].

The exact expression for the probability that the transition

$$F = 1\,,\ m_F = 0 \quad \longrightarrow \quad F = 0\,,\ m_F = 0$$

occurs is given at resonance by

$$P = \frac{1}{2} \frac{T_1 T_2 b^2}{1 + T_1 T_2 b^2} \,. \tag{6.99}$$

Time constants T_1 and T_2 characterise the decrease in population difference and in coherence between states (1, 0) and (0, 0), respectively. They depend on the effects of collisions with the walls and in the gaseous phase, and also on the effusion of atoms through the channel with which the cell is equipped. Typical values of these so-called *relaxation times* are $T_1 = 0.3$ s and $T_2 = 0.5$ s.

When the amplitude of the electromagnetic field increases, and with it the value of b, the probability P tends to 1/2. The populations of energy levels E_1 and E_2 therefore tend to equal out. Analytically, it is the factor $T_1 T_2 b^2$ on the right hand side of (6.99) that accounts for this property, and it is therefore called the *saturation factor*. It is generally of the order of 1 in a hydrogen maser.

Power delivered by the atomic beam. Each atom that makes the transition from energy level E_2 to energy level E_1 releases energy $\hbar\omega_0$. The power P_{at}

delivered by the atomic beam is thus equal to

$$P_{\mathrm{at}} = \frac{1}{2} I \hbar \omega_0 \frac{T_1 T_2 b^2}{1 + T_1 T_2 b^2} \,. \tag{6.100}$$

(c) Oscillation condition

If the power delivered by the atomic beam is sufficient to sustain the electromagnetic field in the cavity, then the power dissipated there is equal to it in the steady state.

When the cavity tuning frequency is equal to the atomic resonance frequency, we obtain from (6.90) and (6.100), using the definition (6.92) for the filling factor,

$$P_{\mathrm{at}} = P_{\mathrm{d}} = \frac{1}{2} \hbar \omega_0 (I - I_{\mathrm{t}}) \,, \tag{6.101}$$

where

$$I_{\mathrm{t}} = \frac{\hbar V_{\mathrm{c}}}{\mu_0 \mu_{\mathrm{B}}^2 \eta Q_{\mathrm{c}} T_1 T_2} \,. \tag{6.102}$$

Equation (6.101) shows that the quantity P_{d} can only be positive, and thus that the oscillation can only be self-maintaining, if

$$I > I_{\mathrm{t}} \,. \tag{6.103}$$

This inequality expresses the oscillation condition. The quantity I_{t} is the threshold value of the atomic flux.

With the typical practical values $V_{\mathrm{c}} = 15\ \mathrm{dm}^3$, $\eta = 2.5$, $Q_{\mathrm{c}} = 3.5 \times 10^4$, $T_1 = 0.3$ s and $T_2 = 0.5$ s, we find that $I_{\mathrm{t}} = 1.1 \times 10^{12}$ atoms per second.

6.5.3 Active hydrogen maser

(a) Oscillation level

Despite appearances from (6.101), the power dissipated in the resonant cavity is not proportional to the atomic flux I. In fact, I_{t} also depends on this flux. Because of collisions between hydrogen atoms, T_1 and T_2 are decreasing functions of the atomic density and hence of I. In addition, the dissipated power depends on the quality factor of the resonant cavity since I_{t} is inversely proportional to Q_{c}. It has been established how P_{d} depends on I and Q_{c} [6.65]. Figure 6.24 shows the dependence on I for different values of Q_{c}.

This figure also shows that P_{d} can only be positive, hence the oscillation condition can only be satisfied, if the quality factor of the resonant cavity is sufficiently large. Indeed, low values of Q_{c} lead to high values of I_{t}, which

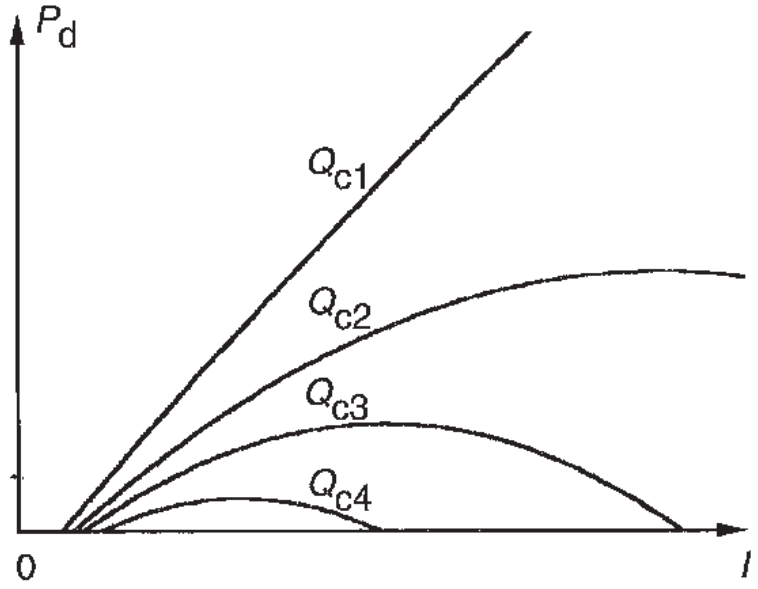


Fig. 6.24. Power dissipated in the cavity as a function of the incident atomic flux, for $Q_{c1} > Q_{c2} > Q_{c3} > Q_{c4}$.

is also increased by the smaller values of T_1 and T_2. Consequently, no values of the atomic flux I can then satisfy the oscillation condition (6.103). The minimum value of the quality factor is around 2×10^4.

In active hydrogen masers, the quality factor of the cavity is typically about 3.5×10^4 and the atomic flux is approximately equal to three times the threshold value. The power dissipated in the cavity is close to 10^{-12} W. The power coupled to external circuits is usually equal to one tenth of this value.

(b) Oscillation frequency

The active hydrogen maser is a self-oscillator delivering a signal at an ultra-high frequency that is mainly determined by that of the atomic transition. This frequency depends slightly on the cavity tuning frequency.

The situation is very different in a caesium clock. In that case, because of the very short interaction time in each oscillatory field, the electromagnetic field in the cavity remains practically unchanged by the passage of the atom beam. In the present case, the atomic medium and electromagnetic field are strongly coupled since it is the presence of the atoms which sustains the field. We must therefore expect a greater dependence of the cavity tuning frequency on the frequency delivered by the instrument.

The oscillation frequency cannot be deduced from the energy relations found above. We need to consider the magnetic moment induced in the atomic medium by the cavity field.

Let $B(\omega) \exp(-i\omega t)$ be the magnetic field to which the atoms are subjected, using complex number notation. The modulus of $B(\omega)$ is equal to $\langle B_z \rangle_b$, intro-

duced previously. A quantum theoretical treatment of the confined hydrogen atoms leads to an expression for the complex amplitude of the magnetic moment of this gas, viz.,

$$M(\omega) = \mu_B T_1 T_2 I \frac{-i[1 + iT_2(\omega - \omega_0)]}{1 + T_1 T_2 b^2 + T_2^2(\omega - \omega_0)^2} \frac{\mu_B B(\omega)}{\hbar} . \tag{6.104}$$

This magnetic moment is highly selective with regard to frequency. When the saturation factor is small, the modulus of $M(\omega)$ is divided by $\sqrt{2}$ for $\omega - \omega_0 = \pm 1/T_2$. The width W of the atomic resonance is thus

$$W = \frac{2}{T_2} , \tag{6.105}$$

and the quality factor is $Q_{at} = \omega_0 / W$, or

$$Q_{at} = \frac{\omega_0 T_2}{2} . \tag{6.106}$$

This quality factor is close to $1\text{–}2 \times 10^9$.

The induced magnetic moment, oscillating at angular frequency ω, reacts on the cavity electromagnetic field. Applying Maxwell's equations to the cavity containing the atomic medium, we find the relation

$$\left[\omega_c^2 - i\omega \frac{\omega_c}{Q_c} - \omega^2\right] B(\omega) = \frac{\mu_0 \eta}{V_c} \omega^2 M(\omega) . \tag{6.107}$$

On the left hand side, we recognise the characteristic factor of a circuit resonating at frequency ω_c and with quality factor Q_c. On the right hand side, the source term is proportional to the oscillating magnetic moment of the atoms.

We now introduce the expression (6.104) for $M(\omega)$ into (6.107) and equate real and imaginary parts on either side of the equation. Using the fact that the angular frequencies ω, ω_c and ω_0 are very close in practice, this yields

$$1 + T_1 T_2 b^2 + T_2^2(\omega - \omega_0)^2 = \frac{\mu_0 \mu_B^2 \eta Q_c T_1 T_2}{\hbar V_c} I , \tag{6.108}$$

and

$$\omega - \omega_0 = \frac{Q_c}{Q_{at}} (\omega_c - \omega_0) . \tag{6.109}$$

At resonance, (6.108) leads back to expression (6.102) for the atomic flux at the oscillation threshold.

Equation (6.109) expresses the oscillation frequency $\omega/2\pi$ in terms of the atomic resonance frequency and the tuning frequency of the resonant cavity. It shows that, for given detuning, the shift in the oscillation frequency of the hydrogen maser is proportional to the ratio of the quality factors Q_c/Q_{at}. Note

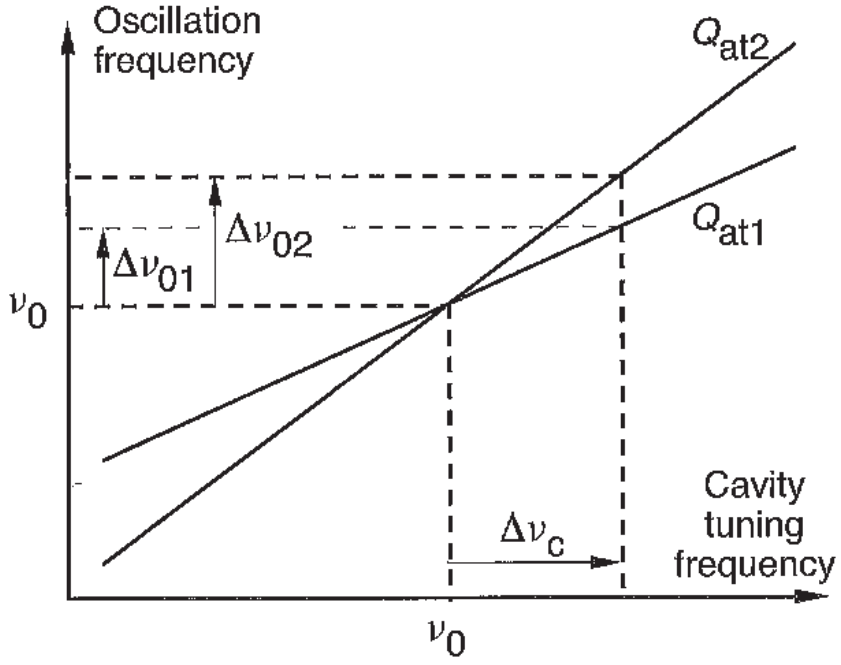


Fig. 6.25. Adjusting the tuning frequency of the resonant cavity by varying the quality factor of the atomic resonance. $Q_{at1} > Q_{at2}$.

that, in the caesium beam resonator, the shift in the transition frequency is proportional to the square of this ratio.

Equation (6.109) also provides a precise way of tuning the resonant cavity. The quality factor of the atomic resonance can be changed by varying the beam intensity. The atomic density in the cell varies, and the relaxation time T_2 varies with it, under the effect of collisions. Therefore, for given $\omega_c - \omega_0$, the oscillation frequency depends on the value of Q_{at}, unless $\omega_c - \omega_0 = 0$.

Figure 6.25 illustrates this method for tuning the resonant cavity. For a cavity detuning $\Delta\nu_c$, the oscillation frequency varies by $\Delta\nu_{02} - \Delta\nu_{01}$ and in proportion to $\Delta\nu_c$, when the quality factor of the atomic resonance is modified from Q_{at1} to Q_{at2}. The cavity is correctly tuned if we obtain $\Delta\nu_{02} - \Delta\nu_{01} = 0$, whatever the value of Q_{at}. Tuning can be carried out manually or automatically (Section 6.5.3e).

The tuning of the resonant cavity must subsequently be held very stable. The relative shift in the oscillation frequency resulting from cavity detuning is

$$\frac{\Delta\nu_{cav}}{\nu_0} = \frac{Q_c}{Q_{at}} \frac{\nu_c - \nu_0}{\nu_0}. \tag{6.110}$$

With $Q_c = 3.5 \times 10^4$ and $Q_{at} = 10^9$, we must achieve $(\nu_c - \nu_0)/\nu_0 = 10^{-10}$ in order to obtain $\Delta\nu_{cav}/\nu_0 = 3.5 \times 10^{-15}$. The resonant cavity must therefore be designed and built with the utmost care. It is thermostatically controlled to

the nearest millikelvin. The tuning frequency is often servo controlled electronically (Section 6.5.3e).

(c) Practical construction of the atomic oscillator

Some general indications have already been given to describe how the various parts of the hydrogen maser are built. We mentioned the atom source, multipole magnet, storage bulb, static magnetic field and magnetic shielding. The aim here is to give a brief practical description of the molecular hydrogen supply, gas pumping system, resonant cavity and various electronic systems required to ensure correct running of the atomic oscillator.

Molecular hydrogen supply. Molecular hydrogen supplies the atomic hydrogen source either from a bottle of compressed gas or from a metal hydride that releases hydrogen when heated.

Gas pumping system. As hydrogen is a permanent gas, a pumping system is required. This involves either one or more ion pumps with capacity $100\,\mathrm{l\,s^{-1}}$, or several pumps containing a material that efficiently absorbs hydrogen, such as titanium foam, assisted by a small ion pump that can evacuate gases other than hydrogen.

Resonant cavity. Several techniques have been used to construct the resonant cavity, with similar results in each case.

One of these corresponds closely to the schematic representation in Figure 6.23 [6.67]. The diameter and height of the cavity are roughly equal (27 cm), which minimises losses in the cavity walls. The quality factor in vacuum is about 50 000, in practice. The storage bulb is spherical with diameter about 16 cm. The cavity is made from a material with very low thermal expansion coefficient, such as solidified silica melt or a glass-ceramic material. A layer of silver deposited on the inner wall limits the volume available to the electromagnetic wave. Thermal control is not applied directly to the cavity walls, but to the walls of the vacuum chamber that contains it, and also to other walls further out. The thermal time constant is large, of the order of 30 to 40 hours.

In a different design, the height of the resonant cavity is equal to about 1.5 times its diameter. The shape of the storage bulb is also somewhat extended [6.68]. This allows better filling of the cavity by the atomic medium. The cavity is made of aluminium which means that it has a high thermal expansion coefficient. Thermal control must therefore be extremely efficient. It acts directly upon the cavity walls, in vacuum. Other walls outside the vacuum chamber are

also thermostatically controlled. Residual fluctuations in the temperature of the resonant cavity are of the order of 1 μK. The thermal time constant is shorter, of the order of 10 hours.

Since the 1980s, one of the main concerns in the construction of hydrogen masers has been to reduce the volume of the resonant cavity, and hence the volume of the whole instrument. To this end, the storage bulb is enclosed in a thin dielectric cylinder. When the additional losses thereby introduced are sufficiently small, the ability to auto-oscillate is maintained. Two different designs have incorporated this feature. In one, the cylinder is a few millimetres thick and made of quartz. This low cost solution cuts the volume of the cavity by around 50% [6.69]. In the other, the cylinder is between 10 and 15 mm thick and made from synthetic sapphire. This material has a high dielectric constant, around 10, and dielectric losses are very low. The cavity volume is thereby reduced by a factor of about 4 [6.70].

In all these designs, the resonant cavity is equipped with two coupling loops. One extracts part of the power delivered by the oscillation, whilst the other contains a varactor used to fine tune the cavity.

Associated electronic devices. In the atomic hydrogen supply, the discharge is maintained by coupling with the resonant circuit of an oscillator delivering a power of about 5 W at a frequency of the order of 100 MHz.

The hydrogen atom flux is held constant by servo controlling the pressure in the discharge tube at a preassigned value. The latter is adjusted by influencing the flow of molecular hydrogen supplying the atom source. To this end, a specific property of nickel, or a palladium–silver alloy, is exploited, viz., the fact that they become more permeable to hydrogen as their temperature rises.

The resonant cavity is held at 45 to 50 °C by separately controlling the various parts of the hydrogen maser. At least ten distinct circuits are involved. Temperature probes are thermistors, selected for their stable characteristics and placed in a Wheatstone bridge. Any imbalance in the bridge determines the power supplied to heating elements distributed over the surfaces to be thermostatically controlled.

Other electronic circuits supply a constant current to the solenoid and an adjustable voltage to the varicap which tunes the cavity.

(d) Controlling quartz oscillator frequency with maser oscillation

When in auto-oscillation, the hydrogen maser is an active frequency standard. However, the signal it delivers cannot be used directly, mainly because its power is so low, of the order of 10^{-13} W. In addition, its frequency has to be transposed to a value which can be exploited directly by the user.

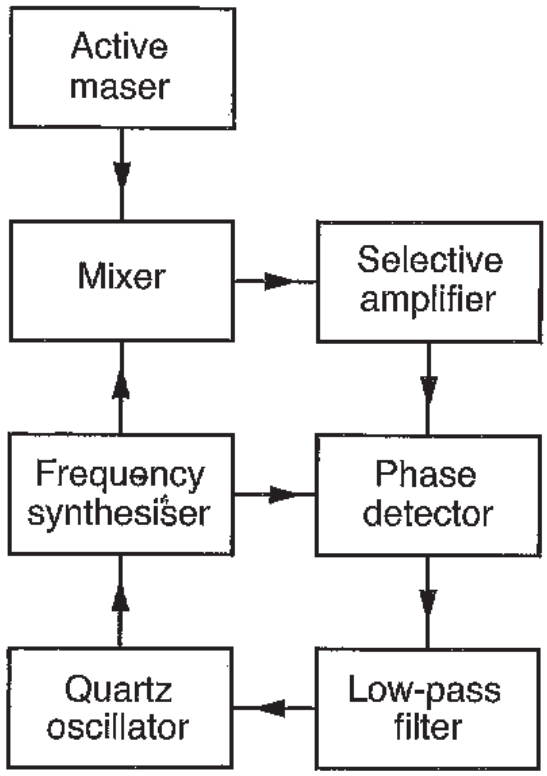


Fig. 6.26. Block diagram of the electronics used to servo control a quartz oscillator on the oscillations of a hydrogen maser.

Quartz oscillator frequencies can be controlled by maser oscillation. Solutions appeal to highly developed heterodyne reception and phase locking techniques.

Figure 6.26 shows a simplified view of the various operations involved. A high-performance quartz oscillator delivers a signal at 5, 10 or 100 MHz. Let ν_{sc} be its frequency when servo controlled by the maser oscillations. This quartz oscillator provides signals that the user can access, via isolation amplifiers. It also drives a frequency synthesiser specially designed for the task at hand. Schematically, it produces a local oscillator signal at frequency $K_1\nu_{sc}$, close to the oscillation frequency ν_r of the maser. The constant K_1 is a rational fraction. The local oscillator signal is mixed with the output signal of the maser, suitably amplified. A signal at the intermediate frequency $\nu_{int} = |\nu_r - K_1\nu_{sc}|$ is thereby obtained. This is amplified and its phase compared with that of another signal, also produced by the synthesiser and of frequency $\nu'_{int} = K_2\nu_{sc}$, where K_2 is a different rational fraction. The phase comparison device produces an error signal if the phases differ. This signal acts on the frequency of the oscillator to be servo controlled, after passing through a low-pass operational filter that determines the servo control transfer function. In a steady regime, the error signal output by the phase comparison device is zero. The phases, and a fortiori, the frequencies of the two input signals are thus equal. Then, $\nu_{int} = \nu'_{int}$

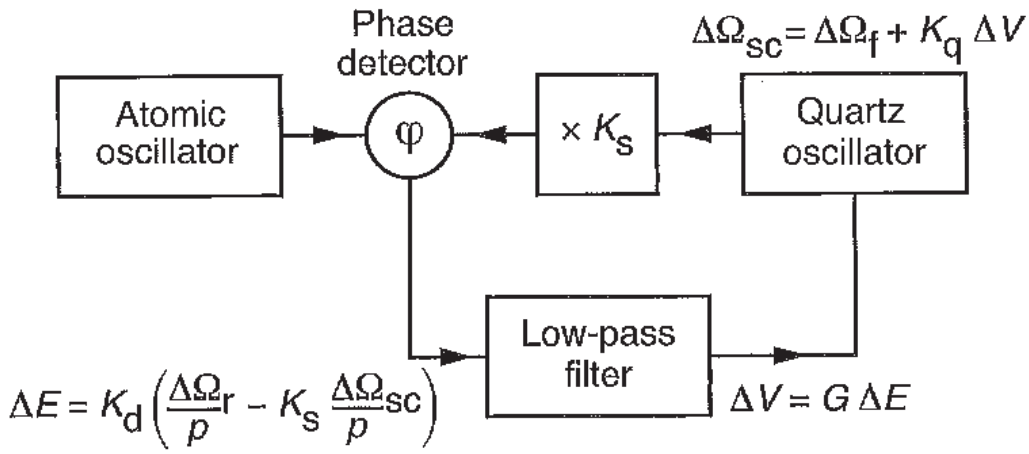


Fig. 6.27. Servo controlling the quartz oscillator frequency by means of an atomic oscillator: functional diagram.

and hence, $\nu_{sc} = \nu_r/K_s$, where $K_s = K_1 + K_2$. The frequency of the servo controlled quartz oscillator is thus tied to that of the atomic oscillator.

Figure 6.27 shows the equivalent functional diagram for servo controlling the phase. The notation is similar to that used in Section 6.4.1c. The phases of the two signals applied to the phase comparison device in the scheme considered are $\Delta\Omega_r(p)/p$ and $K_s\Delta\Omega_{sc}(p)/p$. The error signal is therefore

$$\Delta E(p) = K_d \frac{\Delta\Omega_r(p) - K_s\Delta\Omega_{sc}(p)}{p} . \tag{6.111}$$

Equations (6.47) and (6.48) are both valid, describing the variations in the voltage output by the operational filter and the angular oscillation frequency, respectively. The simplest operational filter that can be used has constant gain A at low frequencies. Hence,

$$G(p) = A , \tag{6.112}$$

and this implies

$$\Delta\Omega_{sc}(p) = \frac{\mathcal{T}p}{1+\mathcal{T}p}\Delta\Omega_f(p) + \frac{1}{1+\mathcal{T}p}\frac{\Delta\Omega_r(p)}{K_s} , \tag{6.113}$$

where

$$\mathcal{T} = \frac{1}{AK_dK_sK_q} \tag{6.114}$$

In practice, the time constant $\mathcal{T}$ is of the order of 0.1 s.

Equation (6.113) has the same form as (6.50), even though the operational filter does not contain a pure integrator. Indeed, when the phase is servo controlled, integration is ensured by the fact that the error signal is proportional to the phase shift of the signals, whilst the control signal acts, not on the phase,

but on the frequency of the quartz oscillator. Note that, in reality, the transfer function of the operational filter is not as simple as we have assumed here. However, this does not affect the general features of the servo loop considered.

The conclusions at the end of Section 6.4.1c also apply. The servo controlled oscillator reproduces the fast frequency variations of the non-controlled quartz and the slow frequency changes of the atomic reference.

(e) Automatic tuning of the resonant cavity

As mentioned in Section 6.5.3b, a severe restriction is imposed on the cavity tuning frequency, if the oscillation frequency is to remain constant. This is why most hydrogen masers are now equipped with cavity auto-tuning systems. We shall give here a brief description of the three procedures used to achieve this goal.

In each case, an error signal is obtained. It is proportional to the extent to which the cavity is detuned with respect to the value procuring an oscillation frequency equal to the atomic transition frequency. The error signal acts on a varicap coupled to the cavity.

The first method is based on measurements of the oscillation frequency of the auto-tuned maser. It requires the use of a second maser as a reference. Generally, the roles of the two masers are exchanged at regular intervals. In the other two methods, however, information concerning cavity tuning is deduced directly from the behaviour of the maser itself.

First method: atomic flux modulation. This method is also known as spin-exchange tuning. It uses the broadening of the atomic resonance that results from collisions between confined hydrogen atoms, with electron, and hence spin exchange. It is the most direct method and makes use of (6.109) giving the change in oscillation frequency as a function of the cavity tuning frequency $\nu_c = \omega_c/2\pi$ [6.71, 6.72].

The atomic flux is slowly modulated, with period of the order of a few hundred seconds, between two values having associated quality factors Q_{at1} and Q_{at2}. Figure 6.25 shows that the variation in the oscillation frequency is proportional to $\Delta\nu_c$, that it changes sign with $\Delta\nu_c$, and that it vanishes when the condition $\nu_c = \nu_0$ is satisfied. The error signal is obtained as the output of a reversible counter. This device counts the beat frequency between the maser to be controlled and the reference maser, counting positively for half the modulation period and negatively for the following half period.

In general, the change $\Delta\nu_{02} - \Delta\nu_{01}$ in the oscillation frequency is smaller than the maser frequency shift associated with cavity detuning. The latter is equal to $\Delta\nu_{01}$ or $\Delta\nu_{02}$, depending on whether the quality factor of the atomic

resonance is Q_{at1} or Q_{at2}, respectively (Figure 6.25). Furthermore, measurement of $\Delta\nu_{02} - \Delta\nu_{01}$ is perturbed by random frequency fluctuations in both the maser to be controlled and the reference maser. Consequently, the latter must also be a hydrogen maser, to ensure a satisfactory signal-to-noise ratio in the measurement of this quantity.

In practice, the time constant for the correction to the cavity tuning frequency of the controlled maser is of the order of one hour. For sampling periods τ less than about one hour, the frequency stability of the maser is degraded. The benefits of this control system are manifested only in the longer term, for τ greater than one day.

Second method: modulation of the cavity resonance frequency. This method uses the inertia of the amplitude of the magnetic moment of the atoms, discussed in Section 6.5.3b [6.69]. It is a consequence of the high values, close to one second, of the relaxation times T_1 and T_2.

The cavity resonance frequency is square wave modulated, with period a few hundredths of a second, much shorter than T_1 and T_2. The electromagnetic field in the cavity represents the cavity response to excitation by the oscillating magnetic moment. Schematically, it is just as though this excitation had constant amplitude and frequency. The response depends on the cavity tuning frequency. The level of oscillation is therefore modulated, unless the variation in the cavity tuning frequency is centered on the atomic transition frequency. An error signal is obtained if this condition is not fulfilled. It can be used to servo control the cavity tuning frequency in such a way that the oscillation frequency equals the atomic transition frequency.

This method guarantees the detection of any cavity detuning with a satisfactory signal-to-noise ratio. The control time constant for the cavity is of the order of one hour. The medium term frequency stability is almost unaffected, and slow drifts in the cavity tuning frequency, and hence in the maser oscillation frequency, are suppressed.

Third method: direct control of the cavity resonance frequency. Suppose for the moment that the resonant cavity contained no hydrogen atoms. We could then use one of the known methods for servo controlling the cavity resonance frequency at a preassigned value. One way would be to apply the method described in Section 6.4.1c, shown schematically in Figure 6.9, but replacing the atomic resonance curve by the cavity response curve. Practically speaking, a frequency modulated signal is injected into the cavity, with frequency variation $\Delta\nu_0'$ about the mean frequency.

For the application to hydrogen masers, care must be taken to ensure that the signal thereby introduced into the cavity does not perturb the oscillation. To this end, it is crucial to suppress the carrier frequency of the applied signal, whose frequency would be equal to the atomic transition frequency. This is achieved very simply when the modulating waveform of the cavity interrogation signal is a square wave. In this case, the carrier frequency is eliminated when the modulation frequency is an even submultiple of the modulation depth $\Delta\nu_0'$ [6.73]. This condition can easily be satisfied using frequency synthesis techniques. In fact, the cavity interrogation signal is synthesised from the signal delivered by the quartz oscillator, whose frequency is servo controlled at the frequency of the maser oscillation.

The cavity interrogation signal may have greater intensity than the maser oscillation. It follows that, when the cavity is detuned, the error signal exhibits a good signal-to-noise ratio. In practice, the cavity control time constant may be short, of the order of 10 s. Long term frequency stability is improved, without significantly affecting medium term frequency stability [6.74].

6.5.4 Passive hydrogen maser

When the auto-oscillation condition is not satisfied, that is, when $I < I_t$, the maser behaves as an amplifier. The resonant cavity is supplied by a signal at 1.42... GHz, produced by an external generator. The response of the standard is just the amplified signal. The maser therefore becomes a passive frequency standard with a highly selective frequency response.

(a) Maser gain

In order to model the external excitation applied to the cavity at angular frequency ω, a source term must be added to the right hand side of (6.107) [6.4]. The complex amplitude of the magnetic field in the cavity is then a solution of

$$\left[\omega_c^2 - i\omega\frac{\omega_c}{Q_c} - \omega^2\right] B(\omega) = \frac{\mu_0\eta}{V_c}\omega^2 M(\omega) - \frac{\omega_c^2}{Q_c} B_e(\omega) \ , \qquad (6.115)$$

where $M(\omega)$ is still given by (6.104). The extra term is such that, when there are no atoms and when we have $\omega = \omega_c$, the magnetic field, with complex amplitude $B'(\omega)$ has the same modulus as $B_e(\omega)$. We set

$$iB'(\omega) = B_e(\omega) \ . \qquad (6.116)$$

The complex gain of the amplifier is the ratio of the value of $B(\omega)$ when there are atoms to its value $B'(\omega)$ when there are none. To simplify, we assume that

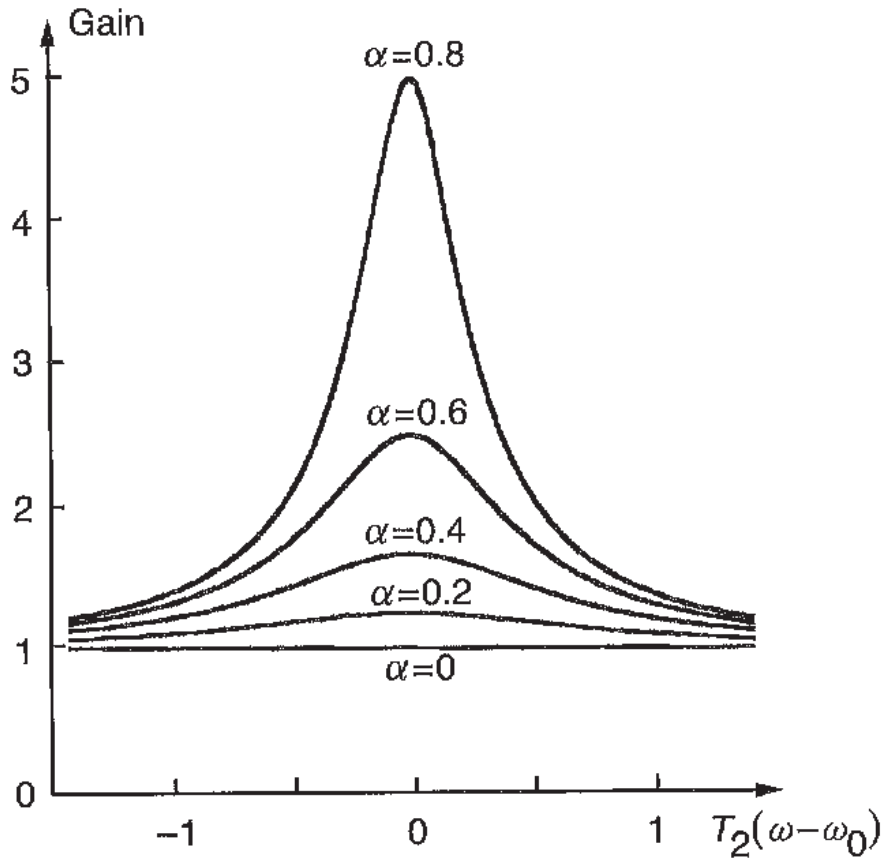


Fig. 6.28. Non-saturated gain of the passive hydrogen maser.

the cavity is tuned to the atomic resonance frequency and restrict discussion to the case where the saturation factor is much smaller than unity. We then obtain

$$|G(\omega)| = \left[1 + \frac{\alpha(\alpha-2)}{1+T_2^2(\omega-\omega_0)^2}\right]^{-1/2} \tag{6.117}$$

for the modulus of the non-saturated gain, where we have defined

$$\alpha = \frac{I}{I_t} . \tag{6.118}$$

Figure 6.28 shows how the modulus of the non-saturated gain varies with the value of $\omega - \omega_0$ for several values of α. At resonance, the gain varies monotonically from 1, when $\alpha = 0$, to infinity, when $\alpha = 1$. In the first case, the maser response is simply that of the resonant cavity. In the second case, the oscillation threshold is reached. The resonance curve narrows as the oscillation threshold is approached. Its width at half-maximum is $2/T_2$ for very small values of α. It tends to zero when α approaches unity.

(b) Resonance frequency

In the passive hydrogen maser, the coupling between the electromagnetic field and the atomic medium is also very strong since the latter amplifies the electromagnetic field injected into the cavity. In practice, the dependence of the atomic resonance frequency on the cavity tuning frequency is also given by (6.109) [6.75]. Here ω must be interpreted as the angular frequency of the atomic transition, as observed when the angular resonance frequency of the cavity is ω_c.

(c) Practical construction of the atomic resonator

Active and passive masers differ essentially through the volume of their resonant cavities. In order to reduce this volume by a factor of the order of 10, materials must be introduced into the cavity whose losses make it impossible to obtain a quality factor greater than 15×10^3.

The oscillation condition can no longer be satisfied and the maser must thereafter be operated passively. (However, it is possible to increase the quality factor electronically, by means of a Q-amplifying circuit, in such a way that the oscillation condition can be satisfied. This method has been implemented, but is not currently used.)

Naturally, every effort is made to reduce the volumes of other parts of the maser, without significantly changing their structure with respect to the active maser. We shall therefore consider only the various solutions adopted for greatly reducing the cavity volume. These are shown schematically in Figure 6.29.

The cavity can be partially filled by a ring-shaped block of alumina or sapphire (Figure 6.29a) [6.76, 6.77]. In this case, the dielectric material serves as a support for an external metallic coating and also for an internal Teflon coating. These define the boundaries of the resonant cavity and the atom storage cell, respectively.

Another alternative is to introduce metal electrodes into the resonant cavity. These surround the storage cell. They have the shape of cylindrical sectors. In one approach they are rigidly fixed to the cell, without touching the outer envelope of the cavity (Figure 6.29b) [6.78]. In another variant, they are connected to the outer walls of the cavity by radial metal struts (Figure 6.29c) [6.79]. The cavity structure then resembles a magnetron structure. It is easier to reproduce and more rigid than the one shown in Figure 6.29b.

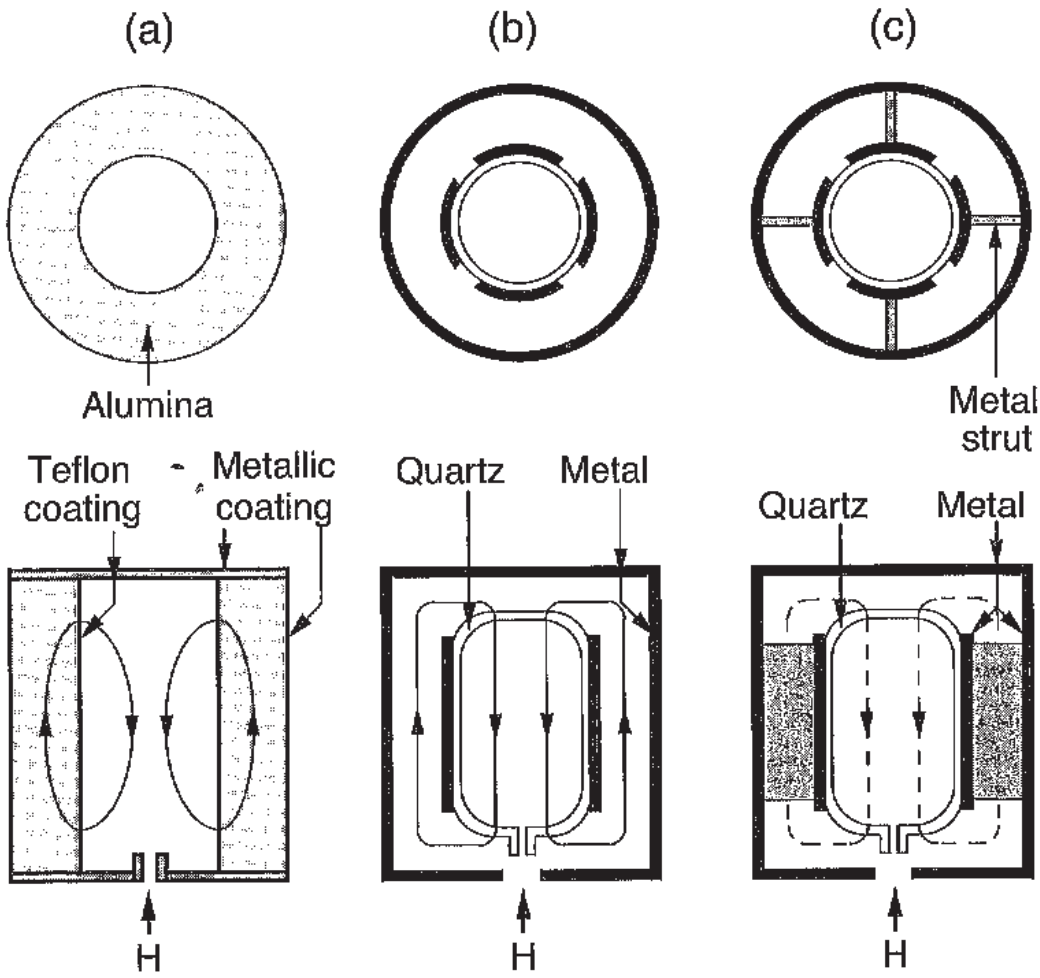


Fig. 6.29. Resonant cavities for passive hydrogen masers. In (c), field lines lie outside the plane of the cross section.

(d) Frequency control of a quartz oscillator and servo control of the cavity resonance frequency

The quartz oscillator could be servo controlled just by transposing the method used in the caesium clock. In this case, the 1.42... GHz interrogation signal would be very slowly modulated, at a frequency of the order of 0.1 Hz, less than the width of the atomic resonance. This method is not the best, because the modulated interrogation signal would be perturbed by flicker frequency noise. Indeed, the closer the spectral components of the modulated signal lie to the carrier frequency, the more they are affected by this kind of noise. In addition, the time constant of the quartz oscillator frequency control must be large compared with the modulation period, and it would then have too great a value.

In a passive hydrogen maser, resonance is observed by the changing amplitude of the electromagnetic field in a resonant cavity. The interrogation

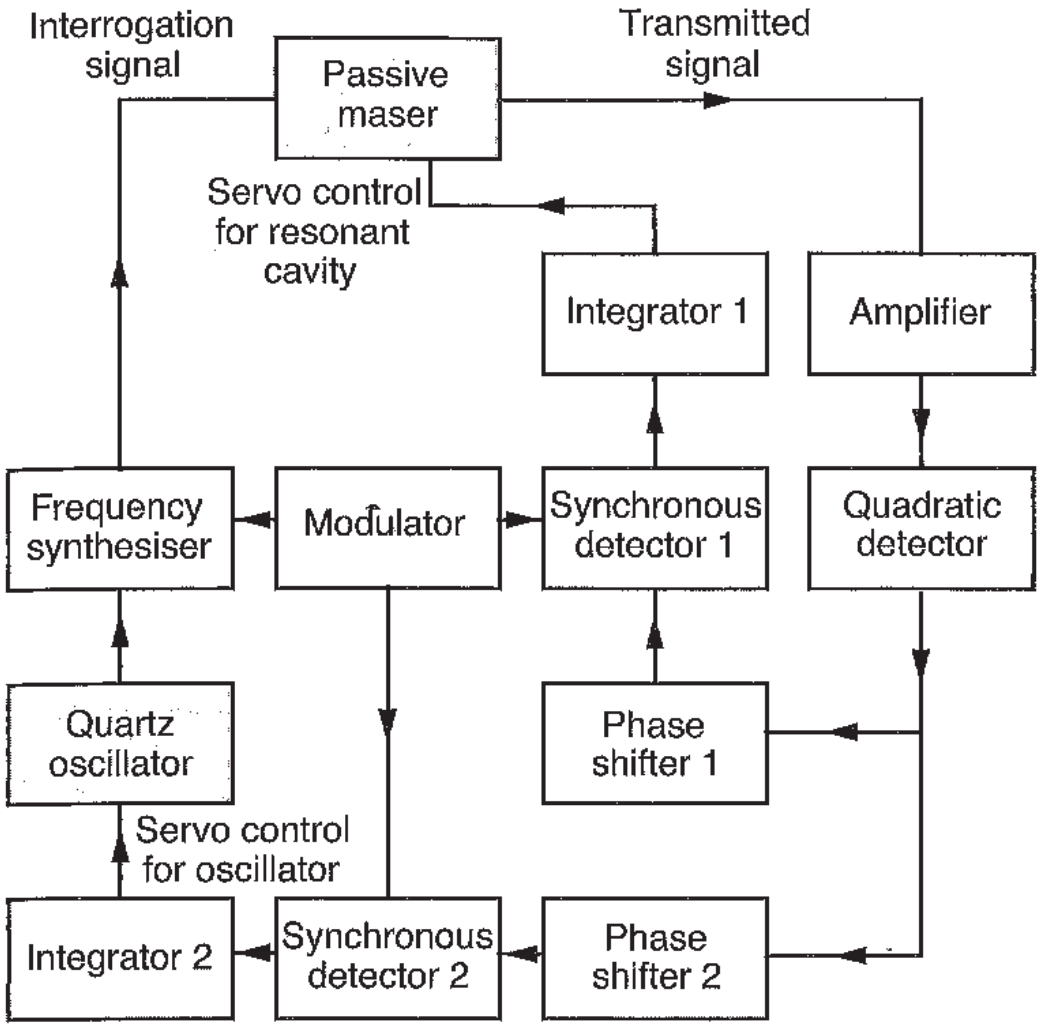


Fig. 6.30. Passive hydrogen maser. Schematic diagram showing how the quartz oscillator and maser cavity can be servo controlled.

frequency can then be stabilised using a variant of Pound's method [6.80]. This method is commonly used at ultrahigh frequencies and in magnetic resonance experiments. In contrast to the method mentioned above, the modulation frequency must be much higher than the resonance width of the reference. The drawbacks described above are thus removed. Using the present variant, both the quartz oscillator frequency and the cavity tuning frequency can be servo controlled at the atomic resonance. The cavity frequency must be stabilised, because its structure means that it has a high thermal expansion coefficient.

Figure 6.30 shows the electronics in a highly simplified way. The interrogation signal, produced via the quartz oscillator, is modulated at a frequency ν_M close to 15 kHz. This is much higher than the atomic resonance width and it has the same order of magnitude as half the pass band of the resonant cavity. The amplitude and phase shift of the transmitted carrier wave depend on the difference between its frequency and the atomic resonance frequency. Likewise, the

amplitude and phase difference of the lateral lines are functions of the cavity resonance frequency.

At the cavity output, the signal is amplified. Its various spectral components interfere in a quadratic detector. The detected signal is modulated and the component at the modulation frequency ν_M is extracted. Two circuits, each introducing an appropriate phase shift, then separate those components expressing the deviation between the carrier frequency and the atomic resonance frequency, on the one hand, and the carrier frequency and the cavity resonance frequency, on the other [6.81]. These two signals are then demodulated in synchronism with the modulation applied to the interrogation signal. Two error signals are thereby produced. After integration, one controls the quartz oscillator frequency and the other the cavity resonance frequency.

6.5.5 Metrological properties of the hydrogen maser

(a) Medium term frequency instability

For sampling times less than about 0.1 s, the frequency instability of the hydrogen maser, whether it is active or passive, is roughly the same as that of the quartz oscillator servo controlled by the atomic resonance. When $\tau > 0.1$ s, we must distinguish the two cases of active and passive masers.

Active maser. In the active hydrogen maser, the electromagnetic field in the resonant cavity induces an atomic magnetic moment which then maintains the electromagnetic field. However, thermal noise occurs in the cavity resonance mode and the associated random electromagnetic field perturbs generation of a coherent oscillation. The noise component in phase with the oscillation creates a frequency fluctuation. It can be shown that the power spectral density of the relative frequency fluctuations (Section 5.2.2) is given by [6.4, 6.82]

$$S_y(f) = \frac{kT}{P_{at} Q_{at}^2} , \tag{6.119}$$

where k is Boltzmann's constant, T is the thermodynamic temperature of the cavity, P_{at} is the power dissipated in the cavity, and Q_{at} is the quality factor of the atomic resonance. This is white frequency noise. Table 5.2 shows that the Allan deviation of the frequency fluctuations is then equal to

$$\sigma_y(\tau) = \frac{1}{Q_{at}} \left(\frac{kT}{2P_{at}} \right)^{1/2} \tau^{-1/2} . \tag{6.120}$$

This expression describes the medium term frequency instability, for τ greater than about 10 s. In the shorter term, for τ between 0.1 and 10 s, frequency

instability is related to the low power supplied by the maser. We must then consider the effect of random signals adding to the oscillation signal. These arise from thermal noise in the cavity itself and from noise contributed by circuits amplifying the signal delivered by the maser. The components of these fluctuations in quadrature with the oscillation produce frequency fluctuations in the observable signal. The power spectral density of the relative frequency fluctuations is then equal to

$$S_y(f) = \frac{4kT}{P_{at}} \frac{f^2}{\nu_0^2} \left(1 + F \frac{P_{at}}{P_r}\right) , \tag{6.121}$$

where ν_0 is the oscillation frequency. The parameter F is the amplifier noise factor and P_r the power effectively received by the amplifier. The Allan standard deviation associated with this white phase noise is

$$\sigma_y(\tau) = \frac{3kTf_h}{2\pi\nu_0^2 P_{at}} \left(1 + F \frac{P_{at}}{P_r}\right) \tau^{-1} , \tag{6.122}$$

where f_h is the cutoff frequency of the device used to observe frequency fluctuations. The standard deviations add quadratically.

Typically, when the cutoff frequency of the device measuring frequency stability is equal to a few hertz, we have $\sigma_y(\tau) \sim 10^{-13}\tau^{-1}$ for $\tau < 10$ s, and $\sigma_y(\tau) \sim 3 \times 10^{-14}\tau^{-1/2}$ for $\tau > 10$ s, as shown in Figure 6.11.

Passive maser. Although the method used to servo control the quartz oscillator by the atomic resonance is not the same as the one used in the caesium clock, equation (6.63) remains valid to the nearest order of magnitude. For the passive maser,

$$\sigma_y(\tau) = \frac{K}{Q_{at}} \left(\frac{kT}{P_{at}}\right)^{1/2} \tau^{-1/2} , \tag{6.123}$$

where K is a constant of the order of 10, that depends on experimental conditions. In practice, we have $\sigma_y(\tau) \sim 1$ to $2 \times 10^{-12}\tau^{-1/2}$ for $1 < \tau < 10^5$ s [6.83].

(b) Residual frequency shifts

The first three frequency shifts that we shall consider are common to all types of atomic frequency standard. The last three are specific to the hydrogen maser.

Second order Zeeman effect. The static magnetic field B_0 applied parallel to the ultrahigh frequency magnetic field shifts the energy levels (Figure 6.2).

This Zeeman effect is second order for levels with $m_F = 0$. The frequency shift is

$$\Delta\nu_Z = 2.7730 \times 10^{11} B_0^2 \,, \tag{6.124}$$

where B_0 is given in teslas. B_0 is usually close to 10^{-7} T. This means that $\Delta\nu_Z/\nu_0 = 1.95 \times 10^{-12}$. The value of this frequency shift is measured by causing low frequency transitions $\Delta F = 0$, $\Delta m_F = \pm 1$ between sublevels of the triplet $F = 1$. In practice, relative uncertainty in the value of this frequency shift is less than 10^{-13}.

Second order Doppler effect. This frequency shift of relativistic origin is related to the motion of the atoms relative to the laboratory reference frame. The hydrogen atoms are thermalised by impacts with the storage cell wall and their kinetic energy is $3kT/2$. From (6.16), the frequency shift is then given by

$$\frac{\Delta\nu_D}{\nu_0} = -\frac{3kT}{2Mc^2} \,, \tag{6.125}$$

where M is the mass of the hydrogen atom. Assuming that the temperature of the storage cell walls is 313 K, this gives $\Delta\nu_D/\nu_0 = -4.31 \times 10^{-11}$. Although this frequency shift is the largest affecting the hydrogen maser, if the cell wall temperature is measured to the nearest 0.1 K, it leads to an extremely low uncertainty in the value of the frequency shift, equal to 1.4×10^{-14}.

Cavity pulling. This is described by (6.109) and has already been discussed in some detail. When the cavity is carefully tuned, either manually or automatically, the uncertainty introduced by this effect is less than 10^{-14}.

Collisions between hydrogen atoms. Collisions between hydrogen atoms induce a shift in the hyperfine transition frequency. This frequency shift can be almost completely compensated for by suitable manual or automatic cavity tuning methods, so that the residual effect on the frequency delivered by the hydrogen maser is a few times 10^{-13}.

Magnetic relaxation. A frequency shift is produced when the hydrogen atoms move through the ultrahigh frequency and static magnetic field gradients. The shift depends on the value of these gradients, the value of the static field and the composition of the atomic beam. When care is taken to produce as uniform a static magnetic field as possible, this frequency shift is less than 10^{-13} for $B_0 = 10^{-7}$ T.

Collisions with the storage bulb walls. Collision with the wall of the storage bulb modifies the phase of the induced dipolar moment of the atom. The corresponding frequency shift has the form

$$\Delta\nu_{\mathrm{w}} = \frac{K}{D}\,, \tag{6.126}$$

where K is a constant depending on the properties of the coating and the temperature, whilst D is the diameter of the storage bulb. As an example, $K \sim -0.4$ Hz cm at 40 °C for FEP 120 Teflon produced by Dupont de Nemours. Hence, $\Delta\nu_{\mathrm{w}}/\nu_0 \sim -1.9 \times 10^{-11}$ for $D = 15$ cm. This shift can be measured, in principle, by varying D. In practice, difficulties involved in reproducing the coating lead to an uncertainty of the order of 10% in the value of the frequency shift due to collisions with the walls.

(c) Long term frequency instability and inaccuracy

Long term frequency instability. In the long term, frequency instability results from systematic effects connected with dimensional instability of the resonant cavity and variations in the wall effect [6.84]. The first mentioned cause of instability leads to a frequency drift of the order of 10^{-16} to 10^{-15} per day. The latter is of little importance for certain applications, such as very long baseline radio interferometry. For timekeeping, an automatic cavity tuning system must be implemented. The second cause of instability is related to the slowly changing properties of the fluorine-containing polymer used to line the storage bulb. The corresponding change in the frequency is also of the order of 10^{-16} to 10^{-15} per day. These variations can occur in opposite directions.

Long term instability of the hydrogen maser depends on whether it is active or passive, and also on the manufacturer. The most stable in the long term are the active type with cavity auto-tuning. Hydrogen masers have been achieved with long term stability similar to that of caesium beam primary standards.

Inaccuracy. The accuracy of the hydrogen maser characterises its capacity to deliver without calibration a frequency agreeing with the hyperfine transition frequency ν_{H} in the unperturbed ground state hydrogen atom.

In practice, inaccuracy is determined by uncertainty in the value of the frequency shift due to hydrogen atom collisions with the storage bulb walls. It is close to $\sigma_{\mathrm{I}} = 2\text{–}5 \times 10^{-12}$.

The value of ν_{H}, resulting from measurement rather than from a definition as in the caesium case, is tainted with uncertainty. Its standard deviation is equal to 3 mHz in absolute value (Table 6.1), or $\sigma_2 = 2 \times 10^{-12}$ in relative value. The standard deviation of the uncertainty in the frequency delivered by a hydrogen

maser is thus σ_H given by

$$\sigma_H = \left(\sigma_1^2 + \sigma_2^2\right)^{1/2} . \tag{6.127}$$

6.5.6 Cryogenic hydrogen masers

In many areas of activity, it would be useful to have a frequency standard even more stable than current state-of-the-art hydrogen masers. For example, a frequency generator with improved medium and long term frequency stability would be very helpful in studying the metrological properties of the cold atom fountain clock (Sections 6.4.3 and 6.4.4) and the confined mercury ion clock (Section 6.7.3). Such a generator could also satisfy the needs of deep space navigation and it could be used for more accurate measurement of gravitational frequency shifts, with a view to distinguishing the various predictions of general relativity.

Since (6.120) and (6.122) show that the short and medium term frequency instability of the active hydrogen maser are increasing functions of cavity temperature, it would clearly be advantageous to cool the resonant cavity. Concerning the long term frequency instability, it is known that the dimensional stability of materials is improved at low temperatures. Likewise, we may hope to improve the stability of the properties of the material lining the storage cell, if there is a way of continually renewing it. These varied reasons have stimulated several studies of the properties of the hydrogen maser at low temperatures, down to 4.2 K.

However, the most decisive results with regard to a highly stable hydrogen maser were obtained from fundamental physical studies of the hydrogen atom. It was observed that recombination and relaxation of this atom are much reduced on a helium film at temperatures below 1 K. In addition, the frequency shift due to collisions on the film is of the same order of magnitude as in masers operating at room temperature, and the broadening of the hyperfine resonance due to collisions between hydrogen atoms is three orders of magnitude smaller, at equal densities, than at room temperature. The atomic quality factor Q_{at} and the power P_{at} dissipated in the cavity can therefore be significantly increased. Equations (6.120) and (6.122) show that such increases would contribute to reducing the frequency instability of the maser, just like a temperature reduction. From the results, it was estimated that the frequency instability of a cryogenic hydrogen maser at 0.5 K could be of the order of 10^{-18} for $\tau = 10^3$ s [6.85]. However, the prediction was made under the assumption that previously presented theoretical results would still apply. In reality, this is not the case. At 0.5 K, the helium pressure is such that the mean free path of the hydrogen

atoms would be of the order of just 1 cm. Hydrogen atoms would no longer average out spatial variations in the static and ultrahigh magnetic fields. Furthermore, additional frequency shifts appear. These result from increased collision time between hydrogen atoms, which itself arises because the average relative velocities of the atoms are smaller at low temperatures. Further studies are thus required.

In order to build a cryogenic hydrogen maser, a helium dilution refrigerator must be used. This is therefore difficult and costly. One of these masers has been used for fundamental physical measurements [6.86]. At the time of writing, it has not been possible to construct two cryogenic hydrogen masers with the qualities required for metrological assessment via mutual comparison.

6.6 Rubidium cell clocks

Optical detection of the hyperfine structure transition in alkali atoms contained in a sealed resonance cell was first studied in the USA at the end of the 1950s [6.87, 6.88, 6.89]. During the same period, the first alkali vapour frequency standards were built [6.90, 6.91]. It was immediately realised that compact atomic frequency standards could be made and the first commercial models appeared at the beginning of the 1960s [6.92]. In these manufactured devices, the rubidium atom was preferred to other alkali atoms for practical reasons.

The market for rubidium cell atomic clocks turned out to be quite sizeable and many companies have produced or still produce this type of atomic frequency standard. Annual production amounts to several thousand instruments. The emphasis has been on high-performance devices that are both reliable and compact. They may be as small as 0.25 dm^3 (Table 6.2) [6.93, 6.94].

By its very structure, the rubidium cell clock is a secondary atomic frequency standard (i.e., its frequency must be calibrated with respect to a primary source). Its long term frequency instability lies between that of a quartz oscillator and a caesium beam clock. It is mainly used in positioning, navigation and telecommunications systems.

6.6.1 Description

The rubidium clock is a passive-type standard. It uses the *double resonance method*, i.e., optical and microwave resonances, and is shown schematically in Figure 6.31.

A small *lamp*, with volume less than 1 cm^3, contains about 1 mg of rubidium 87 or a mixture of isotopes 85 and 87, together with a noble gas such as

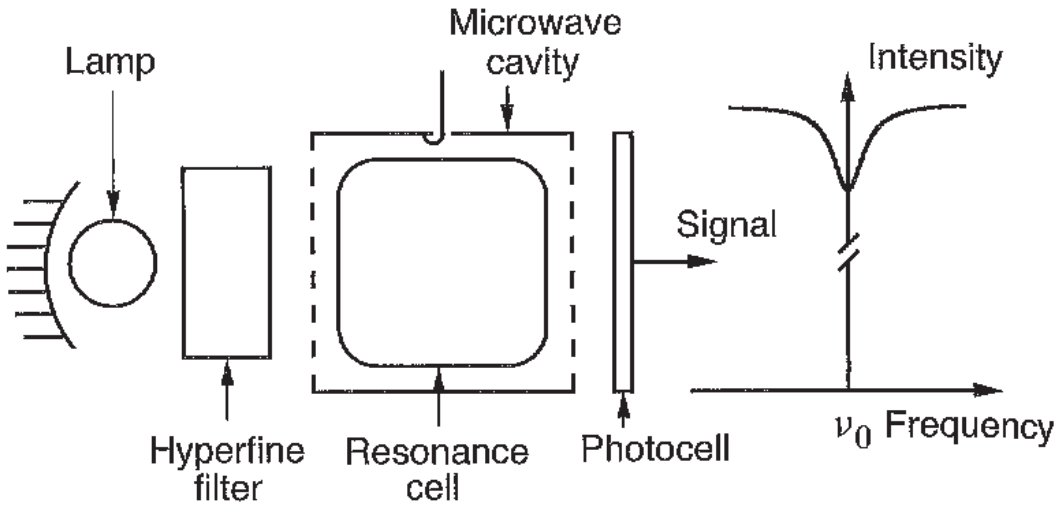


Fig. 6.31. Schematic diagram of the rubidium cell clock.

krypton. A discharge is produced by coupling with a radiofrequency generator. The operating temperature lies between 125 and 140 °C.

In the version shown in Figure 6.30, the light passes through a *hyperfine filter*. This contains rubidium 85 and a noble gas such as argon at a pressure of around 10^4 Pa.

The *resonance cell*, with a volume of a few cm^3, contains rubidium 87. A low pressure *buffer gas* (at about 10^3 Pa) is also present. This consists of a mixture of nitrogen and rare gases.

In the design described here, the hyperfine filter is separate. More compact versions incorporate hyperfine filtering into the resonance cell itself. The latter then contains natural rubidium, composed of 72% rubidium 85 and 28% rubidium 87 and the hyperfine filter is said to be integrated. The quantity of rubidium contained in the resonance cell is very small, of the order of one milligram.

This cell is placed in a *resonant cavity* tuned to the hyperfine transition frequency of the rubidium 87 atom, i.e., about 6834.68 MHz. It receives the microwave interrogation signal. Various procedures are implemented to obtain a suitable electromagnetic field configuration in the cell, whilst reducing the volume of the cavity as far as possible. For example, a cavity structure like the one shown in Figure 6.29c can be used. The two parallel cavity walls have holes pierced in them to let through the light produced by the lamp. The temperature of the resonance cell is maintained at about 80 °C.

The light intensity transmitted by the cell is measured using a silicon *photoelectric detector*. When the frequency of the interrogation signal varies about

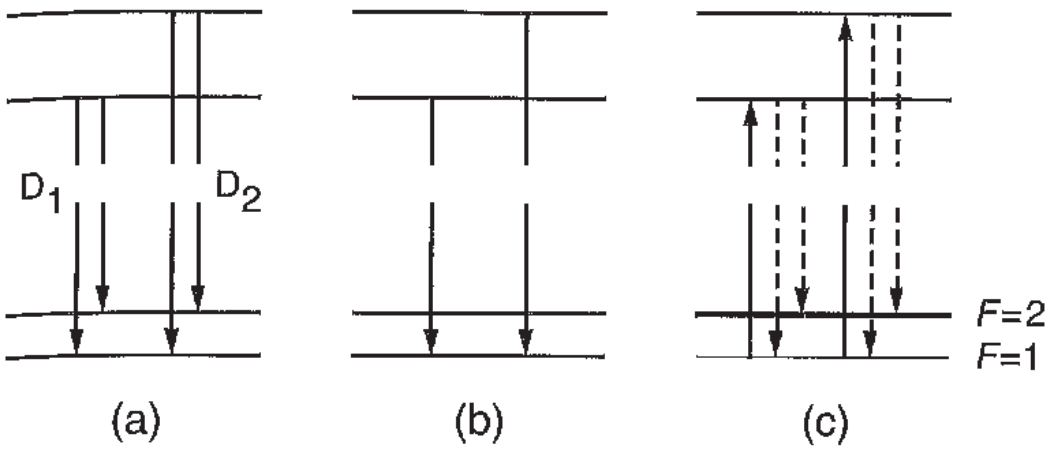


Fig. 6.32. Rubidium clock. (a) Spectrum emitted by rubidium lamp. (b) Filtered spectrum. (c) Optical pumping in the cell.

the rubidium 87 hyperfine transition frequency, the variation in the current delivered by the photoelectric cell has the form of an upside down resonance curve.

Magnetic shielding protects the cell from the surrounding magnetic field and its fluctuations. Suitably arranged coils within this shielding create a magnetic field of about 10^{-5} T.

6.6.2 Operating principles

Complex physical mechanisms take place inside the resonance cell. These arise from the double resonance and the presence of a buffer gas. We shall present only a highly simplified explanation of the atomic resonator, illustrated in Figure 6.32.

In the lamp, rubidium 87 atoms emit resonance lines D_1 and D_2 with wavelengths 795 and 780 nm, respectively. These have a hyperfine structure resulting from the existence of two hyperfine levels $F = 1$ and $F = 2$ in the ground state. Figure 6.32a shows the light spectrum emitted by the lamp in a rather simplified way.

By a favourable isotopic coincidence, rubidium 85 atoms in the hyperfine filter absorb one of the hyperfine components in each of the D_1 and D_2 lines, so that the filter is opaque to these components of the light spectrum emitted by the lamp whilst remaining transparent to the other components. The rare gas contained in the filter shifts and broadens the rubidium 85 resonance lines. Its pressure is adjusted so as to optimise filtering efficiency. Figure 6.32b shows the light spectrum transmitted by the filter.

In the resonance cell, the incident light is selectively absorbed by those rubidium 87 atoms in the $F = 1$ hyperfine level of the ground state (see

Sections 6.2.2b and c). Once the atoms have been excited in this way and thereby enter one of the P states, they relax into the two levels of the ground state, either by fluorescence, or by collisions with nitrogen atoms, the latter being the predominant mechanism.† The absorption–relaxation cycle depopulates the $F = 1$ level in favour of the $F = 2$ level. Figure 6.32c represents the optical pumping cycle thus set up. The cell is transparent to the incident light since the number of atoms capable of absorbing that light has become small.

Up to now, we have not mentioned the ultrahigh frequency field in the cavity. The above description is still valid when this field no longer resonates with the transition between the two hyperfine sublevels of the ground state. Let us now suppose that the frequency of the interrogation signal equals the frequency of the hyperfine transition. The transition $F = 2 \rightarrow F = 1$ is stimulated and the $F = 1$ level is thereby repopulated. A fraction of the incident light is absorbed and the light intensity reaching the photocell is slightly reduced in consequence. Figure 6.31 shows the resonance profile that appears when the microwave signal frequency sweeps across the atomic resonance frequency. This Lorentzian profile can be described by the equation

$$I = I_n \left[1 - \frac{k}{1 + 4(\nu - \nu_0)^2/\Delta\nu^2} \right], \tag{6.128}$$

where I is the flux of photons collected, which equals I_n outside resonance. The width at half-maximum of the resonance is $\Delta\nu$, of the order of 500 Hz, and k is a parameter close to 0.5%.

The buffer gas in the resonance cell plays several important roles. Apart from the non-radiative decay of rubidium 87, it suppresses Doppler broadening of the hyperfine resonance (Section 6.2.3b). In addition, it reduces the rate at which atoms collide with the cell walls, thereby reducing the relaxation and frequency shift associated with such collisions. The buffer gas pressure is arranged so that diffusion velocities of rubidium atoms are very small, of the order of 1 $\mathrm{cm\,s^{-1}}$. To all intents and purposes, the rubidium atoms are stationary during their relaxation time of around 1 ms. They cannot therefore average out the value of the optical and microwave fields crossing the cell. In consequence, the properties of the cell are not homogeneous and this greatly complicates quantitative analysis.

When hyperfine filtering is carried out in the resonance cell itself, as happens in compact designs, unwanted hyperfine components in the incident light are

† Relaxation by collision is not radiative. The nitrogen pressure is such that the fluorescence light intensity of atoms in the cell is very low. The spectrum of this light would be similar to that shown in Figure 6.32a. Such fluorescence would reduce the efficiency of optical pumping since it would excite atoms in the $F = 2$ level as well as the $F = 1$ level.

progressively attenuated along the light trajectory, so that a detailed treatment of the behaviour of the clock is even more complicated.

6.6.3 Electronics

As the rubidium clock is a passive atomic frequency standard, the atomic resonance can be used to control the frequency of a quartz oscillator in the same way as in a caesium beam clock (Section 6.4.1c). Other electrical circuits produce the radiofrequency signal required to excite the lamp discharge. Separate thermal control of the lamp, the resonance cell and the hyperfine filter (if it exists separately) is essential. This makes it possible to adjust the rubidium density and gas pressure contained in the various parts of the instrument, guaranteeing satisfactory operation of the lamp, a sufficient signal-to-noise ratio to allow detection of the hyperfine resonance, and control over undesirable frequency shifts.

6.6.4 Metrological properties of rubidium cell clocks

(a) Medium term frequency instability

As for any passive frequency standard, an approximation to the medium term frequency instability is provided by (6.63). Here, the noise is shot noise in the electrical current supplied by the photocell. To the nearest power of ten, we have $Q_{at} \sim 10^7$ and $(P_S/P_N)^{1/2} \sim 2 \times 10^3$ in a pass band of 1 Hz, implying $\sigma_y(\tau) \sim 10^{-11}\tau^{-1/2}$.

In practice, medium term frequency instability takes values between $5 \times 10^{-12}\tau^{-1/2}$ and $10^{-10}\tau^{-1/2}$, depending on the manufacturer. For sampling periods τ greater than 10^3 or 10^4 s, $\sigma_y(\tau)$ no longer varies as $\tau^{-1/2}$ because of the influence of systematic effects. Figure 6.11 compares the typical frequency instability of rubidium cell clocks with that of other atomic clocks.

(b) Shifts in the atomic transition frequency

Apart from those frequency shifts common to all atomic clocks, two particularly detrimental frequency shifts are specific to the rubidium cell clock. One is typical of double resonance experiments where two transitions sharing the same energy level are simultaneously excited. It is called the *light shift* because it is associated with optical pumping. The other is due to collisions between rubidium atoms and molecules of the buffer gas, where interaction forces induce a change in the hyperfine splitting.

Light shift. This effect vanishes when the pumping light spectrum is centred on the rubidium absorption line. This condition depends on the composition and content of the lamp and hyperfine filter, if there is one, or the resonance cell, otherwise. In practice, the light shift can never be completely suppressed. It amounts to a few hertz.

Shift due to collisions. This frequency shift is proportional to the buffer gas pressure. In typical conditions, it is of the order of one kilohertz. Its value changes from one clock to the next because it is impossible to control the composition and pressure of the buffer gas accurately enough when the cells are being filled and sealed.

However, rubidium cell clocks deliver a signal with predetermined nominal frequency, e.g., 10 MHz. This is obtained by adjusting the frequency produced by the synthesiser included in the microwave excitation signal generator. Such adjustments amount to attributing different values to the factor K_s introduced in Section 6.4.1c for different clocks of the same make. This operation requires calibration with respect to some reference source. The notion of accuracy thus has no meaning as far as the rubidium cell clock is concerned.

(c) Long term frequency instability and the effect of environmental conditions

In the long term, for τ greater than 10^3–10^4 s, the frequency of a rubidium cell clock drifts by 1–5 $\times$ 10^{-11} per month. The reasons are not yet perfectly understood. It is probably related to chemical reactivity between rubidium and glass, and changes in the buffer gas composition and pressure caused by diffusion through the filter and cell walls.

The rubidium standard is the most sensitive of all atomic time and frequency standards to environmental conditions [6.95]. Effects of temperature variations are amongst the most important, amounting to 10^{-10} per °C. Temperature variations affect the light spectrum emitted by the lamp as well as the efficiency of the ensuing hyperfine filtering, and hence the light shift. They also change the probability of collisions between rubidium atoms and buffer gas molecules in the resonance cell, thereby affecting the associated shift in the hyperfine transition frequency.

Technical choices with regard to composition and pressure of the gas used, as well as the temperatures assigned to the lamp, the hyperfine filter (if separate) and the resonance cell, are largely determined by the need to minimise the effects of environmental conditions. Despite the difficulties involved, rubidium clocks showing very low sensitivity to ambient conditions have been produced. They are carried on board satellites of the Global Positioning System (GPS) (see Section 9.2.2), side by side with caesium clocks.

6.6.5 Current studies and prospects

Theoretical and experimental work has shown that the discharge lamp could usefully be replaced by a semiconducting laser emitting at 780 nm [6.96, 6.97, 6.98]. This type of laser is commonly used in CD players and hence cheap to produce. The quasi-monochromatic light of the laser is stabilised on the rubidium D_2 line, and very efficient optical pumping is possible for a much lower flux than is required when a lamp is used. The relative height of the resonance is increased whilst the total current output by the photocell falls. As a consequence, the potential medium term frequency stability is of the order of $10^{-13}\tau^{-1/2}$. It is also simpler to use lasers, from a technical standpoint, and this promises to improve long term frequency stability.

Laboratory studies have produced rubidium 87 masers operating in active mode [6.99, 6.100]. To this end, the production of a population inversion between levels $F = 2$, $m_F = 0$ and $F = 1$, $m_F = 0$ was optimised, as far as such is possible, and the resonance cell placed in a cavity with the best quality factor that could be attained. The main attraction of a rubidium maser is its excellent frequency stability, of the order of 10^{-13}, for sampling periods τ close to 1 s. However, this stability tends to diminish for $\tau > 1$ s, owing to the light shift and the shift caused by collisions with the buffer gas.

6.7 Stored ion clocks

It has been known for several decades that ions can be confined in vacuum by means of electromagnetic fields. This confinement means that the ions are held within a limited volume of space, *without the help of material walls*, for periods of time as long as one day or more.

In a Penning trap, static electric and magnetic fields are used, whilst in a Paul trap, the field is a radiofrequency electric field. In practice, the presence of a magnetic field, as in the Penning trap, is not an advantage for the applications we have in mind here. We shall therefore restrict discussion to confinement by the Paul trap [6.101].

Many ions exhibit a hyperfine structure. Indeed, for the following ions, for example, the hyperfine splitting in the ground state is known to better than one part in 10^8: $^3He^+$, $^9Be^+$, $^{25}Mg^+$, $^{88}Sr^+$, $^{113}Cd^+$, $^{135}Ba^+$, $^{137}Ba^+$, $^{171}Yb^+$, $^{173}Cd^+$ and $^{199}Hg^+$. In 2000, only the $^{171}Yb^+$ and $^{199}Hg^+$ ions had been used to build instruments that could be considered as atomic time and frequency standards. Since there are thus far more results for mercury ions, we shall concentrate on mercury ion standards, with confinement in a Paul trap. The original idea was put forward by Major at the end of the 1960s, in the NASA

laboratories. The hyperfine resonance of confined $^{199}Hg^{+}$ ions was first observed experimentally at NASA [6.102] and subsequently at the University of Mainz (Germany) [6.103].

The main steps in developing the corresponding frequency standard were carried out at the Laboratoire de l'horloge atomique (France) [6.104], Hewlett-Packard (USA) [6.105, 6.106], and the Jet Propulsion Laboratory (USA) [6.107, 6.108, 6.109]. These instruments can be made small enough for use in the field. Frequency instabilities below $10^{-13}\tau^{-1/2}$ have been achieved. More recently, a laser-cooled mercury ion frequency standard has been successfully operated at the National Institute of Standards and Technology (USA), showing a relative frequency instability of $3.3 \times 10^{-13}\tau^{-1/2}$ and an inaccuracy of 3×10^{-15} [6.110]. This is a more complex and bulky device.

We shall begin by giving the principle of ion confinement in a Paul trap. We then go on to describe the mercury ion frequency standard using a cloud of ions that are thermalised by collision with an inert gas at room temperature and optically pumped by means of an isotopic lamp. We shall briefly discuss the ytterbium ion frequency standard and close by presenting the laser-cooling of ions and the particularities of the laser-cooled mercury ion frequency standard.

6.7.1 Confining ions in a radiofrequency trap

(a) Potential distributions

We begin by considering the two potential distributions used to confine charged particles. They correspond to the linear trap (also used in some gas analysers) and the three-dimensional trap, respectively.

Two-dimensional quadrupole field. The potential distribution described in Cartesian coordinates (x, y, z) by

$$V(x, z) = C(x^2 - z^2) \tag{6.129}$$

can be realised physically since the Laplacian of this potential is indeed zero. The quantity C is a constant. The potential is zero on the y axis. In the (x, z) plane, the equipotential curves are rectangular hyperbolas.

If we introduce conducting electrodes that follow these equipotentials, and raise them to the corresponding potential, the potential distribution described by (6.129) can be created at every point between the electrodes. In particular, if the turning points of the electrodes are equidistant from the y axis, the

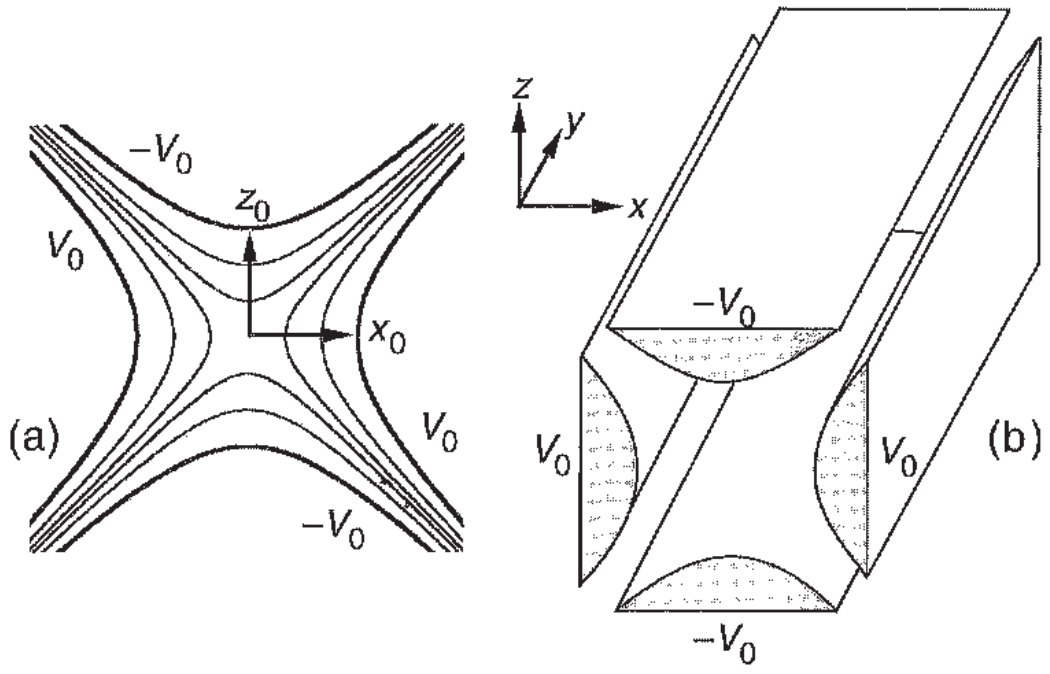


Fig. 6.33. Linear trap (in two dimensions). (a) Equipotentials viewed in a transverse plane. (b) Electrode structure.

configuration shown in Figure 6.33a is obtained, with

$$C = \frac{V_0}{x_0^2} = \frac{V_0}{z_0^2} , \tag{6.130}$$

where V_0 is the absolute value of the potential to which the electrodes are raised and $x_0 = z_0$ is the distance separating the turning points of the electrodes from the y axis.

The components of the electric field between the electrodes are

$$E_x = -\frac{2V_0}{z_0^2}x , \quad E_y = 0 , \quad E_z = \frac{2V_0}{z_0^2}z . \tag{6.131}$$

In practice, the electrodes cannot extend infinitely along the asymptotes of the hyperbolas. They are therefore truncated, as shown in Figure 6.33b. To simplify the construction, they are usually reduced to simple rods of circular cross section. It is then shown that, near the neutral axis, where the ions are in fact confined, the potential distribution tends to the one given by (6.129).

Three-dimensional quadrupole field. The following potential distribution, expressed in cylindrical coordinates,

$$V(r, z) = C(r^2 - 2z^2) , \tag{6.132}$$

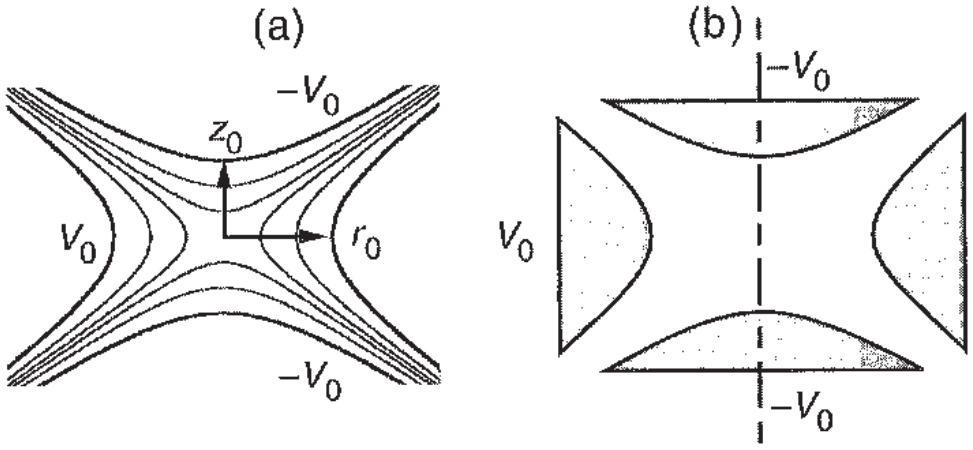


Fig. 6.34. Three-dimensional trap. (a) Equipotentials viewed in a meridian plane. (b) Electrode structure.

can also be realised physically. This potential is zero at the origin and the equipotentials are hyperboloids of revolution about the z axis. The latter comprise a ring-shaped surface centred on the z axis and two other surfaces cut by the z axis, which might be viewed as a kind of lid. The ring component, with minimum distance r_0 from the z axis, and the lids with turning points at distance $z_0 = r_0/\sqrt{2}$ from the centre, have the same absolute value of the potential. If these equipotentials are replaced by material electrodes brought to the potential V with $|V| = V_0$, we obtain the configuration with

$$C = \frac{V_0}{r_0^2} = \frac{V_0}{2z_0^2} , \tag{6.133}$$

as illustrated in Figure 6.34a.

The components of the electric field between the electrodes are

$$E_r = -\frac{V_0}{z_0^2} r , \quad E_z = 2\frac{V_0}{z_0^2} z . \tag{6.134}$$

As in the previous case, the electrodes are actually finite. In practice, they are made as shown schematically in Figure 6.34b. For the sake of convenience, they may not have exactly the theoretical hyperboloid shape. However, close to the centre, the potential is adequately described by (6.132).

(b) Confinement of charged particles

Equations (6.131) and (6.134) show that the orthogonal components of the electric field are proportional to the distance from the axis of the linear trap, or to the centre of the three-dimensional trap, but with opposite signs. If the applied potential difference is constant, a charged particle will therefore be

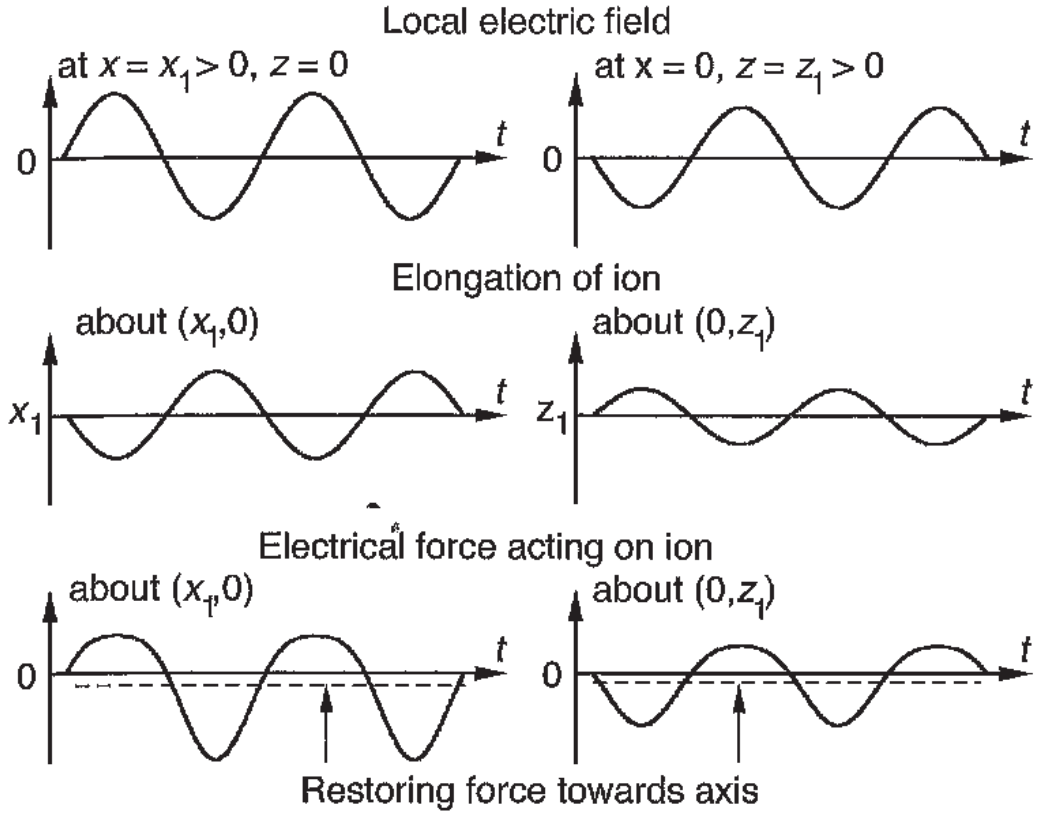


Fig. 6.35. Origin of the restoring force, bringing ions back towards the axis in a linear trap.

subjected to a force with one component tending to restore it towards the axis or the centre, and one component tending to dispel it from the trap.

However, confinement can be achieved if the applied potential difference alternates. The particle motion can, of course, be described analytically via the fundamental equation of dynamics. We thus obtain two Mathieu differential equations. Their solutions are stable for certain values of parameters defined in terms of the amplitude and frequency of the potential difference applied across the electrodes, the trap dimensions r_0 and z_0, and the mass and charge of the particle. In this case, the particle oscillates about the axis or the centre, without reaching the walls. Confinement is thus physically realised. A complete study of the motion of the ions shows that stability conditions are optimised if the applied potential difference includes a d.c. component [6.101].

In the following, we present a simple qualitative argument to explain the origin of the *restoring force* that confines the ions in both directions.

Consider a linear trap and, without loss of generality, imagine that a positive ion is placed in it. The electric field forces the ion into a low amplitude sinusoidal motion called the *micromotion*. It has the same frequency as the electric field, but it is completely out of phase with it. Figure 6.35 shows the time variation of the electric field at two points in the (x, z) plane located on the x and

z axis, with coordinates ($x_1 > 0$, 0) and (0, $z_1 > 0$), respectively. The electric fields there are out of phase since, by (6.131), E_x and E_z have opposite signs. The micromotion executed locally by the ion is also shown. This motion allows it to explore the electric field. However, the latter is non-uniform, being proportional to the distance from the axis, and the electrical force on the ion is thereby distorted, as shown in the figure. Consequently, the average value of the force is not zero and we observe that it is directed towards the axis. This restoring effect on the ion forces it into a two-dimensional transverse harmonic motion about the trap axis. This is called the *macromotion* or *secular motion*. It has a low frequency compared with the electric field itself. Longitudinal confinement is arranged by the end electrodes, positioned on the axis and held at a constant repulsive potential.

By a similar explanation, it can be shown that, in a three-dimensional trap, confinement is obtained in the axial direction and in every radial direction. In this case, confinement is indeed three-dimensional.

Applying a potential difference with an alternating component across the electrodes of a linear or three-dimensional trap, ions can be confined in the vicinity of the axis or centre, respectively. The ions are suspended in the vacuum under the effects of forces of electrical origin that bind them with respect to the trap electrodes. They follow an extensive secular motion, on which is superposed a higher frequency but lower amplitude micromotion. For a trap in which the electrode spacing is of the order of one centimetre, and the applied alternating potential difference has frequency around 100 kHz and amplitude around 100 V, the frequency of the secular motion is of the order of 10 kHz in the case of mercury ions, and the amplitude is less than the internal dimensions of the trap.

By their almost harmonic motion, the ions behave as though they were moving in a parabolic potential well with depth of the order of 10 V.

6.7.2 Confinement of a cloud of ions

(a) Ion cloud density

The density of the ion cloud that can be confined in this way is extremely low, due to repulsive electrostatic forces between the charged particles. This density is of the order of 10^6 ions cm^{-3}, corresponding to a partial pressure in the region of 10^{-8} Pa, comparable with that of the residual gases in an extremely high vacuum. The ionic medium is thus very dilute and collision interactions scarce.

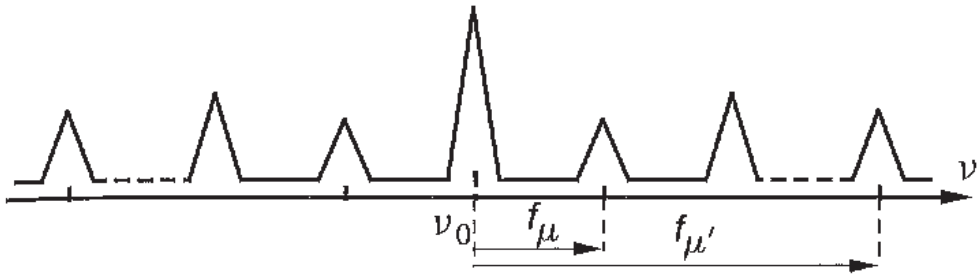


Fig. 6.36. Schematic illustration of the absorption spectrum of an ion confined in a radiofrequency trap.

(b) Absorption spectrum of the ions

When the ions are irradiated by an electromagnetic wave of fixed frequency ν, they feel a frequency modulated wave. This is due to their own motion and the associated Doppler effect. The spectrum of the signal received by the ions thus includes a carrier at frequency ν and lateral lines at frequencies $\nu \pm nf_\mu$ and $\nu \pm mf_{\mu'}$, where n and m are integers and f_μ, $f_{\mu'}$ are the secular motion and micromotion frequencies, respectively. In addition, the ions exhibit a very sharp hyperfine resonance, due to the fact that confinement times can be extremely long. Its width is much less than f_μ and hence also $f_{\mu'}$. If the frequency ν is varied, the ions tend to absorb the incident radiation each time one of the spectral components coincides, up to the width of the resonance, with the hyperfine resonance frequency ν_0.

Switching from the standpoint of the ions to that of the observer, everything happens as though the ionic absorption (or emission) spectrum were composed of very narrow lines at frequencies $\nu_0 \pm nf_\mu$ and $\nu_0 \pm mf_{\mu'}$. The spectrum is shown schematically in Figure 6.36.

When the radiation frequency ν is very close to the hyperfine transition frequency ν_0, the lateral lines are not excited and everything happens as though the Doppler effect were inoperative. This result has already been mentioned in Section 6.2.3. It can be shown that the Dicke condition (6.15), limiting the dimensions of the region occupied by the ions, implies that the central component of the absorption spectrum should be barely attenuated and that the lateral components should be of low amplitude.

(c) Ion energies

The resonance signal that can be obtained from a group of ions or atoms is proportional to the number of particles that can be observed. Efforts are therefore made to increase this number, insofar as this is possible. However, in the case of ions, there is a correlation between the number of ions confined and their kinetic energy. The latter is an increasing function of this number. The

same is therefore true of the second order Doppler frequency shift given by (6.16).

It is easy to find a qualitative explanation for this coupling. Because of Coulomb repulsion, the ionic cloud occupies an ever greater volume as the number of ions increases. Since the electric force acting on the ions is proportional to their distance from the trap axis or centre (by (6.131) and (6.134)), their kinetic energy increases. For the three-dimensional trap, the average kinetic energy is around 2 eV. Despite the large mass of the mercury ion, this leads to a second order Doppler frequency shift of the order of -10^{-11}.

The kinetic energy of the ions is reduced by introducing helium gas at very low pressure, of the order of 10^{-3} Pa, so as not to cause a further frequency shift that might actually worsen the situation. The average kinetic energy of this gas is 4×10^{-2} eV at room temperature. This *collisional cooling method* is not very effective in practice. The second order Doppler shift can be reduced to -5×10^{-12}.

The situation is better in a linear trap because the ions spread out over a cylindrical region about the trap axis. For a given number of ions, the transverse extent of the cloud can be decreased if the cylinder length is greater than its diameter. The average kinetic energy is thereby diminished [6.107, 6.108, 6.109]. Using helium gas at a partial pressure of 10^{-3} Pa, the second order Doppler shift falls to around -10^{-12}. It can also be shown that, in a linear trap, the average kinetic energy is less sensitive to fluctuations in the a.c. potential difference applied across the electrodes.

These advantages explain the current preference for using linear traps in confined mercury ion frequency standards.

6.7.3 Description of a frequency standard using a cloud of mercury ions

(a) Spectroscopic properties of $^{199}Hg^+$

The nuclear spin of the 199 isotope of mercury is 1/2. In addition, when the Hg^+ is produced by removing an electron from the outer electron shell, it has an electronic structure very similar to the alkali metals, with an unpaired electron on the outermost electron shell.

The $^{199}Hg^+$ ion therefore has a ground state hyperfine structure as simple as the hydrogen atom itself, with just four hyperfine sublevels, viz., $F = 0$, $m_F = 0$ and $F = 1$, $m_F = -1, 0, 1$. The proportion of levels useful for operating a frequency standard, i.e., the levels (0, 0) and (1, 0), is therefore as large as possible. This makes it easier to observe the hyperfine resonance signal of a limited number of ions. Note also that the hyperfine

splitting is large, corresponding to about 40.5 GHz. For a given hyperfine resonance width, the quality factor of this resonance will thus be higher than for the atoms considered previously. This same property also favours frequency stability.

(b) Optical pumping of $^{199}Hg^+$ and observation of the hyperfine resonance

The hyperfine states of an ion cannot be separated by magnetic selection because the deflecting force associated with the electrical charge is far greater than that arising from the magnetic moment of the unpaired electron. State selection must therefore be carried out by optical pumping.

In the present case, there is a fortunate isotopic coincidence between the absorption spectrum of $^{199}Hg^+$ and the emission spectrum of $^{202}Hg^+$. The latter spectrum is emitted by a lamp containing the 202 isotope of mercury and excited by a radiofrequency discharge. This isotope has no hyperfine structure and it just happens that the wavelength of the D_1 line of $^{202}Hg^+$, near 194 nm, lies very close to the wavelength of the transition between the $F = 1$ hyperfine level of the $^{199}Hg^+$ ground state $^2S_{1/2}$ and its first excited state $^2P_{1/2}$. The situation is thus very similar to the one illustrated in Figure 6.4a, except that the incident light excites ions in level β (here, the $F = 1$ level) and optical pumping populates level α (here, the $F = 0$ level).

Fluorescence light emitted by ions when they spontaneously decay from the first excited level down to the ground state is not very intense because the population of the $F = 1$ level is reduced. If the ions are now irradiated by an electromagnetic wave with frequency close to the hyperfine transition frequency, the $F = 1$ level repopulates and the fluorescence light at 194 nm intensifies. Its intensity is measured and, when the applied signal frequency sweeps across the hyperfine transition frequency ν_0, a resonance curve is obtained with a maximum at that frequency.

If optical pumping were accomplished using monochromatic light, it would be more efficient and the resonance would be more intense. However, at the time of writing (2000), there is no far ultraviolet continuous-wave laser. Coherent radiation has been produced at wavelengths as short as 194 nm using optical frequency synthesis techniques, i.e. by multiplying and mixing frequencies in non-linear crystals [6.111]. The experimental setup is complex and costly, so the $^{202}Hg^+$ lamp is still used in relatively compact devices.

(c) Experimental setup and resonance curve

The experimental setup, designed by the Jet Propulsion Laboratory (USA), is shown schematically in Figure 6.37.

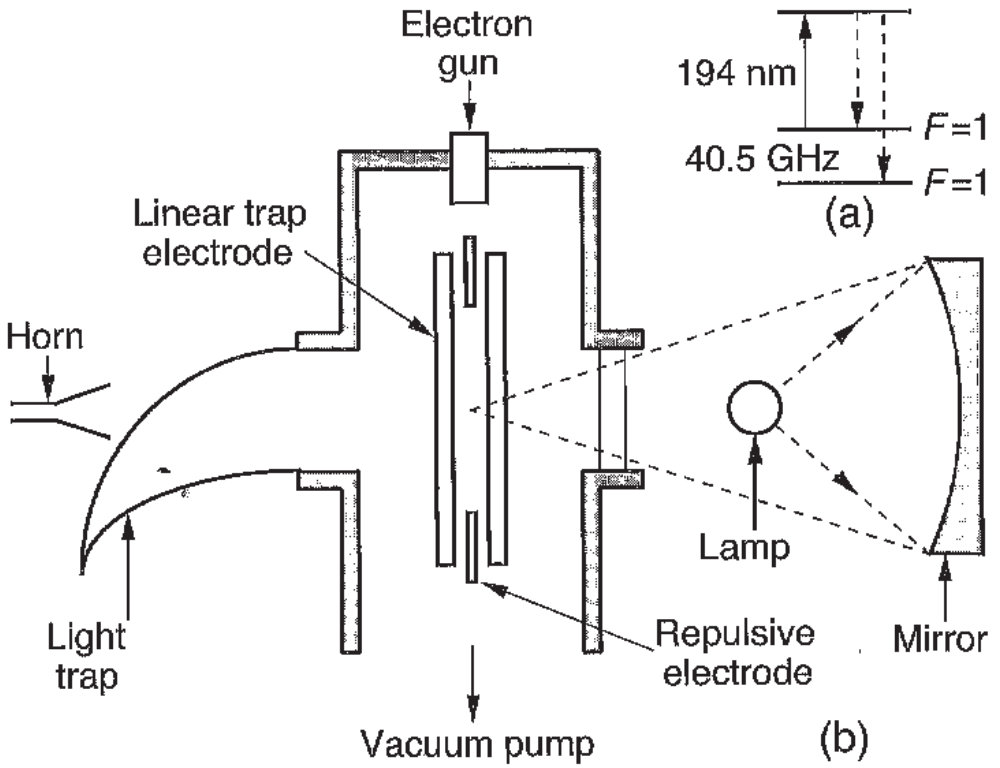


Fig. 6.37. Confined mercury ion resonator. (a) Useful part of the $^{199}Hg^{+}$ spectrum. (b) Simplified representation of the confined mercury ion resonator. Fluorescence light is observed at right angles to the plane of the figure.

A neutral mercury 199 vapour is introduced in a controlled manner, at very low pressure, by heating mercury oxide. The vapour encounters a blast of electrons from an electron gun. The ions thereby produced are confined by applying a suitable potential difference to the linear trap electrodes. The latter consist of four regularly spaced cylindrical rods, about 1 cm apart. Helium gas at a pressure of about 10^{-3} Pa reduces the kinetic energy of the ions.

Light from the $^{202}Hg^{+}$ lamp is directed towards the ion cloud. After crossing the trap, it is collected in a light trap in order to reduce the intensity of spurious light that might impede observation of fluorescence light from the ions. The fluorescence light is focused onto a photomultiplier, in a direction perpendicular to the plane of the figure.

The microwave interrogation signal at 40.5 GHz is applied using a horn. The electromagnetic wave propagates at right angles to the trap axis.

Magnetic shielding attenuates the Earth's magnetic field and its fluctuations. It contains coils creating a uniform and stable magnetic field, with strength similar to that in a caesium beam clock. A high vacuum pump and various other devices are incorporated.

The resonance is observed with the help of Ramsey's method. In this case,

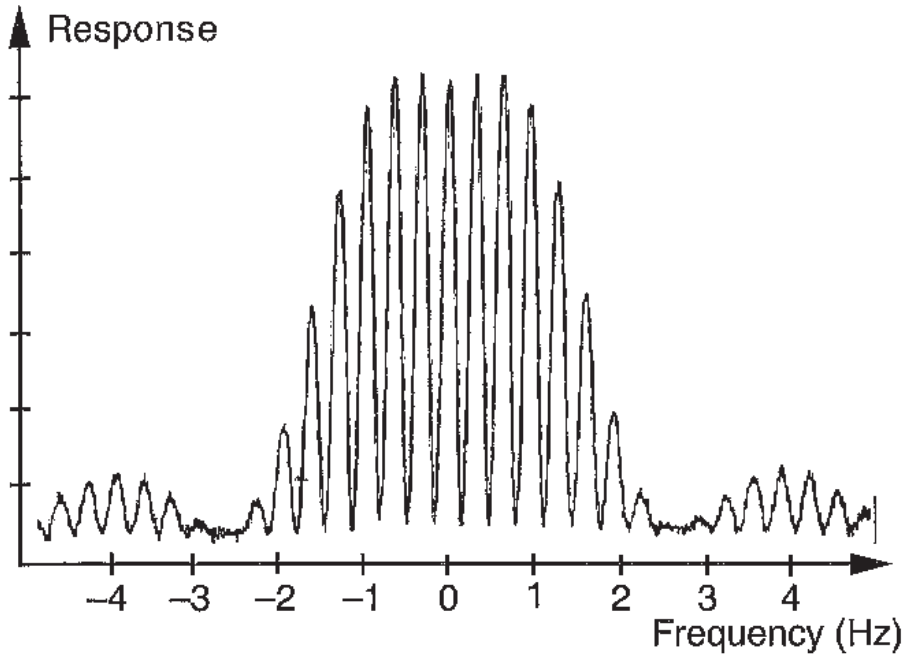


Fig. 6.38. Response of a linear mercury ion trap. The frequency origin is positioned at the ground state hyperfine transition frequency. From [6.107, 6.108, 6.109].

however, the two successive interactions do not result from motion of the atoms (or ions), as in a caesium clock. In the $^{199}Hg^{+}$ frequency standard, two successive interactions are obtained by pulsing the oscillating field. During an observation sequence, this field irradiates the region occupied by the ions for two brief instants separated by a time interval T. The Ramsey method therefore operates in the time dimension here.

An observation sequence typically begins by reloading the trap in order to maintain the number of ions (and hence their energy) at a constant level from one cycle to the next. The electron gun is therefore activated briefly. The lamp is then lit for about 1.5 s, by which time the ions have been prepared in the state $F = 0$. The microwave radiation is applied for two time intervals τ lasting about 0.3 s, separated by a time T, anywhere between 2 and 30 s, during which irradiation ceases. The lamp is subsequently relit and the fluorescence light intensity measured, to discover the effect that the microwave radiation has had. Figure 6.38 shows the resonance curve thus obtained, with $T \doteq 2.5$ s [6.107, 6.108, 6.109]. It represents the transition probability as described by (6.43).

The quality factor Q_{at} of the central fringe of the resonance curve in Figure 6.38 is 2×10^{11}. It can be increased, in proportion to T. As a general rule, the temporal Ramsey method applied to confined ions is currently the method of high resolution spectroscopy that produces the highest quality factors for atomic resonances.

(d) Controlling the frequency of an oscillator

The response of the atomic system can only be observed at discrete times, at the end of an interrogation sequence. The frequency of a quartz oscillator is thus controlled by the ionic hyperfine resonance in the same way as it would be by a cold caesium atom fountain.

6.7.4 Metrological properties of frequency standards using a cloud of mercury ions

(a) Medium term frequency instability

When sampling periods are long compared with the time constant for servo control of the local oscillator, and hence longer than T, the medium term frequency instability of the confined ion frequency standard is characterised by (6.86).

Detection noise arises from shot noise in the detected light flux. There are two components in the detected flux, viz., fluorescence light from ions and light produced by unwanted reflection of the incident light. In practice the latter is more intense than the fluorescence light. This introduces an inordinately large shot noise. However, the high value of the quality factor Q_{at} favours frequency stability. An Allan deviation of $7 \times 10^{-14}\tau^{-1/2}$ has been obtained, putting the mercury ion standard into the ranks of the most stable, as can be seen from Figure 6.11 [6.107, 6.108, 6.109].

(b) Residual frequency shifts

The main frequency shifts affecting the mercury ion frequency standard come from the quadratic Zeeman effect, the second order Doppler effect, and collisions with the helium buffer gas.

The static magnetic field is typically 3.5×10^{-6} T. The shift in the hyperfine transition frequency of the $^{199}Hg^{+}$ ground state is then equal to 3×10^{-12}. It is much smaller than in a caesium beam clock, where it would have a value around 1.7×10^{-10}.

The second order Doppler shift and the shift due to collisions with the buffer gas are related since it is the density of the buffer gas that determines the kinetic energy of the ions. For a helium partial pressure of 10^{-3} Pa, the frequency shift due to these two associated effects is close to 10^{-12}.

Note that the shift arising from the quadratic Stark effect in the presence of the confining electric field is as low as a few times 10^{-14}.

(c) Long term frequency instability and inaccuracy

The long term frequency instability depends on the sensitivity of the observed transition frequency to environmental conditions and operating parameters. These instabilities may arise from fluctuations in the surrounding magnetic field, the potential difference across the electrodes, the number of confined ions, and the temperature and pressure of the buffer gas.

Experience shows that these parameters can be sufficiently well controlled to ensure a frequency instability of the order of 10^{-15} for $10^4 < \tau < 10^6$ s [6.107, 6.108, 6.109].

Theoretical and experimental analyses of the various frequency shifts provide an estimate of a few times 10^{-13} for the inaccuracy of the confined $^{199}Hg^+$ ion time and frequency standard [6.107, 6.108, 6.109].

6.7.5 Ytterbium ion frequency standard

The ion $^{171}Yb^+$ has been used to make a confined ion frequency standard. Its nuclear spin is also 1/2, and its hyperfine transition frequency in the ground state is close to 12.6 GHz. Optical pumping is carried out on the D_1 line at 369.5 nm. Standards have been constructed in different ways. In one case, the radiofrequency trap is three-dimensional and the light source provided by a dye laser with doubled frequency [6.112]. In another case, the radiofrequency trap is linear and the pumping light obtained from a titanium–sapphire laser [6.113].

The measured frequency instability is close to $10^{-13}\tau^{-1/2}$ for $100 < \tau < 3000$ s.

The ground state hyperfine transition frequency for $^{171}Yb^+$ has been measured to within 2 mHz [6.113].

6.7.6 Laser-cooled mercury ion frequency standard

Ions can be cooled by laser irradiation in the same way as atoms [6.114]. In fact, the experimental method is simpler for ions than for atoms because they are held close to the trap centre or axis. The first radiative cooling of confined ions was accomplished in 1978 at the National Bureau of Standards, since renamed the National Institute of Science and Technology (NIST).

The cooling mechanism can be understood intuitively. The optical absorption spectrum of a confined ion has lateral lines due to the periodic component of its motion, in exactly the same way as its microwave absorption spectrum, shown in Figure 6.36. The ion can therefore absorb a photon when illuminated by a laser beam with frequency lower than the optical transition frequency. By

fluorescence, it then emits a photon whose mean frequency is equal to the optical transition frequency. Each absorption–fluorescence cycle thereby removes energy from the kinetic energy of the ion. A single laser suffices to achieve this result, and residual thermal energies corresponding to a temperature of about 10 mK are commonly obtained. The second order Doppler shift is then reduced to a completely negligible level, of the order of 10^{-17} or 10^{-18}.

Such low temperatures can only be attained if the number of confined ions is small. They are then localised very close to the trap axis where they form a linear crystal of a few ions. Since the ions lie along the nodal line of the radiofrequency field, the second order Stark effect becomes negligible. Unfortunately, frequency stability is reduced by the small number of ions. This is counterbalanced by a reduction in the hyperfine resonance line width, thanks to an increase of up to 100 s in the Ramsey interaction time T. The frequency shift due to collisions with residual gases must be reduced as far as possible by rarefaction, and this requires the use of cryogenic techniques.

A laser-cooled mercury ion frequency standard has been built and operated at NIST according to the principles just described [6.110]. Coherent radiation is produced at 194 nm by optical frequency synthesis [6.111]. It is used to cool the ions and also to create a population difference between the hyperfine sublevels. Ions are stored in a 4 mm long linear radiofrequency trap with an electrode separation close to 1 mm. Under the reported operating conditions, a set of seven cooled ions form a linear crystal along the trap axis. The trap is placed in an enclosure cooled to the temperature of liquid helium. The clock transition is probed using the Ramsey technique in the time domain. During one elementary cycle of operation, the ions are first cooled for 300 ms. They are then optically pumped in the $F = 0$, $m_F = 0$ level of the ground state. The light is switched off and the Ramsey microwave interrogation is accomplished. For this purpose, two 40.5 . . . GHz pulses of duration 250 ms are applied with a separation of up to 100 s. The clock transition is then detected optically. The frequency control of the quartz crystal oscillator is very similar to that implemented in atomic fountains. The frequency instability and inaccuracy achieved are $3.3 \times 10^{-13}\tau^{-1/2}$ and 3.5×10^{-15}, respectively [6.110]. Improved performance is expected by increasing the number of stored ions or by preparing the ions in special quantum states [6.115, 6.116].

6.8 Other atomic frequency standards

Hyperfine structure transitions have no exclusive claim as atomic frequency references. Other transitions may be acceptable, provided they have the required qualities.

6.8.1 Magnesium beam frequency standard

The alkaline earth atoms have no hyperfine structure in the ground state. However, there are interesting transitions between levels of their excited states. For example, in magnesium, the frequency of one of these transitions is 601 GHz and the lifetime of the levels is such that the quality factor is close to 10^{10}. A beam frequency standard has been built on the basis of this transition and results are encouraging [6.117]. A more efficient preparation of the relevant atomic states, together with cooling of the magnesium atoms, would be required to improve this device.

6.8.2 Optical frequency standards

A great many frequency standards have been constructed, or are under development, in the optical region of the spectrum. They use atomic or molecular transitions between the ground state and an excited state, or between two excited states. The light source used to observe such a transition is always a laser, with frequency controlled by the frequency of the selected transition. These are therefore passive frequency standards. The references are transitions in the following molecules, atoms and ions, e.g., I_2, CH_4, CO_2, OsO_4, Rb, Ca, Hg^+, Yb^+, Ba^+, Sr^+, Ca^+, In^+. The whole available range of CW lasers is concerned here, including He–Ne, CO_2, semiconductor, neodymium-doped YAG, titanium–sapphire, and so on.

Remarkable results have been obtained. At frequency 29 THz (10 μm), the CO_2 laser stabilised on a transition of the osmium tetroxide molecule OsO_4 has a frequency instability below 10^{-14} for $10\ \text{s} < \tau < 10^4$ s, a long term (1 yr) reproducibility of 1.3×10^{-13}, and a potential inaccuracy of the order of 10^{-14} [6.118]. Similar results have been obtained at 88 THz (3.39 μm) with the He–Ne laser stabilised on a transition of the methane molecule [6.119].

6.8.3 Optical frequency standards based on individual ions at rest

The ion confinement technique is likely to play a key role in the development of extremely high performance optical frequency standards. Indeed, in 1973, Dehmelt put forward the idea of using a *single ion at rest* [6.120].

Consider an ion in which two transitions with very different properties can be identified. One couples the ground state to a level of very short lifetime, of the order of 10 ns. This means that about 10^8 absorption–fluorescence cycles can be carried out per second and the ion can thus be very efficiently cooled. The other transition is associated with a very low probability. It connects the

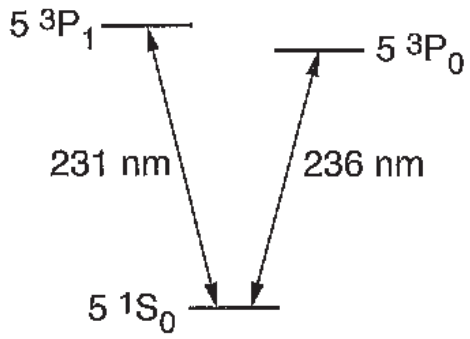


Fig. 6.39. Energy levels in the indium ion chosen to build an optical atomic frequency standard.

ground state to a metastable state with a lifetime of about 1 s. The corresponding resonance therefore exhibits an intrinsic width of the order of 1 Hz. It constitutes the optical reference frequency with a quality factor somewhere between 10^{14} and 10^{15}. Since the cooling transition provides something like 10^8 fluorescence photons per second, the presence of a single ion in the trap is easy to detect. Indeed, this has been demonstrated experimentally in several laboratories.

A single cold ion resides on the axis of a linear Paul trap, with an extremely low kinetic energy. We thus achieve a good approximation to the idea of a single atom (or ion) at rest. In such a situation, the causes of unwanted frequency shifts are reduced to an absolute minimum.

The possibility of using the indium ion $^{115}In^+$ has been considered [6.121, 6.122]. The relevant energy levels are shown in Figure 6.39. The $^3P_0$ level is the one with a long lifetime. This ion has no electron angular momentum in the ground state $^1S_0$, nor in the metastable state $^3P_0$. Hence, the Zeeman frequency shift is extremely small. Moreover, the reference transition frequency is not shifted by the Stark effect, and it is in fact completely immune to the presence of a confining electric field. The production of two emissions at 231 and 236 nm appears feasible by twice doubling the emission frequency of semiconductor lasers.

The reference transition can be detected by the so-called *shelving method*, which works as follows [6.120]. If the ion is in the ground state $^1S_0$, it is detected as a result of the 231 nm fluorescence light released after irradiation by the cooling light beam. On the other hand, if the frequency of the radiation at 236 nm is tuned to the reference transition frequency, the ion is carried into the state $^3P_0$, and the fluorescence light disappears.

An excellent frequency stability is expected of this frequency standard, because of the high value of the quality factor and the high signal-to-noise ratio for detecting the transition. The inaccuracy may be as small as 10^{-17}, be-

cause the frequency shifts affecting the reference transition are significantly reduced.

Convincing experimental results have been obtained recently at NIST, using a single laser-cooled $^{199}Hg^{+}$ ion stored in a cryogenic linear radiofrequency trap [6.123]. Ion cooling is achieved by inducing transitions at 194 nm between the $^{2}S_{1/2}$ ground state and the short-lived $^{2}P_{1/2}$ first excited state. The 282 nm reference transition occurs between the ground state and the metastable $^{2}D_{5/2}$ state having a lifetime of 90 ms. The shelving method is implemented and the detection of the 282 nm transition is accomplished by observing the fluorescence light at 194 nm. A line width of 6.7 Hz for the 282 nm transition (1.06×10^{15} Hz) has been observed. The corresponding line quality factor is equal to 1.6×10^{14}. It is the highest ever achieved in microwave and optical spectroscopy. This result proves the true potentiality of optical frequency standards to compete successfully with frequency standards based on microwave transitions.

It should be noted, however, that in order to observe a resonance in the optical region with width as small as 1 Hz, the light source itself must have a spectral width well below this value. Specific studies will be needed to construct such a source.

6.8.4 Optical frequency standards for time measurements

The frequency of radiation at a wavelength of 0.6 μm, for example, is 500×10^{12} Hz. This is about 5000 times higher than frequencies in the most advanced systems ($\sim$ 100 GHz), used for telecommunications and radar detection. Although several specialised laboratories are able to compare the frequencies of a microwave signal and those of an optical signal, using the frequency synthesis technique, the experimental setups are delicate, cumbersome and costly. At the time of writing, no practical way has yet been discovered for using optical frequency standards to produce a reliable and durable time scale.

Research is in progress with regard to improving this situation. New results have opened the way to much easier comparisons of frequencies in the microwave and optical domains [6.124, 6.125]. They have been obtained using a mode-locked laser which regularly emits very narrow (a few femtoseconds) and intense pulses. This periodic radiation creates a frequency comb in the optical domain. The repetition frequency of the laser, and hence the frequency of each component of the comb, can be controlled by a microwave frequency standard.

Table 6.2. *Practical information concerning atomic time and frequency standards. Numerical values indicate orders of magnitude. They correspond to different performance levels or degrees of development.*

Type of standard	Volume (dm^3)	Mass (kg)	Cost (10^3 Euro)
Laboratory caesium	500–3000	500–1000	1000
Commercial caesium	10–30	10–30	50–80
Active hydrogen maser	150–250	75–250	150–450
Passive hydrogen maser	20–80	20–60	75–150
Rubidium	0.2–1	0.4–1.5	3–8
Mercury ion	10–250	10–250	150–2000

Table 6.3. *Typical inaccuracies for various types of atomic frequency standard, as of 2000.*

Type of standard	Inaccuracy
Laboratory caesium: cold atom fountain	1×10^{-15}
Laboratory caesium: continuous beam from oven	6×10^{-15}–10^{-13}
Commercial caesium	1.5×10^{-12}–3×10^{-11}
Hydrogen maser	2–5×10^{-12}
Trapped mercury ion	3.5×10^{-15}–2×10^{-13}
Rubidium	No meaning (calibration required)

6.9 Conclusion

In this chapter, we have described the state-of-the-art as far as atomic time and frequency standards are concerned. Table 6.2 gives some details concerning the volume, mass and cost of such instruments.

Figure 6.11 gives an overview of current frequency instabilities. The other important characteristic is inaccuracy, summarised in Table 6.3.

We have also mentioned some current developments. It may be hoped that, within a short space of time, standards will be produced with frequency insta-

bilities of 10^{-16} for $\tau = 1$ day, and inaccuracies approaching 10^{-16}. There are possibilities for further improvements, too. Referring to Figure 4.3, which shows schematically how the quality of artificial clocks has evolved, we may thus fully justify an optimistic extrapolation beyond the year 2000.

7

Atomic time measurement

The scientific and technical work that serves as a foundation to atomic time measurement has been described in the previous chapters. Chapter 4 showed how such measurements have gradually displaced astronomical measurements. In the present chapter, after recalling the atomic definition of the second and defining the main time scales issuing from it, we shall be concerned with the way atomic time measurements are put to use. We describe all the steps involved in setting up and disseminating a worldwide time scale that can be used both for synchronisation and as a way of providing a unit of time. We also discuss time scales suitable for dynamical purposes.

7.1 Definitions

7.1.1 Definition of the second

Let us recall the definition adopted by the 13th General Conference on Weights and Measures (CGPM) in 1967:

> The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom.

As explained in Chapter 3, the second should be considered as the unit of proper time along the worldline of the relevant caesium atom. Since the atoms are moving relative to the device that observes their frequency, i.e., the clock, corrections must be made to allow for this motion in order to ascertain the proper time of the clock. At present levels of accuracy, the only relativistic effect that needs to be taken into account is the second order Doppler effect of special relativity. However, the time may soon come when the whole clock

will need to be treated within the framework of general relativity, in order to provide the proper time at a specified point, such as a connector.

If the clock is in the immediate vicinity of, and fixed with respect to the observer, it can be used as in classical physics, provided that we accept Einstein's principle of equivalence. If it is at some distance, carried by a satellite, for example, and communicates data by radio, then this data must be processed as indicated in Chapter 5.

7.1.2 International Atomic Time and its relationship with geocentric and barycentric coordinate times

As explained in Chapter 3 (Section 3.3.2e), International Atomic Time (TAI) can be viewed as an average of clock readings. The clocks in question here operate in accordance with the definition of the second. They are assumed fixed relative to the Earth and located on the rotating geoid. These definitions specify the *scale unit* of TAI, but the whole system of graduations remains to be fitted together somehow. This problem is referred to as *fixing an origin*. The task was accomplished by assigning the same date in TAI and Universal Time (UT) to a fictitious event, occurring on 1 January 1958 at 0 h UT. (To be precise, the form UT2 of Universal Time was used, as defined in Section 8.1.3.) UT is based on the Earth's rotation and is not measured as accurately as TAI. Moreover, it was subsequently recognised that systematic errors were made when determining UT in 1958, and UT values had to be corrected by several milliseconds. However, these corrections were not carried over to TAI. We may therefore say that the origin of TAI is simply fixed by the very existence of this time scale, the coincidence with UT on 1 January 1958 being merely approximate.

In Chapter 3, we defined the geocentric coordinate time. The scale unit for this ideal time is obtained theoretically from the definition of the second. However, an origin must also be fixed in this case. In 1991, the International Astronomical Union (IAU) did just this in its Resolution A4, by defining a particular geocentric coordinate time denoted TCG such that

$$TCG - TAI = 32.184\ \text{s} \quad \text{on 1 January 1977, 0 h TAI}\,. \tag{7.1}$$

The number 32.184 appearing here should be treated as exact by convention, as should the date. Note, however, that there is an uncertainty in the TAI reading, something like 1 μs at that epoch, so that TCG can be said to float by this amount. In practice, this has little importance.

We also saw in Chapter 3 that a variant geocentric coordinate time was defined having a scale unit with duration very close to the duration of the (proper)

second on the rotating geoid. This variant is Terrestrial Time (TT), related to TCG by the expression for the difference in seconds,

$$TCG - TT = L_G \times 86\,400 \times \Delta D \,, \tag{7.2}$$

where ΔD is the time elapsed in TAI days since 1 January 1977 at 0 h TAI and L_G has a conventional value†

$$L_G = 6.969\,290\,134 \times 10^{-10} \,. \tag{7.3}$$

TAI is thus a realisation of TT, up to the 32.184 s offset. We may therefore write

$$TT(TAI) = TAI + 32.184\,\mathrm{s} \,, \tag{7.4}$$

adopting the commonly used notation whereby $U(x)$ represents the realisation of the theoretical quantity U identified by the symbol x. Hence, TT(TAI) should read 'realisation of TT based upon TAI'.

The reader may wonder why there is a difference of 32.184 s between TT and TAI. In 1958, it was justifiable to fix TAI onto UT rather than Ephemeris Time (ET), because the latter was rather poorly determined. Later, however, astronomers wanted TT to replace ET. All attempts either to shift TAI or to change the argument of the astronomical ephemerides failed. The number 32.184 represents the best estimate of the deviation in seconds between ET and TAI on 1 January 1977.

Let us now consider the barycentric coordinate time mentioned in Section 3.3.2b. This is used to study planetary motions and the trajectories of space probes. The word 'barycentric' is intended to mean 'centred on the centre of mass of the Solar System'. As for TCG, the scale unit is based on the (proper) second. The origin of one particular barycentric coordinate time denoted TCB is such that TCB has the same reading as TCG on 1 January 1977 at 0 h TAI at the centre of mass of the Earth. It is essential to specify 'at the centre of mass of the Earth' because the relationship between these two coordinate times is four-dimensional. To order c^{-2}, it is given by‡

$$TCB - TCG = \frac{1}{c^2}\left\{\int_{t_0}^{t}\left[\frac{v_E^2}{2} + U_P(\mathbf{x}_E)\right]\mathrm{d}t + \mathbf{v}_E\cdot(\mathbf{x}-\mathbf{x}_E)\right\} \,, \tag{7.5}$$

where $\mathbf{x}_E$ and $\mathbf{v}_E$ denote here the barycentric position and velocity of the centre of mass of the Earth, whilst $\mathbf{x}$ is the barycentric position of the observer and

† Relation (7.2) is approximative. According to decisions taken by the IAU in 2000, ΔD should be reckoned in TCG days since 1 January 1977 at 0 h TAI. The difference is of order 10^{-18} in relative frequency. In time, it amounts to 3 ns in one century.

‡ A more accurate expression is provided by IAU Resolution B1-5 (2000).

U_P the Newtonian gravitational potential due to all the bodies in the Solar System except for the Earth itself. In the integral, $t = TCB$ and t_0 is chosen in accordance with the origins indicated previously. On the Earth's surface, the last term reaches a maximum value of 2.1 μs.

For terrestrial observers, the integral in (7.5) gives rise to an increment proportional to the time, rising at a rate of about 0.5 s per year, to which periodic terms must be added. The largest of the latter is due to the fact that the Earth's orbit is elliptical. It is thus annual, with half amplitude equal to 1.7 ms. To reach an accuracy of 100 ns, more than one hundred periodic terms must be taken into account [7.1].

Towards 1970, it became necessary to use a relativistic model to calculate the ephemerides of bodies in the Solar System. In 1976, the IAU defined the temporal argument of the ephemerides by carrying out a scale unit transformation on the barycentric coordinate time in such a way as to leave only the periodic terms in the difference with TT. In 1979, the new time scale was given the name TDB (Temps dynamique barycentrique). At this time, some astronomers had raised doubts about general relativity, so that all reference to this theory and its terminology was scrupulously avoided. The relation between TDB and TCB, in seconds, is given by

$$TCB - TDB = L_B \times 86\,400 \times \Delta D \,, \tag{7.6}$$

where ΔD has the same meaning as in (7.2) and L_B is currently estimated to be

$$L_B = 1.550\,519\,768 \times 10^{-8} \pm 2 \times 10^{-17} \,. \tag{7.7}$$

Like TT, which was given the name TDT (Temps dynamique terrestre) between 1979 and 1991, both TCB and TDB can be represented on the basis of TAI, with the help of relations (7.2), (7.5) and (7.6). The use of TDB instead of TCB gives rise to theoretical complications. However, a great deal of work has been carried out taking TDB as time argument, particularly the construction of the Jet Propulsion Laboratory ephemerides, which are used for planetary exploration. For this reason, it is difficult to make the desired transition to TCB.

Figure 7.1 sums up the relationships between the various coordinate time scales.

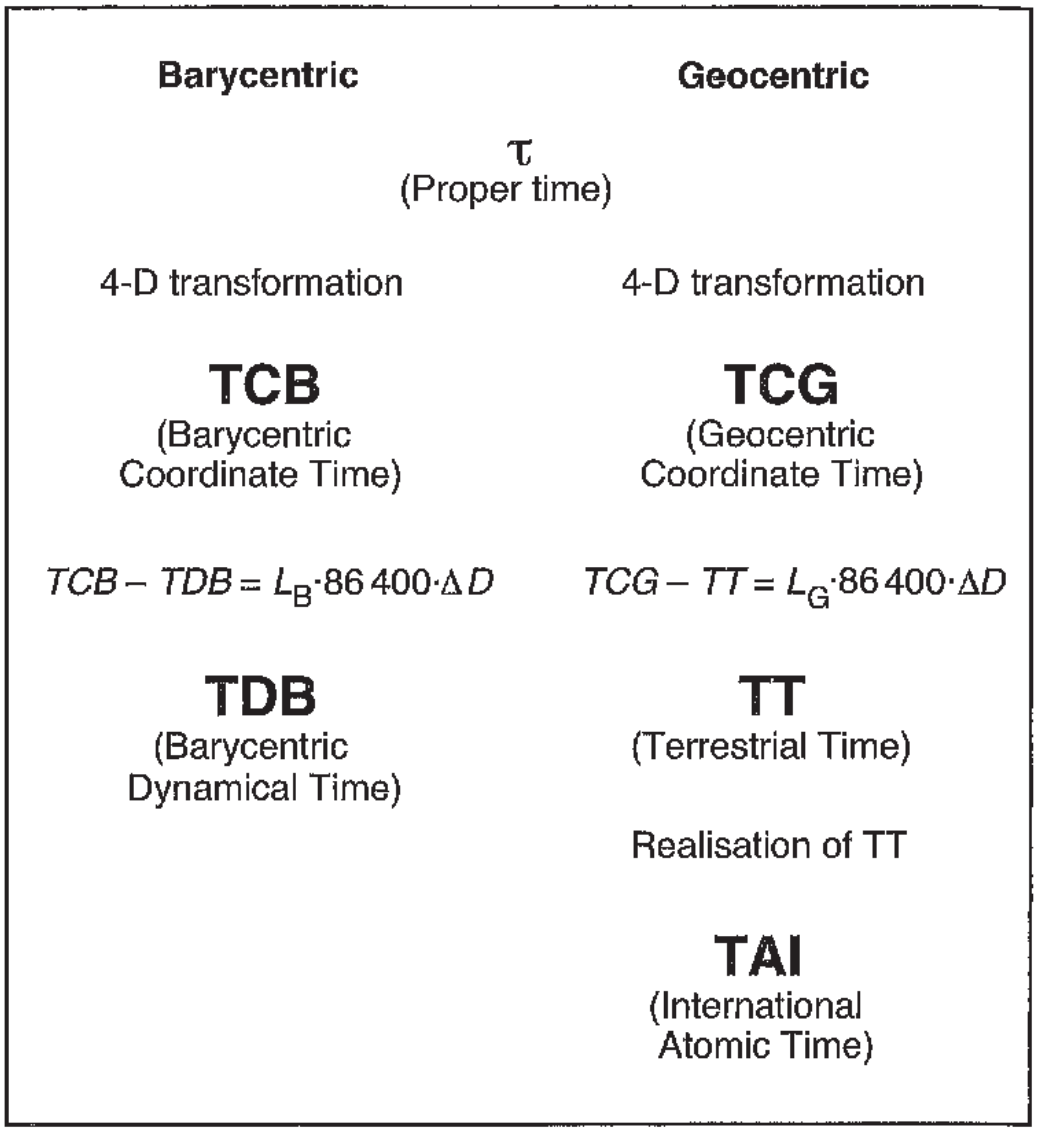


Fig. 7.1. Relationships between coordinate time scales starting from the proper time τ of an observer. Barycentric and geocentric channels are related by four-dimensional transformations, so that TAI can be used to realise all theoretical times. Quantities L_B, L_G and ΔD are defined in the text.

7.2 Establishing International Atomic Time

7.2.1 Time scale algorithms

When several atomic clocks are available, their readings can be used to set up a single time scale. This is certainly more reliable than time scales based on the contributing clocks taken separately, and hopefully has better characteristics than any of them.

If the clocks are sufficiently local, in the sense that they are not too widely separated, they can be used to establish the proper time of their laboratory without it being necessary to take into account any other relativistic effects than the second order Doppler effect from special relativity. If the clocks are spread across a country or around the world, a realisation of TT is usually constructed.

This is what happens, for example, when TAI is established. In this case, the proper time τ_C of each clock C must be transformed into an approximation t_C of TT, before the data can be combined as would have been done to establish a local time scale. This transformation introduces an arbitrary additive constant, but we shall see that it presents no obstacle. In order to simplify notation, t_C will be written t_C throughout the present chapter.

Before setting up an algorithm for some time scale that we intend to construct, at least the following set of questions should be asked:

- Over what sampling period do we wish to optimise frequency stability?
- Do we wish to impose a certain frequency accuracy, and if so, what would be an acceptable compromise in the case where this affects stability?
- What should be the interval between clock comparison dates (the data acquisition period)?
- How long is the interval between the date at which the calculated scale becomes available and the last date for which this scale has been calculated?
- How often is the calculation to be updated?

The various time lapses involved are shown in Figure 7.2.

Answers to these questions are interdependent. For example, if we seek long term frequency stability, we can accept a longer interval between clock comparisons than when short term stability is required. A wide range of situations nevertheless remains open. In some applications, the scale must be obtained almost in real time, whilst for TAI, a two month delivery time is acceptable. For the study of pulsars, where the highest possible quality is sought, nobody would complain about waiting a whole year. It is not therefore possible to describe a typical algorithm. Besides, algorithms must yield to technical constraints, such as the invention of new clocks or new methods of time comparison. All algorithms nevertheless exhibit certain common features, although they may be hidden to some extent by the mathematical formalism. These can be exposed by examining the algorithm set up by the BIH in 1973 and still used by the BIPM to establish TAI, apart from a few adaptations to accommodate technical advances [7.2, 7.3].

7.2.2 Algorithm for International Atomic Time

(a) General organisation

International Atomic Time is currently based upon data from over 200 atomic clocks and frequency standards in about sixty different laboratories and compared via the Global Positioning System, its Russian counterpart GLONASS,

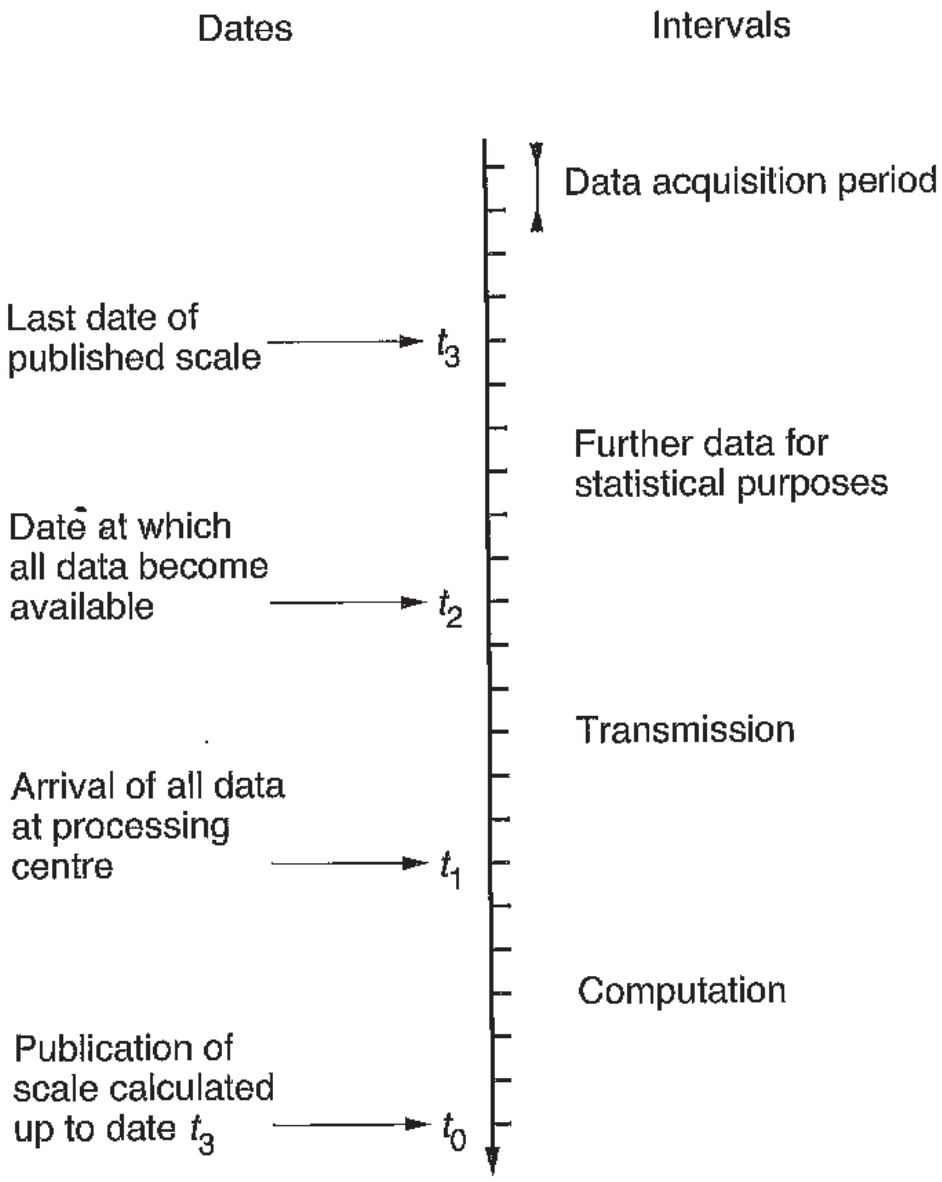


Fig. 7.2. Schedule for carrying out time scale algorithms. Data comprises either clock time differences, or other information that can be used to calculate such differences (e.g., reception dates of GPS signals).

and the Two-Way Satellite Time and Frequency Transfer (TWSTFT) method (see Section 5.6.3).

It is set up in such a way as to favour long term frequency stability, over periods of a month or more, and frequency accuracy. Because of the long term stability, it becomes acceptable to allow a rather long time for publication of TAI, between one and two months. Indeed, time differences between local clocks and TAI can be extrapolated to yield an approximation of TAI in real time.

TAI is established in two stages. A *stability algorithm* produces what is known as the *Echelle atomique libre* (EAL), a free-running mean time scale whose stability is optimal for the relevant sampling period. Then a long term servo control based on the frequency of primary standards constructs TAI by

applying a time dependent correction $[TAI - EAL](TAI)$ to the EAL. This so-called *steering* aims to ensure the frequency accuracy of TAI.

The schedule is as follows (see Figure 7.2).

- Computation period: one month. The date t_0 is around the middle of month m.
- Data acquisition rate: every five days (at 0 h UTC on dates ending in 4 or 9 when expressed in Modified Julian Date (MJD)). Data was previously acquired every 10 days, up to the end of 1995.
- Last date covered by the publication: t_3 somewhere around the end of month $m-1$.
- There is almost no further data (i.e., $t_2 - t_3 = 0$).
- Transmission takes a few days (including preparation of data in participating laboratories). This involves a certain level of commitment on the part of the various laboratories and the BIPM.

(b) Stability algorithm

We begin by considering an imaginary situation in which all participating clocks I (I = J, K, etc.) are read simultaneously, according to the convention of coordinate simultaneity, at the date t in geocentric coordinate time (TCG, TT or TAI). The readings $\tau_I(t)$ are transformed into $t_I(t)$, introducing arbitrary integration constants. Letting $\{\}(t)$ denote a weighted mean over the clock ensemble at the date t, EAL could be defined by

$$[EAL](t) = \{t_I\}(t) , \tag{7.8}$$

assigning statistical weights in accordance with the chosen stability criteria.

Definition (7.8) can serve the required purpose as long as all the clocks continue to operate, weights are not modified, and new clocks are not introduced. This is summarised by saying that the clock ensemble must remain invariable. If the ensemble is altered in some way, it is clear that this will cause a time step to occur, since the clocks do not all have the same reading, and it will also cause a frequency jump, since the clocks do not all have the same frequency. In order to avoid this situation, definition (7.8) is replaced by

$$[EAL](t) = \{t_I\}(t) + A + B(t - t_0) , \tag{7.9}$$

where t_0 is some arbitrarily chosen origin. A and B are constants which are modified when the clock ensemble is altered, in such a way as to maintain time and frequency continuity. Let t_m be the date at which some modification is made. Let us denote the old values of A and B by A_- and B_-, and the new

values by A_+ and B_+. The requirement of time continuity is expressed by

$$A_+ = A_- + \{t_I\}_-(t_m) - \{t_I\}_+(t_m) - (B_+ - B_-)(t_m - t_0)\ , \qquad (7.10)$$

the ensemble averages $\{t_I\}_-(t_m)$ and $\{t_I\}_+(t_m)$ being calculated at t_m for the old and new ensembles, respectively. Consider the case where one clock has stopped running. Then t_m is the last date for which that clock still gave data. It is therefore included in $\{t_I\}_-(t_m)$ but not in $\{t_I\}_+(t_m)$. As far as continuity of frequency is concerned, this imposes

$$B_+ = B_- + \left\{\frac{dt_I}{dt}\right\}_- - \left\{\frac{dt_I}{dt}\right\}_+\ , \qquad (7.11)$$

where derivatives are calculated at t_m.

Whereas it is a straightforward and rigorous matter to calculate A_+ once B_+ has been found, it is actually impossible to obtain B_+ from (7.11) for the simple reason that we cannot evaluate instantaneous frequencies. (We can change origin, choosing $t_0 = t_m$. The Bs are then not needed to calculate A_+.) We must work with mean frequency values. Before t_m, a duration must be chosen over which to calculate these mean frequencies. After t_m, no data is available if we wish to produce the time scale in real or near real time. Consequently, we must *predict* the frequency of each clock on the basis of its past behaviour.

A subtle problem arises when we consider (7.11), because the derivatives are with respect to the unknown ideal time t. However, only a difference between two such derivatives is required. This means that t can be replaced by a physically realised time scale, provided that its frequency is stable enough and lies close enough to that of t. In practice, the EAL itself can fulfill this purpose. Hence, using past frequency differences with respect to the EAL for each clock, we attempt to predict future values of this difference.

Frequency prediction and weighting form the core part of time scale algorithms. Another important task must be added, namely, the detection of anomalous behaviour in one of the clocks, which leads to its being withdrawn from the ensemble. Such statistical problems are challenging. The type of solution brought to bear can significantly influence the quality of the scale that is finally produced.

We now transform the above equations so that they can be applied to real measurements. These measurements are proper time differences between clocks, obtained by the methods described in Chapter 5. Let us recall that these are differences in readings at the same date t, but that, since they vary only slightly with time, they can be dated by a reference scale θ, for which there remains a great freedom in the choice. In practice, Coordinated Universal Time (UTC) is used. Starting with the proper time differences for two clocks J and

K, we calculate the differences in the coordinate times t_J and t_K, fixing the constants of integration in an arbitrary manner. This gives

$$[t_J - t_K](\theta) = \delta_{JK}(\theta) , \tag{7.12}$$

where δ_{JK} is the measured value. Putting

$$x_I(\theta) = [EAL - t_I](\theta) , \tag{7.13}$$

equations (7.9) and (7.12) become

$$\{x_I\}(\theta) = A + B(\theta - \theta_0) , \tag{7.14}$$

$$x_J(\theta) - x_K(\theta) = -\delta_{JK}(\theta) , \tag{7.15}$$

where the factor $t - t_0$ is replaced by $\theta - \theta_0$, since B is extremely small (of the order of 10^{-13}). The time comparison network is organised in such a way that the system formed by (7.14) and equations of type (7.15), written for various pairs of clocks, can be solved exactly, because the processing of superfluous links brings with it more drawbacks than advantages. When the clock ensemble changes, A and B are reestablished using

$$A_+ = A_- - \{x_I\}_-(\theta_m) + \{x_I\}_+(\theta_m) - (B_+ - B_-)(\theta_m - \theta_0) \tag{7.16}$$

and

$$B_+ = B_- - \left\{\frac{dx_I}{dt}\right\}_- + \left\{\frac{dx_I}{dt}\right\}_+ , \tag{7.17}$$

where, as before, t is replaced by EAL in derivatives.

Having obtained the solution for x_I, it is transformed to EAL $- \tau_I$ using the theoretical relation between τ_I and t_I, with the same value for the constant of integration as for the inverse transformation. EAL is then available at the location of the participating clocks through a correction to their readings.

When calculations are made, we use the fact that all clocks currently participating in the construction of the EAL are fixed on the Earth. The transition from their proper time to coordinate time only involves a fixed frequency shift. This shift is not applied explicitly. It is contained in the correction term B. It follows that the stability algorithm provides EAL $- \tau_I$ directly. However, if we wanted to incorporate satellite-borne clocks, a full treatment would be needed. The previously noted simplification whereby θ_0 is made to coincide with θ_m, leads to a change in θ_0 each time the ensemble is modified.

(c) Frequency steering

A stability algorithm like the one just described optimises stability over a sampling period that has been fixed by choosing the way frequencies are predicted. In its application to the EAL, we shall see that this optimal value is somewhere between 30 and 60 days. However, the algorithm contains no guarantee against longer term fluctuations. Step by step adjustment of the Bs may even lead to an unbounded frequency drift.

The first problem to solve is how to evaluate the frequency of the EAL, using data from primary standards. In this discussion, all frequencies are referred to the level of the rotating geoid.

A primary standard provides an estimate of the mean frequency of the EAL between dates θ_i and θ_f. As time goes by, whilst this calibration becomes part of the past, the EAL retains a recollection of it, thanks to its frequency stability. Nevertheless, this recollection is not perfect and a contribution from the frequency instability of the EAL must be added to intrinsic uncertainties in the standard.

A *filter* has been devised which yields an optimal value for the mean frequency of the EAL over a given interval by assigning coefficients (with sum 1) to all available calibrations, whether they come before or after. The word 'filter' is used, rather than weighting, because some of the coefficients may be negative. This filter takes into account the following factors:

- a statistical model for frequency noise in the EAL, sum of a white phase noise (arising from time comparisons), and white, flicker and random walk frequency noise (due to the clocks);
- random and systematic uncertainties in frequency calibrations;
- initial and final dates of each calibration and of the period of estimation.

The theory behind this is rather complex and will not be discussed here. An account can be found in [7.4].

Steering consists in finding a function $P(\theta)$ allowing us to deduce TAI from the EAL by

$$TAI(\theta) = EAL(\theta) + P(\theta) \,, \tag{7.18}$$

in such a way that the TAI frequency remains close to the results of the accuracy filter, without its stability being degraded with respect to that of the EAL. For example, if the TAI frequency were immediately changed each time a new calibration occurred, frequency jumps would be unacceptable. To avoid this situation, the frequency instability of $P(\theta)$ is arranged to be smaller than the estimated frequency instability of the EAL.

(d) Application

The following is a highly schematic description, since steady advances are being made in the design of atomic clocks and time comparisons, and the BIPM algorithms must have a certain level of flexibility in order to adapt. Our discussion refers to the situation at the end of 2000. However, the initial assessment of the role to be played by TAI, as carried out by the BIH in 1973, has been regularly confirmed by international organisations. For this reason, the basic principles have not changed since then.

When calculations are carried out, whole seconds are omitted. They are fed back in when the results are obtained. We may thus consider that the following explanations apply equally to either TAI or UTC.

Time comparison network. Let k denote a laboratory contributing to TAI. Each laboratory k has a *master clock* which supplies an approximation UTC(k) to Coordinated Universal Time UTC. This clock serves as a reference for all local dating procedures. In particular, it measures the time differences $UTC(k) - \tau_C$ for all clocks C of laboratory k. The BIPM only uses values measured every five days on *standard dates*. These are defined to be Modified Julian Dates (MJD) ending with a 4 or a 9, at 0 h UTC.

Remote time comparisons concern the UTC(k). Although continuous, the BIPM estimates them on standard dates by averaging or filtering. Figure 7.3 shows the network of time transfer links. A few laboratories act as centres for links up to about 1000 km, using GPS or GLONASS. These centres are linked together by the best available techniques, namely, TWSTFT and GPS, using for the latter precise ephemerides, computed retrospectively, and measured values for ionospheric corrections. In certain countries, there is a further level in the hierarchy, whereby a central laboratory provides the link with several national laboratories. This has long been the case in France, for example.

Stability algorithm. The stability algorithm, called Algos, calculates the values of $EAL - UTC(k)$ at the standard dates in one month batches. The initial and final dates of these batches are the last standard date in the month. (The duration of one batch is normally 30 days, but it is sometimes 35 or 25 days.) Quite arbitrarily, these beginning and end dates of the batches have been chosen as ensemble modification dates, i.e., the previously defined θ_m. Consequently, clock weights are held constant over each monthly interval. A clock weight is zero when a clock has provided incomplete data, either because it has stopped or because it has entered the ensemble after the beginning of the interval.

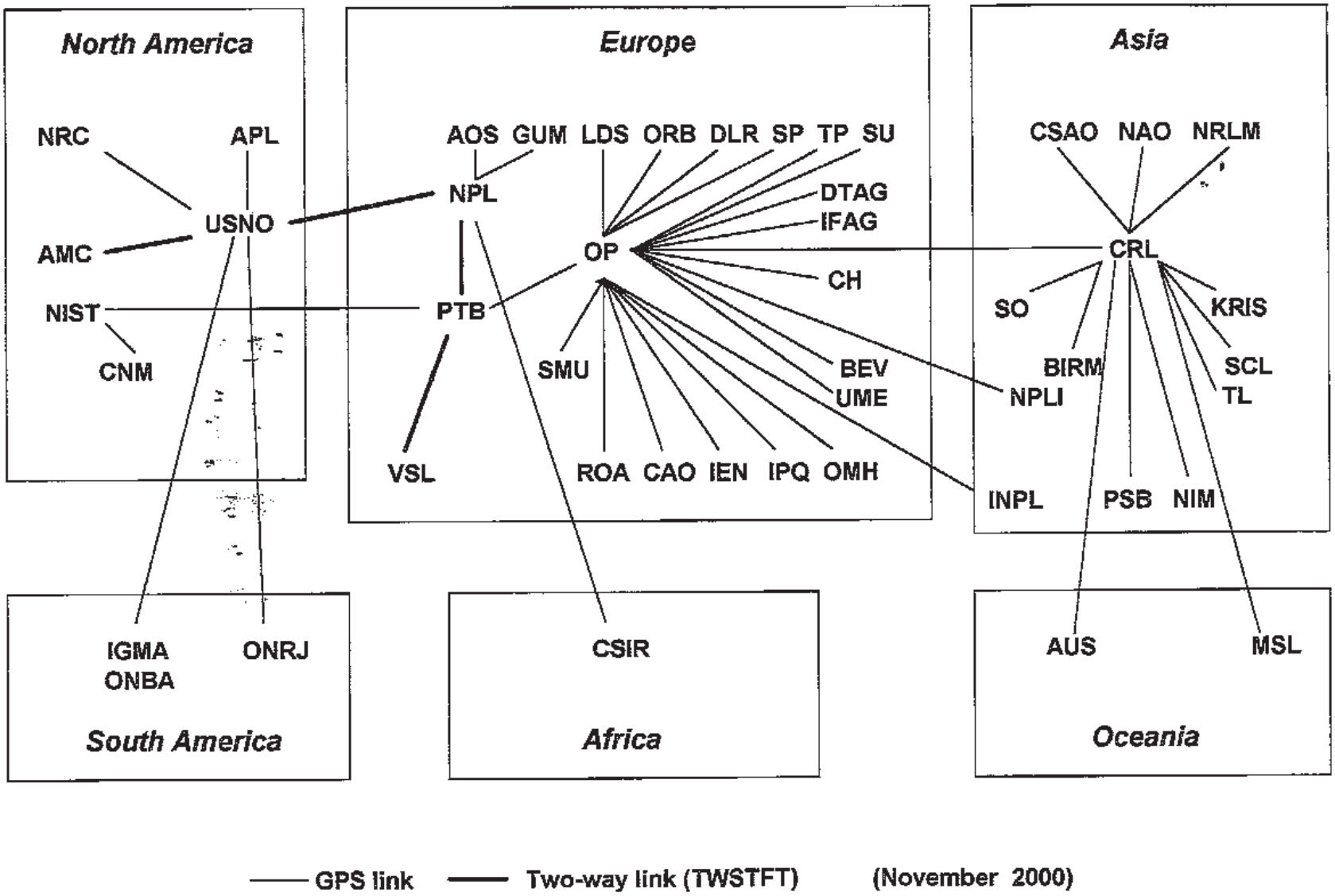


Fig. 7.3. Time transfer links for establishing TAI (as of November 2000). Acronyms are given in Appendix 1.

At the present time (2000), Algos processes the data from more than 200 clocks, mainly commercially manufactured caesium clocks which have excellent long term stability. Since only stability is relevant here, other types of clock can contribute. There are about forty hydrogen masers and their number now remains stable.

The ensemble average is a weighted mean. The weight of a clock C is proportional to the reciprocal of the variance $\sigma_C^2(12, \tau)$ of 12 mean frequency samples taken over one month, relative to the EAL.† This variance has been chosen, rather than the usual two-sample variance $\sigma_{y,C}^2(\tau)$, in order to reduce the weight of clocks exhibiting frequency drift, or annual frequency fluctuations, as has sometimes been observed. In the flicker noise frequency model, which is suited to the values of τ considered,

$$\sigma_C^2(12, \tau) = 2\sigma_{y,C}^2(\tau) \,. \tag{7.19}$$

As frequency fluctuations of clocks are evaluated with respect to the EAL to which they contribute, some clocks may progressively gain an excessive statistical weight and unduly dominate the ensemble. The problem is avoided by using an upper limit for weighting. This has been modified several times, in order to cope with clock improvements. Since January 1998, the rule is that no clock can receive a weight exceeding 0.7% of the total weight. However, with the increasing role played by the latest commercial caesium clocks, this rule is becoming unsatisfactory because it does not sufficiently discriminate between these clocks. A relative upper limit L in the form L/N is currently under study (November 2000), where N is the total number of clocks and L lies between 2 and 3.

In order to detect anomalous behaviour in a clock, a comparison is made between its mean frequency with respect to the EAL over the last month, that is, over the month for which the TAI is currently being calculated, and its mean frequency with respect to the EAL over the preceding 11 months. If the variation is more than three times what might be expected from random fluctuations, the weight of the clock is set to zero. All clocks undergo this test. The calculation is repeated once offending clocks have been eliminated. Four iterations are applied.

Frequencies are predicted by simple continuation of the observed frequency, always relative to the EAL, over the preceding month. Many studies of real data have failed to yield a better prediction for the aim in question.

Any new clock remains under observation for at least four months before effectively taking part in TAI. Over this period, it enters the calculations,

† In the variance, through long-established usage, τ represents the duration over which the mean frequency is calculated, i.e., the sampling period, as discussed in Section 5.2.3.

but with zero weight. When the number of frequency samples n is less than 12, $\sigma_C^2(12, \tau)$ is estimated from $\sigma_C^2(n, \tau)$, assuming a flicker noise frequency model.

With the Algos algorithm, it is possible to apply different weighting and prediction rules according to the type of clock. For example, it would be feasible to take into account a frequency drift in the hydrogen masers. Up to now, this option has not been implemented.

As an example, consider the laboratories and clocks taking part over the period 28 August 2000 to 27 September 2000. The BIPM received data from 221 clocks (179 caesium clocks and 42 hydrogen masers) spread over 39 laboratories or national institutes, the latter incorporating several laboratories each. 121 clocks were assigned the maximal weight, including 98 of the latest commercial caesium clocks, 4 primary caesium clocks (built in research centres), and 17 hydrogen masers. Zero weights were attributed to 44 clocks, mainly new ones in their probationary period. Finally, 56 other clocks received intermediate weights, generally rather low.

Clocks assigned the maximal weight constitute 85% of the total, and this proportion is too high, as noted above.

Frequency steering. In 1969, using the Loran-C navigation system, the BIH was able to drop the mean frequency approach to calculating TAI (then known as TA(BIH)), replacing it by one averaging over time. However, it did not have access to data from individual clocks. TAI was an average of atomic time scales set up by a small number of large national timing centres. It was calculated by means of an algorithm similar to Algos [7.5]. Originally, the correction B in the defining equation of type (7.9) was zero for the three participating timing centres, viz., the United States Naval Observatory (USNO), the Physikalisch-Technische Bundesanstalt (PTB) in Germany, and the Commission nationale de l'heure (F) at the Paris Observatory. No attempt was made to connect frequencies with previous values of TAI, nor to align with data from primary standards, since such data was too sparse and too uncertain.

In 1973, having organised direct access to individual clock data, the BIH implemented the Algos algorithm. Care was taken to avoid any frequency jump by suitably adjusting the initial value of B. TAI was produced directly by Algos, in other words, it was just equal to the EAL. However, primary standards had seen much progress and showed right from the start that the TAI frequency was too high by about 10×10^{-13}.

This frequency error was corrected on 1 January 1977 by a jump of exactly 10×10^{-13} in the TAI frequency, following a resolution adopted by the IAU in 1976.

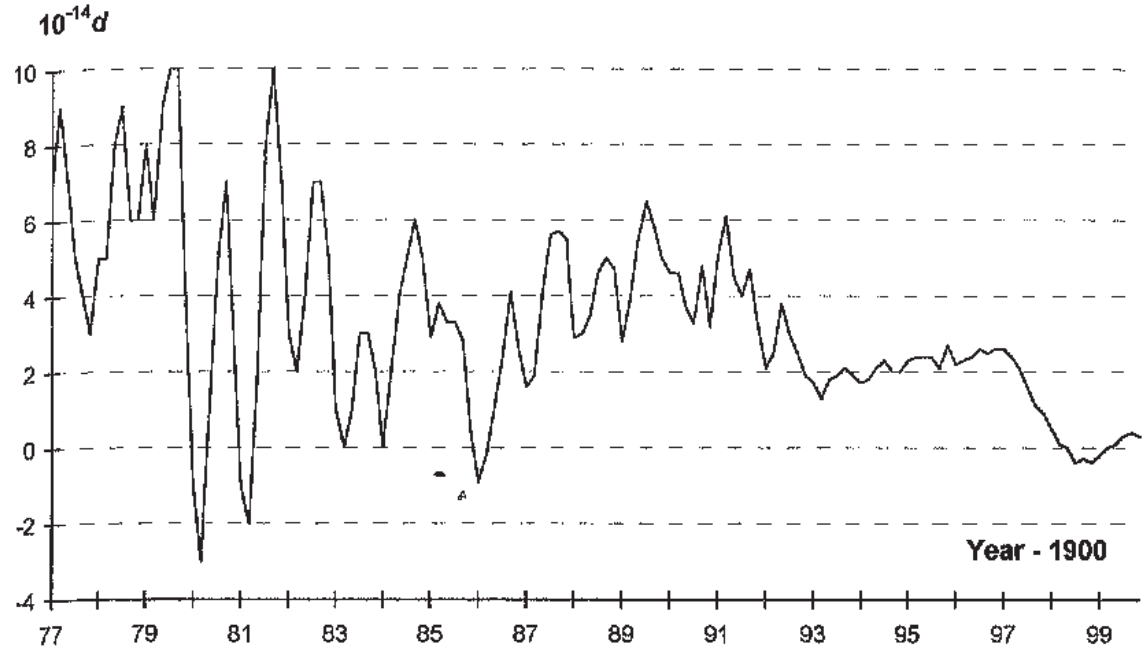


Fig. 7.4. Duration $(1 + d)$ s of the TAI scale unit on the rotating geoid (bimonthly values).

Having done this, it became possible to introduce frequency steering, as had already been envisaged in 1974 by the Consultative Committee for the Definition of the Second (CCDS). Application had been postponed because the frequency error to be compensated was too great. At the beginning, the function $P(\theta)$ was a succession of linear functions, with frequency jumps of 2×10^{-14} occurring, when needed, at intervals of more than two months. The size of the discontinuities was reduced in several stages to reach 10^{-15} in 1996, and these jumps may be made at monthly intervals. Figure 7.4 shows the bimonthly mean values of the duration of the TAI scale unit in seconds on the rotating geoid. The offset of about 2×10^{-14} prior to 1996 results from the omission of a correction for blackbody radiation (see Section 6.4.1e). The decision to apply this correction to all primary standards was taken in March 1996. The frequency offset has been gradually reduced by steering.

Whereas Algos optimises stability over one or two months, steering guarantees stability over much longer periods. It can be considered as a way of introducing a new weighting. As the sampling period increases, each clock contributing to the EAL affects TAI less and less, whilst primary standards have more and more relative importance. For stability over about 5 years, all the weighting is carried by primary standards.

Form of results. After steering corrections, TAI is known from the values of $TAI - UTC(k)$ at standard dates. In fact, the BIPM publishes the values of

$UTC - UTC(k)$ every month in its Circular T, available by electronic mail. An example is shown in Figure 7.5. Apart from this, by tracking GPS and GLONASS satellites for time comparisons, values of $UTC - GPS\ Time$ and $UTC - GLONASS\ Time$ are provided with similar uncertainties to those in $UTC - UTC(k)$. Timing centres k, and GPS and GLONASS satellites form the network of primary access points to TAI and UTC.

It is important to note that TAI and UTC become definitive once published. Small corrections have sometimes been allowed (amounting to a few nanoseconds), but never any global revisions. Even these small corrections are now avoided. It follows that, if errors, from whatever source, are not detected during calculations, they are incorporated forever into TAI. The recognition and elimination of these errors are a key task for Algos and preprocessing programmes, to which only a few days can be devoted.

7.3 Properties of TAI and UTC

7.3.1 Reliability

Reliability is guaranteed by the international basis of TAI. Imagine for a moment what would happen if, just as a practical joke, someone found a way to stop all atomic clocks, just for a short time. This would cause such a tremendous disturbance in world affairs that the loss of TAI would be a totally insignificant matter! Furthermore, when it came to setting it up again, the phase of TAI could be retrieved to within a few tenths of a microsecond by observations of rapidly rotating pulsars (see Chapter 8).

7.3.2 Accuracy of readings

The uncertainty in TAI readings using values of $UTC - UTC(k)$ in one of the primary access laboratories k is approximately equal to the uncertainty introduced by this laboratory in time comparisons used to form TAI. It represents the smallest possible uncertainty when dating an event with TAI in the laboratory under consideration. In the best equipped laboratories, this uncertainty is estimated (in 2000) at about ten nanoseconds, taking uncertainties in instrumental delays into account.

Access afforded to TAI by receiving GPS signals and using values of $UTC - GPS\ Time$ introduces uncertainties of the same order, in a laboratory possessing the best available equipment, by averaging over a great many trackings for at least one day. The same can be said for GLONASS.

Circular T 154 (2000 November 17)

1 - Coordinated Universal Time UTC. Computed values of UTC-UTC(k).
(From 1999 January 1, 0h UTC, TAI-UTC = 32 s)

Date 2000 0h UTC		Sep 27	Oct 2	Oct 7	Oct 12
MJD		51814	51819	51824	51829
Laboratory k		UTC-UTC(k) (Unit is one nanosecond)			
AOS	(Borowiec)	-2757	-2793	-2736	-2659
APL	(Laurel)	6676	6716	6740	6771
AUS	(Sydney)	284	271	249	253
BEV	(Wien)	-383	-372	-371	-371
BIRM	(Beijing)	222	213	200	194
CAO	(Cagliari)	-3306	-3284	-3292	-3290
CH	(Bern)	-95	-115	-139	-142
CNM	(Queretaro)	-157	-161	-163	-171
CRL	(Tokyo)	-44	-50	-53	-50
CSAO	(Lintong)	-23	-26	-8	2
CSIR	(Pretoria)	-526	-591	-646	-704
DLR	(Oberpfaffenhofen)	-13618	-13790	-13974	-14149
DTAG	(Darmstadt)	-695	-701	-681	-681
GUM	(Warszawa)	-759	-767	-770	-765
IEN	(Torino)	59	55	58	55
IFAG	(Wettzell)	-3186	-3213	-3236	-3247
IGMA	(Buenos Aires)	119	112	108	104
INPL	(Jerusalem) (1)	-81	-88	-97	-105
IPQ	(Monte de Caparica)	4734	4787	4842	4891
JATC	(Lintong)	-4130	-4171	-4211	-4257
KRIS	(Taejon)	13	13	2	26
LDS	(Leeds)	-	-	-	-
MSL	(Lower Hutt)	10612	10694	10839	10949
NAO	(Mizusawa)	-2316	-2292	-2269	-2229
NIM	(Beijing)	-2709	-2717	-2722	-2721
NIST	(Boulder)	-7	-8	-8	-13
NPL	(Teddington)	1	2	4	3
NPLI	(New-Delhi)	7953	7981	-	-
NRC	(Ottawa)	17	20	24	25
NRLM	(Tsukuba)	3514	3546	3563	3597
OMH	(Budapest)	4993	5044	5054	5051
ONBA	(Buenos Aires)	17	-47	-52	-55
ONRJ	(Rio de Janeiro)	2904	2912	2922	2934
OP	(Paris)	-39	-35	-39	-38
ORB	(Bruxelles)	31	37	41	36
PSB	(Singapore)	1999	2023	2028	2062
PTB	(Braunschweig)	11	10	18	20
ROA	(San Fernando)	-35	-38	-41	-43
SCL	(Hong Kong)	-11	-3	-13	-3
SMU	(Bratislava)	-3655	-3665	-3679	-3690
SO	(Shanghai)	-	-	-	-
SP	(Boras)	-442	-423	-412	-395
SU	(Moskva)	84	88	89	90
TL	(Chung-Li)	-4	-47	-57	-68
TP	(Praha)	98	107	114	118
UME	(Gebze-Kocaeli)	-600	-712	-203	-243
USNO	(Washington DC) (MC)	-3	-3	-2	-5
VSL	(Delft)	8	14	11	11

Fig. 7.5. First page of BIPM Circular T.

7.3.3 Frequency stability

Ordinate values in Figure 7.4 represent, after sign change, differences in normalised frequencies between TAI and the set of primary standards (on the geoid). An annual fluctuation is observed. Its total amplitude was close to 10^{-13} over the period 1977–1982 but then gradually fell to around 10^{-14} in 1993–1995, before virtually disappearing altogether. Is this due to TAI or to the primary standards? When the amplitude was large, it seems that fluctuations were mainly due to environmental influences on commercial caesium clocks, with humidity being the dominating factor. This weak point was gradually rectified, either by improving clocks, or by giving them better protection. The tiny magnitude of the fluctuation remaining in 1995 may still be caused by environmental influences acting while time comparisons are made. The temperature sensitivity of certain GPS receivers has already been mentioned. There may also be inadequacies in the way refraction has been modelled, in particular, tropospheric refraction, which is never measured. Primary standards appear to be exempt from detectable annual variations.

The stability of TAI can be estimated by comparison with other atomic time scales maintained by several timing centres. Instabilities shown in Figure 7.6 were calculated for sampling periods τ ranging between 10 and 160 days. In fact, these are instabilities in the EAL in 1999 and 2000 [7.6], but for these values of τ, they differ only slightly from the instabilities in TAI. For longer τ, instability is estimated under the assumption that current properties of clocks and primary standards remain the same indefinitely. It levels out at around 5×10^{-15} as a result of frequency steering. Frequency accuracy imposes an upper limit on very long term instability.

7.3.4 Frequency accuracy

Let d_{TAI} be the relative deviation of the TAI scale unit from the TT scale unit (the SI second on the rotating geoid), i.e., the relative frequency deviation of TAI with the opposite sign ($d_{\mathrm{TAI}} = -y_{\mathrm{TAI}}$). The monthly average of d_{TAI} is normally less than $\pm 10^{-14}$. Its value is supplied retrospectively by the BIPM, together with its standard uncertainty u. The latter is estimated with great care, taking into account uncertainties in the primary frequency standards and in their links with TAI. At the end of 2000, the value of u lay in the range 2–3×10^{-15}. For example, for the interval 28 August 2000 to 27 September 2000, $d_{\mathrm{TAI}} = 7.3 \times 10^{-15}$ and $u = 2.2 \times 10^{-15}$.

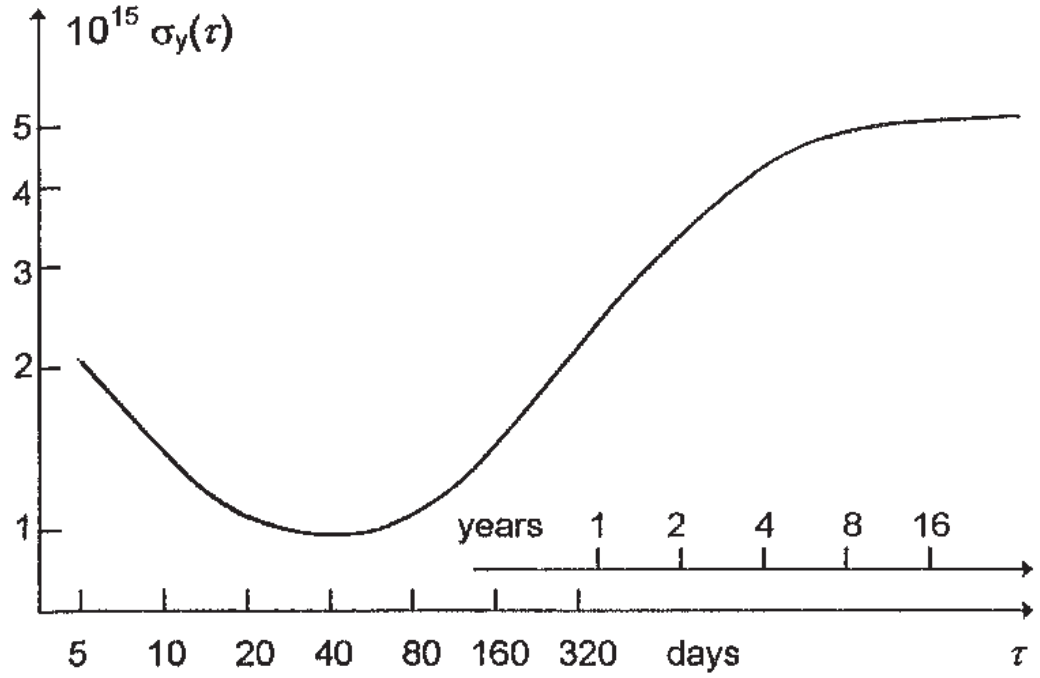


Fig. 7.6. Frequency instability of TAI (see text for explanation).

7.4 World organisation of time measurement

A worldwide organisation of time measurement gradually came into being as progress was made with atomic clocks and standards, together with improved means for comparing instruments and time scales. Once this organisation had reached a certain maturity, it was recognised and stabilised by the relevant international organisations. Chapter 4 gives an overview of the evolutionary phase. In this chapter, we shall describe the present state of the organisation, which would appear to represent the future at least for the next few years, or even a decade or more.

7.4.1 Coordinated Universal Time

As explained in Section 4.5, Coordinated Universal Time (UTC) in its present form is none other than International Atomic Time corrected by a whole number of seconds in order to follow approximately the Universal Time UT1 tied to the Earth's rotation.

Since 1 January 1972, UTC has been run according to the guidelines in Recommendation ITU-R TF 460-4 of the International Telecommunications Union (ITU). This recommendation took root in Recommendation 460 adopted in 1970 by the International Radio Consultative Committee (CCIR), an ITU

committee that has since been dissolved. Let us now outline several of the directives and explain how they are put into practice.

(a) Leap seconds

Jumps of one whole second, referred to by the ITU as 'the introduction of one positive or negative leap second', must be introduced at the end of a UTC month, preferably at the end of December or June, otherwise at the end of March or September. It is stipulated that a positive leap second should begin at 23 h 59 m 60 s and end at 0 h 0 m 0 s the following day. There is then no ambiguity when events are dated in this way. However, an ambiguity appears in other systems, for example, when fractional days are used. If a positive leap second has been introduced, the date $N.000\,005\,79$ could mean either day $(N-1)$ at 23 h 59 m 60.5 s, or day N at 0 h 0 m 0.5 s. This ambiguity does not arise in the case of a negative leap second. However, the probability that negative leap seconds will be needed is almost zero.

Dates for leap seconds are fixed by the International Earth Rotation Service (IERS) and announced at least 8 weeks beforehand.

The UTC system works well. On 1 January 1972, $TAI - UTC$ was equal to 10 s. Positive leap seconds have always been introduced at the end of December or the end of June, at a rate of between one per year and one every 2.5 years. From 1 January 1999 up until a date that cannot yet be specified, $TAI - UTC = 32$ s. The relation between TAI and UTC can be found in the Annual Reports of the IERS and the Time Section of the BIPM.

(b) Time signals. Code for Universal Time UT1

Time signal emissions conform as closely as possible to UTC. An ITU recommendation fixes a tolerance of 1 ms. In reality, the discrepancy is much smaller than this. It is also recommended that the carrier frequency be tuned to the TAI frequency, with relative frequency offset less than 10^{-10}, and that the relationship between time signals and the phase of the carrier wave should be known (e.g., if the carrier has frequency 10 MHz, the second pulse should occur every 10^7 periods).

It is recommended that time signals should carry audible coded information about the difference $UT1 - UTC$. This information is a quantity $DUT1$, in the form of an integral multiple of 0.1 s, whose value is provided by marking a certain number of second pulses, either by broadening, splitting or any other method. Figure 7.7 shows how the code works. Values of $DUT1$ and their application date are provided one month beforehand by the IERS and they are the same for all emissions.

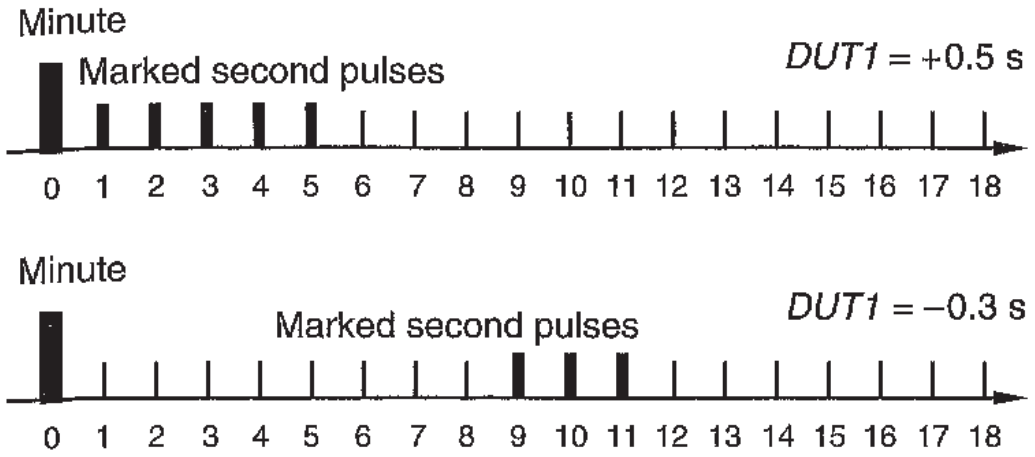


Fig. 7.7. Code for $DUT1 = UT1 - UTC$ transmitted with time signals. Second pulses are emphasised by broadening, splitting, or some other audible means. For negative values of $DUT1$, the relevant number of pulses is marked from second 9.

Some time signals also carry coded information $dUT1$ giving the value of $UT1 - UTC$ to the nearest 0.02 s, but this is unofficial. Even the coded information $DUT1$ is rarely used and some emissions no longer include it.

The Annual Report of the BIPM Time Section provides a list of international time signals with their main characteristics.

(c) Using UTC

Coordinated Universal Time is used as a practical reference worldwide. Its use, like the acronym UTC which is the same in every language, is recommended by international organisations, including the 1971 General Conference on Weights and Measures (CGPM) and the ITU.

In different countries around the world, local time is attached to UTC corrected by a whole number of hours. This is sometimes stipulated by written law, as in France (by decree, on 9 August 1978). Legal institutions sometimes prefer to use the national approximation to UTC, as happens in Germany. There are countries in which UTC is not legally recognised, although it is actually used, since no other time scale is readily available.

Let us just recall that the still widespread use of the acronym GMT is not correct when it is intended to refer to UTC, as is the case when expressing the time in general usage. This practice has often been criticised, especially by the IAU. Strictly speaking, it should be replaced by UTC. However, for everyday usage, where inaccuracies of more than one second are of no real import, UT is an acceptable appellation.

7.4.2 Local representations of UTC and independent local atomic times

As mentioned in Section 7.2.2, the approximation UTC(k) to UTC is maintained at a timing centre k. This is usually realised by an atomic clock with an attached device which adds a linear time correction, whilst allowing adjustment of its two parameters. In other cases, this physical device is replaced by calculation, allowing in principle a more complex correction, with a quadratic term, for example. In both cases, every effort is generally made to keep UTC(k) as close to UTC as possible. In 1993, the Consultative Committee for the Definition of the Second recommended that timekeeping laboratories supply the information required to facilitate synchronisation with UTC in real time, to within 100 ns, if feasible. One way of achieving this is to ensure $|UTC - UTC(k)| < 100$ ns.

The BIPM's monthly cycle for producing and disseminating $UTC - UTC(k)$ requires one to two month extrapolations in order to servo control the UTC(k) on UTC. To this end, laboratory k can base its calculations on extrapolations from those of its clocks with the best stability over a sampling period of around one month. However, if this laboratory has several atomic clocks, it is better to construct, on this local level, a time scale known as *independent local atomic time*. Such a time scale, denoted TA(k), can be set up much more quickly than TAI/UTC, because local time comparisons are extremely accurate. This then provides a more solid foundation for extrapolating UTC(k).

In 2000, about fifteen laboratories were producing a TA(k) by various means. Let us consider some examples. At the US Naval Observatory, the scale TA(USNO) is based on several tens of commercially manufactured caesium clocks and also on hydrogen masers. In Germany, TA(PTB) is directly produced by one of the PTB's primary standards. In France, TA(F) is calculated by the Laboratoire primaire du temps et des fréquences at the Paris Observatory, using data from about twenty commercial caesium clocks spread across the country. The BIPM publishes the values of $TAI - TA(k)$ in the same way as the values of $UTC - UTC(k)$ in its Circular T.

The TA(k) are not independent of TAI (or UTC) since the same clocks participate in both undertakings. But they are independent of one another and this makes it possible to deduce their instability from the mutual instability of each pair of time scales (using the so-called *three-corner hat method* mentioned in Section 5.2.8, or indeed the *n-corner hat method*).

In 1995, only two laboratories in the United States were maintaining their UTC(k) to within 100 ns of UTC, namely, the National Institute of Standards and Technology (NIST) and the USNO already mentioned. In 1996,

Table 7.1. *Synchronisation between UTC and UTC(k) for four timing laboratories and GPS over the period January to September 2000.*

Laboratory or system	Extremal value of $UTC - UTC(k)$ (ns)	Mean value of $UTC - UTC(k)$ (ns)	Rms value of $UTC - UTC(k)$ (ns)
GPS	−39	−7	13
NIST	28	15	17
NPL	80	20	29
OP	−49	−31	31
USNO	−20	2	9

the National Physical Laboratory (NPL) in the United Kingdom and the Paris Observatory (OP) could be added to the list. On 27 September 2000, 20 out of the 46 $UTC - UTC(k)$ published by the BIPM were, in absolute values, less than 100 ns. Table 7.1 illustrates the quality of synchronisation for the four laboratories quoted and for GPS, from January to September 2000.

7.4.3 Other forms of atomic time

By convention, TAI is the international reference. However, there is nothing to prevent other time scales being set up using data available the world over, if some advantage is to be gained for specific projects. It might be thought that a more stable and accurate atomic time than TAI would be realised if a longer publication time were possible and if corrections were authorised on the basis of the latest information received. For example, the effect of blackbody radiation on the frequency of primary standards was only taken into account in a uniform manner for TAI in September 1995, and no retrospective correction could be made to TAI to improve past values.

We shall see in Chapter 8 that certain shortcomings of TAI are critical in studies that involve timing fast pulsars. Aware of the need for an improved time scale for such studies, the BIPM set up a dated time scale called TT(BIPMxx), where xx represents the two last figures of the year in which it was produced. Although this scale is based on the EAL, data from primary standards is not used in the same way as for TAI. Annual fluctuations, mentioned previously, are eliminated [7.7]. Successive versions of TT(BIPMxx) are not straightforward extensions, since they may differ over the whole length of the time scale from their starting date on 1 January 1977, which is the same for all scales. TT(BIPMxx) is supplied in the form of differences with TAI at standard dates

(every ten days, then every five days), upon request to the BIPM. The offset between TT(BIPM96) and $TAI + 32.184$ s, zero on 1 January 1977, can reach 25 μs.

7.5 Dissemination of time and frequency

7.5.1 Frequency

(a) Standard frequency emissions

Carrier frequencies of specialised time signal and standard frequency emissions are based on TAI frequency to much greater accuracy than is suggested in ITU recommendations. Most announce uncertainties of 10^{-11}, whilst in some cases it is pushed down as far as 10^{-12}. The BIPM provides a list of these emissions in the Time Section Annual Report.

Certain other radio signals, used for various purposes, have very accurate frequencies. Loran-C emits at 100 kHz to within $\pm 10^{-12}$. Some radio broadcast signals have very accurate frequencies, such as Allouis in France, at 162 kHz, to within $\pm 2 \times 10^{-12}$.

These signals cover most frequency reference requirements. They are easily accessed using quite ordinary receivers.

(b) Access to primary frequency standards

For fundamental metrology, the subject of this book, it is often useful to have access to the frequency of a remote primary frequency standard E. When the greatest possible accuracy is desired, time comparisons are used, as explained in Section 5.6.4a. If A is a local oscillator, $y_E - y_A$ is obtained via (5.67) and (5.68).

If time is not a factor, $y_E - y_A$ can be obtained without asking anything of the primary laboratory housing E, simply by consulting BIPM publications. These give access to TAI locally (e.g., by reception of GPS signals). They also provide, under the notation d, the relative difference between the duration of the TAI scale unit and the proper second as realised by the standard E, referred to the geoid. We use here the notation $d_{\mathrm{TAI,E}}$ for this quantity, which is an average over an interval (TAI_1, TAI_2). Hence,

$$[\langle y_E \rangle - \langle y_{TAI} \rangle]|_{geoid} = d_{TAI,E} \,, \tag{7.20}$$

where $\langle\ \rangle$ represents the average over the relevant interval. For the user's clock A, with proper time τ_A, we measure

$$N_A = \frac{[\tau_A - TAI]\,(TAI_2) - [\tau_A - TAI]\,(TAI_1)}{TAI_2 - TAI_1} \,, \tag{7.21}$$

and then, by (5.73),

$$\langle y_A \rangle - \langle y_{TAI} \rangle = N_A + \langle \underline{h}(TAI) \rangle \ . \tag{7.22}$$

The function $\underline{h}$ is given by (3.30) or (3.32) and its average value is taken along the worldline of A between TAI_1 and TAI_2. Hence,

$$\langle y_A - y_E \rangle = N_A - d_{TAI,E} + \langle \underline{h}(TAI) \rangle \ . \tag{7.23}$$

For a fixed clock A on the Earth, we find

$$\langle y_A - y_E \rangle = N_A - d_{TAI,E} - 1.091 \times 10^{-13} h_A/\text{km} \ , \tag{7.24}$$

where h_A is the altitude of the user above the geoid, in kilometres.

Either by making direct time comparisons or by using BIPM publications, uncertainties involved in accessing primary frequency standards can be significantly less than 10^{-14}. However, such uncertainties are too great for the latest standards, where frequency uncertainties are as small as 10^{-15}. The problem of how to gain remote access to the frequencies of these standards has not yet been solved.

7.5.2 Time

GPS Time is related to TAI by

$$[TAI - GPS\ Time] = 19\ \text{s} + C_0 \ , \tag{7.25}$$

where the synchronisation discrepancy C_0 is usually maintained within ± 100 ns limits. From March 1995 to the present (October 2000), it has remained within these bounds, and has often been below 50 ns. The values of C_0 are published for each day at 0 h UTC, in BIPM Circular T. Uncertainty is 10 ns. Hence, access to TAI and UTC is possible everywhere to within ± 10 ns, although deferred by one or two months, using commercially available GPS time receivers. In order to attain such accuracies, a daily average must be taken over a great many satellite trackings. In addition, coordinates of the antenna must be known to within one or two metres in the WGS84 or ITRF systems. GPS time receivers can give these coordinates, but to insufficient accuracy.

The other highly accurate way of accessing TAI or UTC, once again deferred by one or two months, is to establish a link with some laboratory k whose clocks take part in TAI and for which $UTC - UTC(k)$ is published. Uncertainties may be slightly less than 10 ns. This method raises the problem of short range time transfer links. GPS can be used (some errors decreasing with distance), otherwise cable or direct view television represent further possibilities.

Today (2000), access through GLONASS has the same qualities as access via GPS. Putting

$$[UTC - GLONASS\ Time] = C_1\ , \tag{7.26}$$

values of C_1 are a few hundred nanoseconds. They are published by the BIPM for each day with uncertainties around 10 ns.

In real time, we may use either the values of UTC transmitted by GPS ($\pm$100 ns), or a link with a laboratory where $UTC - UTC(k)$ can be extrapolated, again with the possibility of uncertainties attaining values around $\pm$100 ns.

To lesser accuracy, there are a great many ways of accessing UTC or official time, such as speaking clocks or radio broadcast pips. Some coded emissions can be used to display the time. It is not certain that radio time signals, aimed mainly at navigators, will be able to hold out against the facilities offered by other rapidly developing means of disseminating time and positioning, viz., GPS, GLONASS or their civilian counterparts.

7.6 Summary and prospects

The current world organisation of atomic timekeeping was set up at the beginning of the 1970s. At the time, choices were guided by three main considerations.

- There were many caesium atomic clocks in existence, commercially built and possessing excellent long term stability, but falling short in terms of accuracy. Systematic frequency differences persisted throughout the lifetime of these clocks.
- Primary frequency standards only operated on an occasional basis, and doubts remained concerning their stated accuracy.
- To reduce uncertainties in frequency comparisons due to time comparisons, averaging was required over one or two month periods.

At the beginning of the 1980s, questions were raised about the way atomic time measurement was being organised. The primary standard NRC Cs V at the Canadian National Research Council had been running continuously (as a *primary clock*) since 1976. It was joined in July 1978 by the German primary clock PTB CS1, and then in 1980 by three other primary clocks at the NRC, viz., NRC Cs VI A, B and C. There was good frequency agreement between these standards, and the unrivalled accuracy of PTB CS1 ($\pm 3 \times 10^{-14}$) was confirmed in 1986 by PTB CS2, which proved to be even more accurate ($\pm 1.5 \times 10^{-14}$). These developments made it reasonable to ask whether TAI

should not be based directly and solely on such primary clocks. The calculation of TAI from the clock readings of many laboratories was nonetheless maintained, since it guaranteed more reliable results. However, there was another reason, namely, the desire that TAI should remain a collective enterprise, involving the commitment of a significant number of laboratories, both large and small. The number of countries joining the TAI club had grown to thirty by 2000. This spirit of cooperation in the area of time measurement has never failed. It is through this cooperation that such close synchronisations have been possible with UTC, including the synchronisation of GPS which is not required for the system to work, but which simply serves the world community. It is this same spirit that has made it possible to set up time comparison networks that form the very foundation of the whole edifice.

Current research is likely to produce even more accurate and stable primary standards, with uncertainties being decreased by at least a factor of ten. But since 1994, most national timing laboratories have equipped themselves with new commercial caesium clocks in which long term instabilities have also been reduced by a factor of ten, compared with the previous generation of clocks. These instruments are thus progressing at the same rate. However, the accuracy of time comparisons via GPS had made it possible to reduce the frequency sampling period upon which calculation of TAI is based from two months to one month.

Apart from this modification and despite the pressures of progress, the original 1970 organisation continues to operate. It is often reassessed by the Consultative Committee for Time and Frequency but, up to now, the conclusion has always been to leave things as they are.

To conclude, let us recall that, today, the time in TAI or UTC can be made available everywhere, to within a hundred nanoseconds in real time, and to within about ten nanoseconds one or two months after the event. Taken as an entity, the second, that is, the duration of one second alone, can be realised at best to within a relative error of about $\pm 5 \times 10^{-14}$ by a cooled caesium atom standard, owing to the short term frequency instability of this device. But this standard can express in seconds a duration of several hours or more to within 10^{-15} (relative value). Through access to TAI, the second can be obtained everywhere to within a few multiples of $\pm 10^{-15}$, on average over a few days.

8

Astronomical times

Although astronomical times are no longer the best measure of time, they continue to play a role in current research. Universal Time is witness to the Earth's rotation, whilst serving also to establish Coordinated Universal Time (UTC), the practical basis for unifying times the world over. Ephemeris Time (ET) is the best representation of dynamical time before the advent of atomic time in 1955 and is still used to process astronomical observations made prior to that date. And finally, some have hoped that *Pulsar Time* might exhibit better long term stability than atomic time. In the present chapter, we shall explain how these times are defined, realised and used.

8.1 Universal Time

8.1.1 Conceptual definition

A historical view of the developments leading up to Universal Time might resemble a series of metamorphoses of apparent solar time (see Section 4.2). It is simpler to give a conceptual definition, based upon the Earth's rotation, which brings out its current role.

There are various forms of Universal Time. The one directly related to the Earth's rotation is UT1. Other forms will be defined in Section 8.1.3.

In Section 3.3.2b, we defined geocentric coordinate systems. One is rotating in space (*celestial system*) and the other is rotating with the Earth (*terrestrial system*). The rotation of the terrestrial system with respect to the celestial system is the *rotation of the Earth*. We showed how these systems are realised as frames of reference ICRF, denoted (C) here, and ITRF, denoted (T) here. It is the rotation of (T) relative to (C) that is described. To do so, we appeal to classical kinematics. The time t is treated as absolute and realised by International Atomic Time with negligible uncertainty.

Although three functions of time (the three Euler angles) are sufficient to orient (T) with respect to (C), a five parameter representation is preferred, which includes the direction (I) of the rotation axis. The latter moves relative to both (T) and (C).† These parameters are as follows.

- The two angular coordinates giving the position of (I) in (C) describe *lunisolar precession–nutation*. Observed values are usually given in the form of deviations from a theoretical model of precession–nutation, chosen by convention. At present, deviations from the model are of the order of $0.01''$. They will be reduced to about $0.001''$ when a new model adopted in 2000 is put into practice.
- The two angular coordinates of (I) in (T) describe the *terrestrial polar motion*. They are given in the form of deviations between the direction of (I), this axis being oriented towards the north, and the X^3 axis of (T), denoted by Z in IERS publications: $x(t)$ is measured along the prime meridian and $y(t)$ along the meridian 90° west. Values of $x(t)$ and $y(t)$ do not exceed $1''$, corresponding to 30 m on the Earth's surface.
- An angle $A(t)$, measured around (I), completes the description. The position of (I) is the one it occupies at date t.

The angle A can be defined in two ways, as we shall see. UT1 is expressed as a function of A by mathematical relations in which numerical parameters are chosen so that, whichever choice is made, the same values of UT1 are obtained. These relations are set up in such a way as to satisfy the following conditions.

(i) The derivative of UT1 with respect to time must be proportional to the modulus of the (variable) rotation vector. Then UT1 fulfills its historical role as a uniform time under the assumption of uniform rotation, and its modern role as a simple representation of the Earth's rotation.

(ii) The constant of proportionality implied by (i) and the phase are chosen so that, over the millennia, the Sun transits the prime meridian (the Greenwich meridian) at 12 h UT1, on average.

Strictly speaking, the second condition is not compatible with (i) because, even if the Earth's rotation is uniform, the mean duration of an apparent solar day varies very slowly. However, priority has been given to (i), and condition (ii)

† The instantaneous axis of rotation (IAR) has diurnal and sub-diurnal motions, both in (T) and (C). Although their amplitude is less than $0.02''$, they are inconvenient in practice. The IAR is replaced by an axis close to it, which is not subject to these motions. This axis was first defined by the IAU in 1976 and the definition was subsequently improved in 2000. The corresponding poles were referred to as the Celestial Ephemeris Pole (CEP) and Celestial Intermediate Pole (CIP), respectively. In the following, the CIP and its equator are used implicitly.

is only approximately fulfilled. The shift between UT1 and the alternation of day and night resulting from this approximation is quite negligible, being well under one second over a few centuries.

If the Earth's rotation vector is $\boldsymbol{\omega}(t)$ with length $\omega(t)$, we must therefore have

$$\frac{\mathrm{d}\,(UT1)}{\mathrm{d}t} = \frac{\omega(t)}{\omega_0}\,, \tag{8.1}$$

where ω_0 is a constant angular speed corresponding to $\mathrm{d}\,(UT1) = \mathrm{d}t$, whose value is chosen to satisfy (ii). In integral form, this means that

$$[UT1](t) = [UT1](t_0) + \frac{1}{\omega_0}\int_{t_0}^{t} \omega(t)\,\mathrm{d}t\,, \tag{8.2}$$

the constant of integration also being chosen to satisfy (ii).

Equation (8.2) expresses the conceptual definition of UT1, whilst the relationships between the two forms of A and UT1 given in Section 8.1.2 are practical definitions.

Let us review some other useful relations. As the values of UT1 are published in the form of corrections to TAI, as a function of the TAI (or UTC) date, it is sometimes convenient to write (8.2) in the form:

$$[UT1 - TAI](t) = [UT1 - TAI](t_0) + \int_{t_0}^{t} \frac{\Delta\omega(t)}{\omega_0}\,\mathrm{d}t\,, \tag{8.3}$$

where $\Delta\omega = \omega - \omega_0$.

The quantity $\omega(t)$ can be deduced from published values of $UT1 - TAI$ by

$$\omega(t) = \left[1 + \frac{\mathrm{d}(UT1 - TAI)}{\mathrm{d}(TAI)}\right]\omega_0\,. \tag{8.4}$$

Instead of ω_0, it is common to use a more suggestive quantity $D(t)$, called the *length of day*, which is the duration in TAI seconds corresponding to a 24 hour increase in UT1. Then,

$$D(t) = \frac{\omega_0}{\omega(t)} \times 86\,400\ \mathrm{s}\,. \tag{8.5}$$

8.1.2 Practical definition of UT1

We shall now tackle the subtle problem of defining the directions between which the angle $A(t)$ is measured, distinguishing two options.

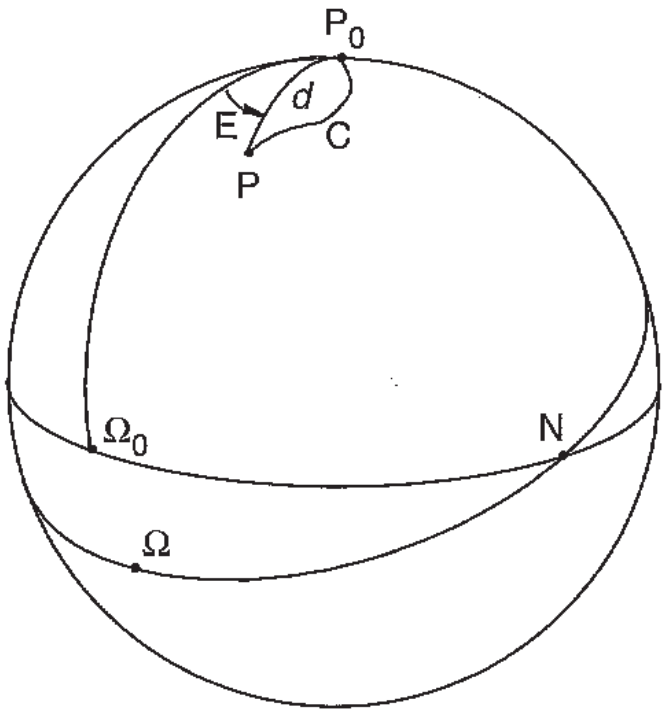


Fig. 8.1. Defining the instantaneous longitude origin Ω in the moving equator.

(a) Use of non-rotating origins

A form $A_S(t)$ of $A(t)$, sometimes called the *stellar angle*, has recently been found whereby this angle is linearly related to UT1 [8.1, 8.2].

Following the method generally used by astronomers, we represent directions by straight lines originating in some arbitrarily chosen point O. These lines intersect a sphere of unit radius (the *celestial sphere*) in points that can be taken to represent these directions. The equator is a great circle on this sphere and $A_S(t)$ is represented by an arc on the equator at date t.

Let us begin with the determination of the origin of $A_S(t)$ in (T). To simplify, we assume that the X^3 axis corresponds to the position P_0 of the pole at date t_0. Let Ω_0 be the origin for longitudes in (T), i.e., the point representing X^1. At date t, the pole has moved to P along the trajectory C (see Figure 8.1). The equator of P cuts the one associated with P_0 at N (the *ascending node*). In order to position the *instantaneous longitude origin* Ω, the following condition is imposed: the rectangular triad attached to OP and OΩ has no instantaneous component of rotation about OP when P moves. If P is located via its polar coordinates as shown in Figure 8.1, it is easy to show that the quantity $s = \Omega N - \Omega_0 N$ is given by

$$s = \int_C (\cos d - 1)\, \mathrm{d}E \ . \tag{8.6}$$

The position of Ω thus depends on the whole history of the Earth's rotation, a worrying but unavoidable conclusion. In this case, s is small, because d never exceeds $1''$. This is fortunate, because s has always been taken as zero in practice. The situation is quite different in (C).

When treated in (C), the same problem leads us to define an origin in the moving equator, called the *non-rotating origin* σ. However, the motion of P in (C) includes a secular term (precession) of about $20''$ per year and periodic terms (nutation) with total amplitude around $20''$, in such a way that s can take on significant values. For example, for precession we find

$$s_P = 36.28''(t - t_0)^3 - 0.04''(t - t_0)^4 + \cdots , \tag{8.7}$$

where $t - t_0$ is measured in thousands of years. The full expression for s including precession and nutation, valid to within a few times $10^{-6''}$ until 2100, is given in [8.3].

Having established the origins Ω and σ, we can define $A_S(t)$ modulo 2π as the arc $\Omega\sigma$. Since every component of rotation about the polar axis is zero by definition of the non-rotating origins, $A_S(t)$ represents the integral in (8.2) directly. The values of UT1 are proportional to A_S. It is useful to deduce the relation between UT1 and A_S from the expression for A_S at 0 h UT1, as provided each day. Let d'_u be the number of UT1 days elapsed since 2000 January 1 at 12 h UT1 (d'_u is a positive or negative whole number plus or minus 0.5). Then setting $T'_u = d'_u/36\,525$ (expressed in Julian centuries of 36 525 days) and taking one second of hour angle, viz., $15''$, as unit of angle, a common practice amongst astronomers, we have

$$A_S \text{ at } 0\,\text{h UT1} = 24\,110.548\,41 + 8\,639\,877.317\,38\,T'_u . \tag{8.8}$$

This equation is established using the value of ω_0 given by

$$\omega_0 = 7.292\,115\,146\,706 \times 10^{-5} \text{ rad s}^{-1} , \tag{8.9}$$

chosen so as to satisfy the conceptual definition of UT1.

To obtain A_S at a general instant of time, d_u must be calculated in the same way as d'_u, but in UT1 days and fractions of a day for this instant, together with the corresponding T_u, adding $86\,400 \times 36\,525$ seconds to the coefficients of T'_u. For example, with A_S in revolutions,

$$A_S = 0.779\,057\,273\,264 + 1.002\,737\,811\,911\,354\,d_u . \tag{8.10}$$

At its General Assembly in 2000, the IAU recommended the use of the non-rotating origins associated with the Celestial Intermediate Pole, under the names of Celestial Ephemeris Origin in (C) and Terrestrial Ephemeris Origin in (T). This recommendation should take effect in 2003.

(b) Use of the equinox

Although the stellar angle A_S is becoming more and more frequently used in practice, let us recall the developments based on the traditional origin in (C). This is the *vernal equinox* Υ, the intersection between the equator and the ecliptic. The arc $\Omega\Upsilon$ is *Greenwich Sidereal Time* GST. It is related to UT1 in a much more complicated way than the stellar angle. It involves periodic terms that have to be subtracted to define the *Greenwich mean sidereal time* GMST. Then, according to an IAU decision [8.4, 8.5], with the previous definition of T'_u and in seconds of hour angle, we find

$$\begin{aligned} GMST \text{ at } 0\text{ h UTI} = 24\,110.548\,41 + 8\,640\,184.812\,866\,T'_u \\ + 0.093\,104\,T'^2_u - 6.2\times 10^{-6}T'^3_u\ . \end{aligned} \tag{8.11}$$

The relation at arbitrary times is deduced in the same way as for A_S.

8.1.3 Other forms of Universal Time

It is sometimes necessary to estimate the Universal Time when the instantaneous position of the Earth's pole is not known. We then take $x = 0 = y$ at the time of estimate, and this leads to the form UT0. The difference $UT1 - UT0$ is given by

$$UT1 - UT0 = -(x \sin L + y \cos L)\ \tan\phi\ , \tag{8.12}$$

where ϕ and L are the latitude and longitude (positive towards the east) of the observer, and x and y are the coordinates of the pole. At medium latitudes, this difference is less than a few hundredths of a second.

There is also a regularised form of Universal Time, denoted UT2, which corrects for an annual irregularity. The difference $UT2 - UT1$ varies between ± 30 ms. UT2 does not represent the Earth's rotation any more than UT0 does. Only UT1 has a physical meaning.

Also defined is UT1R, which is UT1 corrected for zonal tidal effects affecting the polar moment of inertia and having periods less than 35 days. Corrections are less than 2 ms. Finally, UT1R′ is corrected for the effects of all tides, the main term having amplitude 0.16 s and period 18.6 years.

8.1.4 Measurement techniques

Until about 1970, our knowledge of the Earth's rotation was based entirely upon optical observation of the apparent motion of the stars relative to terrestrial reference directions. After a transition period of about ten years, this

approach was superseded by methods using space geodesy techniques and interferometry.

Several of these new methods coexist. They provide complementary data for establishing the celestial and terrestrial reference frames and monitoring the five parameters that specify the Earth's rotation. But they are also competitive in certain domains where they lead to results of comparable quality. In the following brief review, we shall see that all these methods are based on time and frequency techniques.

(a) Very long baseline radio interferometry (VLBI)

VLBI was first developed to study the structure of celestial radio sources. Around the beginning of the 1980s, it was used to measure the direction of these sources relative to a terrestrial baseline.

The measurement method involves receiving quasar radiation at centimetre and decimetre wavelengths using two antennas several hundred or several thousand kilometres apart. These antennas define the baseline. The characteristics of the radiation are recorded on magnetic tape as a function of the time given by clocks placed near the antennas. Let us assume that the clocks are synchronised. The magnetic tapes are compared and correlations sought between the signals received by the two antennas. The time lag between them can thus be found and, if the distance between the antennas is known, the angle between the baseline and the source direction can also be determined (Figure 8.2).

When this method is implemented, clocks cannot be synchronised sufficiently accurately by external methods like the ones described in Chapter 5. They therefore remain independent and time comparisons have to be modelled, for example, by a second order polynomial in the time, throughout the period of the measurements to be treated in a single batch (often 24 hours). Model parameters, such as the coefficients of this polynomial, are left as unknowns to be deduced from the VLBI observations themselves. It is therefore essential to use clocks that are as stable in frequency as possible over sampling periods up to the duration of the relevant batch of observations, so that the model of their time difference involves a minimum of parameters. This explains why hydrogen masers are used, and indeed, this is one of their principal applications.

Likewise, the antenna coordinates are not known a priori, at least, not with sufficient accuracy, and neither are the source positions. When observations of several sources are analysed globally using several baselines, the various unknown parameters can be fitted. VLBI can supply all five Earth rotation parameters and construct both celestial and terrestrial reference frames. However, the terrestrial frame is only obtained up to a translation. The centre of mass of the Earth is not located.

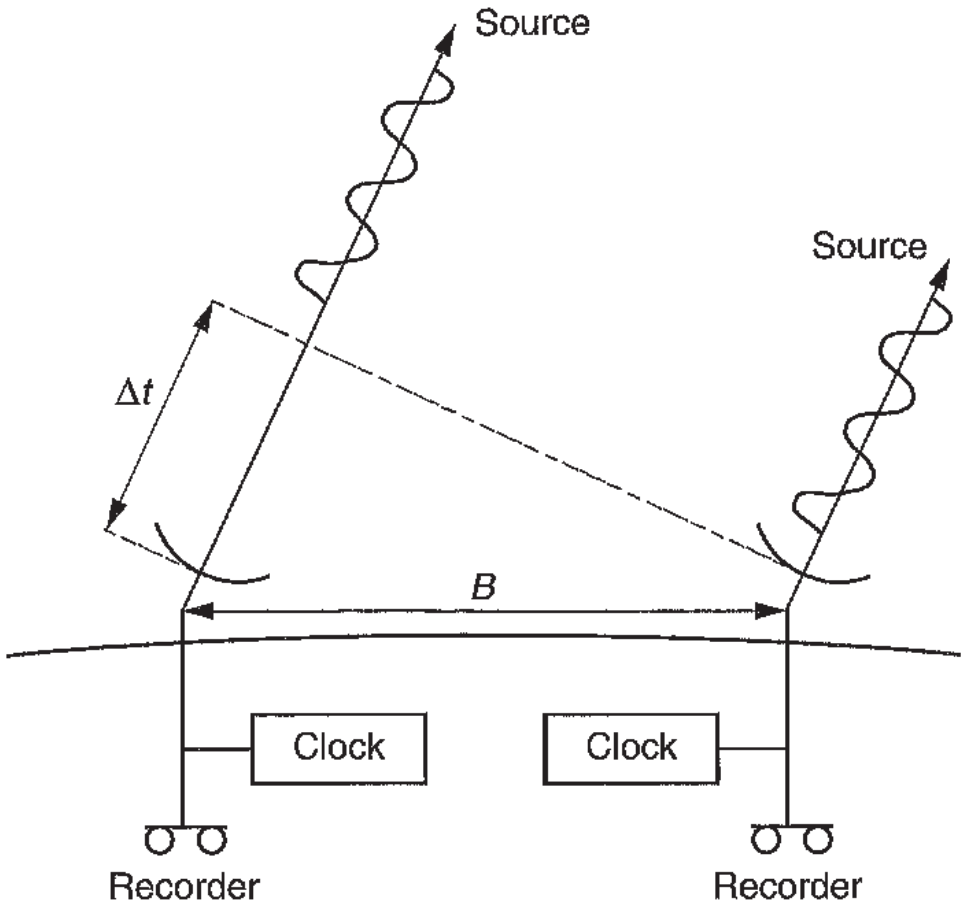


Fig. 8.2. Astrometric use of VLBI. By measuring the geometrically produced time lag Δt, we can find the orientation of the base vector B relative to the source direction.

The application of VLBI to astrometry and geodesy is coordinated by the International VLBI Service (IVS). We have already noted in Section 5.6.3c that clock comparisons could be a spin-off of VLBI if instrumental delays were measurable.

(b) Dynamical methods

In VLBI, nature directly yields up the non-rotating celestial frame. The way is then open for geometry. For the methods considered here, observations refer to objects in orbit around the Earth, viz., artificial satellites and the Moon. The non-rotating reference frame is realised by dynamical studies of the motion of such objects. Let us adopt a highly schematic approach in order to bring out the essential features of these so-called dynamical methods. We imagine that the Earth is spherical, homogeneous and isolated in space, and that a satellite is deployed in circular orbit around it. The plane of the orbit has fixed orientation in space. In addition, the orbital period of the satellite is constant and can be measured with great accuracy by pursuing observations as long as necessary from various points on the Earth's surface. We can then identify fixed geocentric axes in the orbital plane and a non-rotating frame is thus established.

In practice, dynamical methods are extremely complex because significant perturbations have to be accounted for. These include the presence of neighbouring celestial bodies, irregularities in the shape and rotation of the geoid, residual friction effects, solar radiation pressure, and refraction in the Earth's atmosphere. All the corresponding problems have been resolved and dynamical methods now attain comparable accuracies to VLBI, wherever both approaches are applicable.

Dynamical methods can determine the Earth's polar motion, provide data concerning precession–nutation, and realise the terrestrial reference frame in relation to the centre of mass of the Earth (something VLBI cannot do). However, residual terms not included in corrections for the above-mentioned perturbations lead to a drift in UT1 values. This drift becomes more pronounced when perturbations to the satellite increase. It is very slow for the Moon, but not negligible. For this reason, these methods are well suited to studying short term variations in UT1. They can be categorised according to the object observed and the way it is observed.

The coordinates of the pole were first obtained in 1967 by tracking satellites in the Navsat system using the Doppler effect. An improved version of this technique is the DORIS system developed by the Centre national d'études spatiales (CNES) in France and used to track Spot satellites, as well as the Franco-American TOPEX–POSEIDON oceanographic satellite.

For extremely precise applications, several spherical satellites were launched by France (Starlette, 1975), the United States (Lageos, 1976), then Japan and Russia. These are very dense in order to reduce the relative importance of non-gravitational perturbations and covered with retroreflectors. They are tracked by laser ranging and uncertainties in measured distances are of the order of one centimetre. Apart from their direct contribution to measuring the Earth's rotation and constructing the terrestrial frame, these satellites are of great importance in representing the geoid. This in turn is a major contribution to geophysics and all those techniques based upon precise orbitography of artificial satellites. It is unfortunate that laser telemetry is so expensive and sensitive to cloud conditions, because it is a very accurate method which can also be used for calibrations. An International Laser Ranging Service (ILRS) coordinates this activity.

The dynamical method which seems to be the most efficient involves tracking the twenty-four GPS satellites. Coordination and analysis of observations is the task of an International GPS Service (IGS) which produces outstanding results for the motion of the terrestrial pole and geodetic coordinates.

Three reflecting panels were deployed on the Moon from 1969 by Apollo astronauts, and two more (French made) by Soviet unmanned Lunakhod

spacecraft. One was never relocated but the other four are used for laser ranging. This telemetry also achieves centimetre accuracy but remains a state-of-the-art technique only mastered by two ground stations, in France and the United States, where observational data is regularly retrieved. For this reason, the method has not produced a global assessment of the Earth's rotation, only some useful values of UT1. It contributes rather to studies of the Earth–Moon system.

(c) Coordination and results

Results obtained by each of these technical centres (IVS, DORIS, ILRS, and IGS) are communicated to the IERS, where they are combined. The IERS publishes, among other results, the final values for the five Earth rotation parameters, currently on a monthly and annual basis [8.6, 8.7]. In 2000, uncertainties in the orientation of (I), on 5 day averages, were around $0.000\,05''$ relative to (T), i.e., ± 1.5 mm for the position of the rotational pole at the surface of the Earth, and $0.0002''$ relative to (C). Uncertainties in $UT1 - TAI$, also averaged over five days, were around 10 μs.

8.1.5 Irregularities in the Earth's rotation

If the Earth is treated as a rigid body, its rotational parameters are predictable once a very small number of initial conditions have been established by measurement. According to this hypothesis, the terrestrial pole would follow a circular motion with period 305 days, and its amplitude and phase would only need to be measured once. UT1 would vary linearly with time. Lunisolar precession–nutation, due to the effects of the Moon and Sun on the Earth's equatorial bulge, is certainly a difficult theoretical problem in astronomy, but it has been resolved with uncertainties well below those remaining in observations. Only one parameter would require observations, namely, the dynamical flattening of the Earth.

Discrepancies between observation and the theory of rotation of a rigid Earth arise from displacements in the fluid components of the planet, i.e., atmosphere, oceans, and core, with respect to the lithosphere, and also elastic and viscous deformations of the so-called solid components. These displacements and deformations may cause rotational energy to be lost, as we shall see. In fact, certain irregularities are not yet well understood. Even when the basic mechanism has been ascertained, it is impossible to make quantitative predictions about their effects on the Earth's rotation. We shall restrict discussion to UT1. The reader interested in such geophysical phenomena should consult [8.8, 8.9].

(a) Secular slowing of the Earth's rotation

The earliest available estimates of the length of day D go back 400 million years. Various phenomena have recorded the diurnal, lunar and annual cycles. These phenomena are mainly the growth of corals and bivalves, and the laying down of sediments. The year is a duration that has remained practically unchanged, whereas variations in the lunation, expressed in uniform time, are fairly well known. By counting the number of days in the year and the lunation, it can be deduced that the length of one day has increased on average by about 2 ms per century.

This value agrees with what can be deduced from observations of eclipses in ancient times. The appearance of an eclipse is governed by orbital motions and therefore depends on the uniform Ephemeris Time (see Section 8.2). However, the place from which an eclipse is observed depends on the Earth's rotation. Since the size of this geographical shift is sometimes cause for surpise, we shall outline the way it is calculated.

For example, suppose that D increases by 2 ms per century. From (8.5),

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = -5.35 \times 10^{-22} \text{ rad s}^{-2} . \tag{8.13}$$

The integration in (8.2) is carried out over the last 2000 years, putting $t = -2000$ yr, $t_0 = 0$ (the present), in two different ways. The first time we put $\omega = \omega_0$, and the second,

$$\omega = \omega_0 + \frac{\mathrm{d}\omega}{\mathrm{d}t} t .$$

Taking the difference, it is observed that the cumulated lag since the beginning of Christianity is 4.1 h. The narrow band in which a solar eclipse would have been total at this epoch is thus shifted by 61° longitude towards the east, relative to what would have been calculated if UT1 had been uniform. Information about the Earth's rotation can thus be derived by studying the archives.

The secular slowing down of the Earth's rotation is mainly attributed to ocean tides. Rotational energy is lost in shallow seas. There is also energy transfer to the orbital motion of the Moon which, as a result, is moving away from the Earth by about 4 cm per year. Recent research has revealed significant dissipation in deep oceans [8.10].

(b) Decennial fluctuations

The difference between UT1 and Ephemeris Time can be restituted over the past three centuries or so. Figure 8.3 shows the corresponding values of D. Fluctuations are large enough to hide the secular variation. The details of these

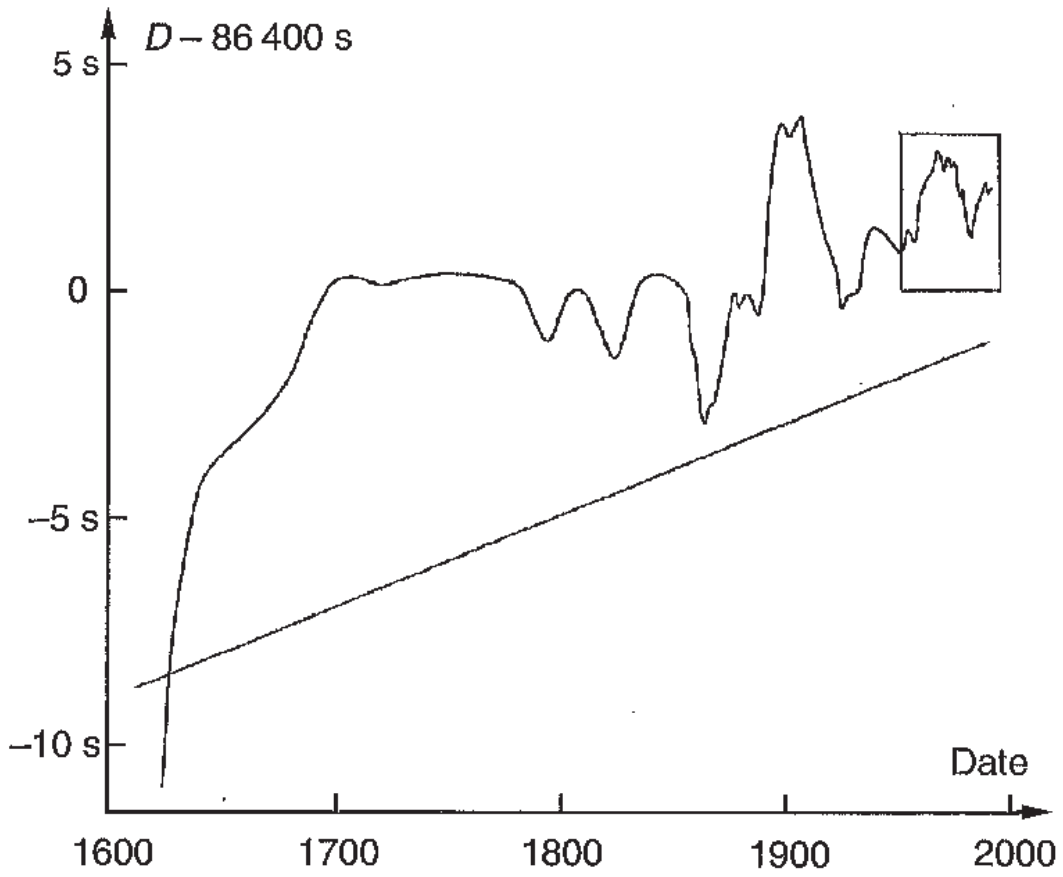


Fig. 8.3. Fluctuations in the length of day D, measured in seconds of Ephemeris Time up until 1955 (from [8.11]), then in atomic seconds (inset). The sloping straight line represents the secular slowing of the Earth's rotation.

fluctuations are becoming more and more clearly revealed. Since 1955, TAI has been available as a much better reference, but the period is too short to study variations with typical period of the order of ten years (inset in Figure 8.3). The origin of these fluctuations is still not well understood, but there is little doubt that they must be due to interaction between the Earth's fluid core and its rigid mantle. Another contribution may also arise from the oceans. They remain totally unpredictable. It is observed that they lead to relative fluctuations in D, and hence in the length of the UT1 second, with magnitude reaching about 10^{-7}.

(c) Short term fluctuations

Figures 8.4 and 8.5 show fluctuations in UT2 and D with the best temporal resolution that can be achieved. These fluctuations are essentially due to atmospheric motions, via angular momentum exchange with the solid Earth. An annual component is recognisable in Figure 8.5. For the main part, this component has been subtracted from the curve in Figure 8.4.

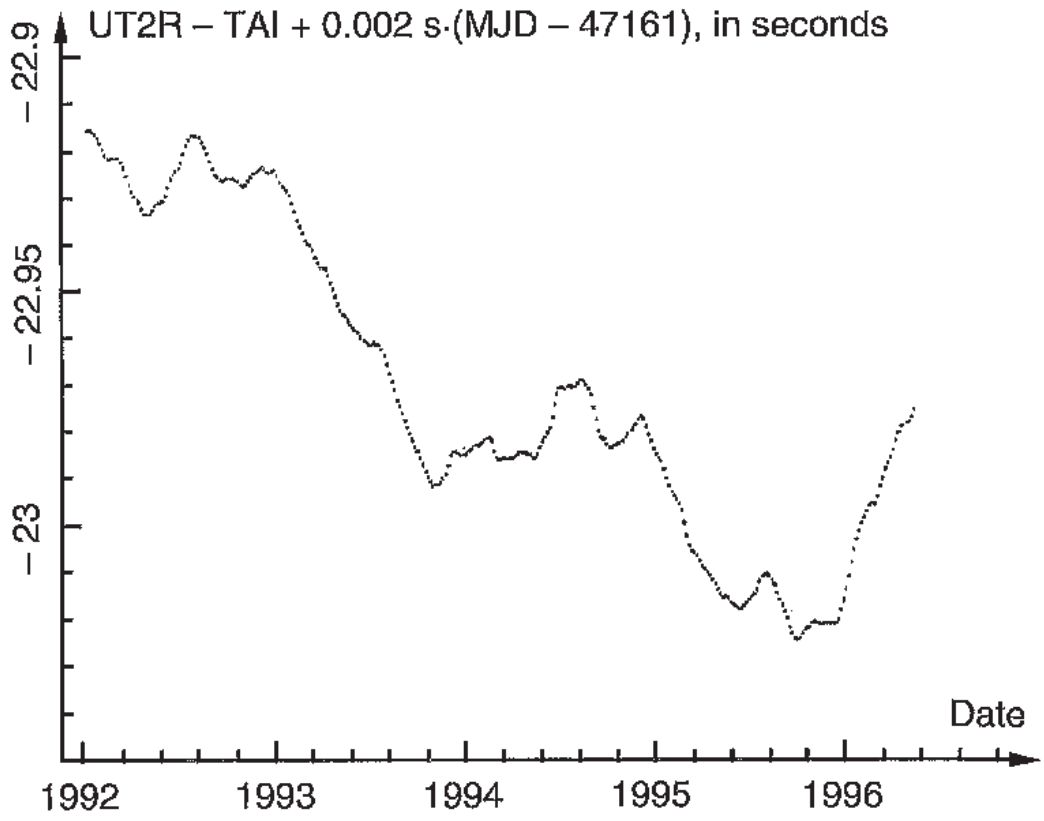


Fig. 8.4. Observed values of Universal Time, after subtracting an annual variation, periodic terms due to tides and a linear function of time. The details are real fluctuations, not observational errors. (Figure kindly communicated by the IERS.)

(d) Frequency instability of Universal Time

The frequency instability of Universal Time is shown in an approximate way in Figure 8.7, at the end of this chapter. In the very long term, it is dominated by secular slowdown of the Earth's rotation.

8.1.6 Use of Universal Time

UT1 is used to adjust Coordinated Universal Time, which requires very little precision. When the utmost accuracy is involved, one of its main uses is to provide a link between celestial and terrestrial frames as a function of TAI (see Section 3.3.2b). It is then used in combination with the other four Earth rotation parameters. These parameters also have an intrinsic value to geophysics. Although it is hardly satisfactory to isolate the factors contributing to a description of the Earth's rotation, we shall add a few brief comments on each.

The non-rotating celestial frame serves to establish dynamical models of the motions of bodies in the Solar System, whether they be natural or artificial, using observations relative to this frame. This produces an *ephemeris* for

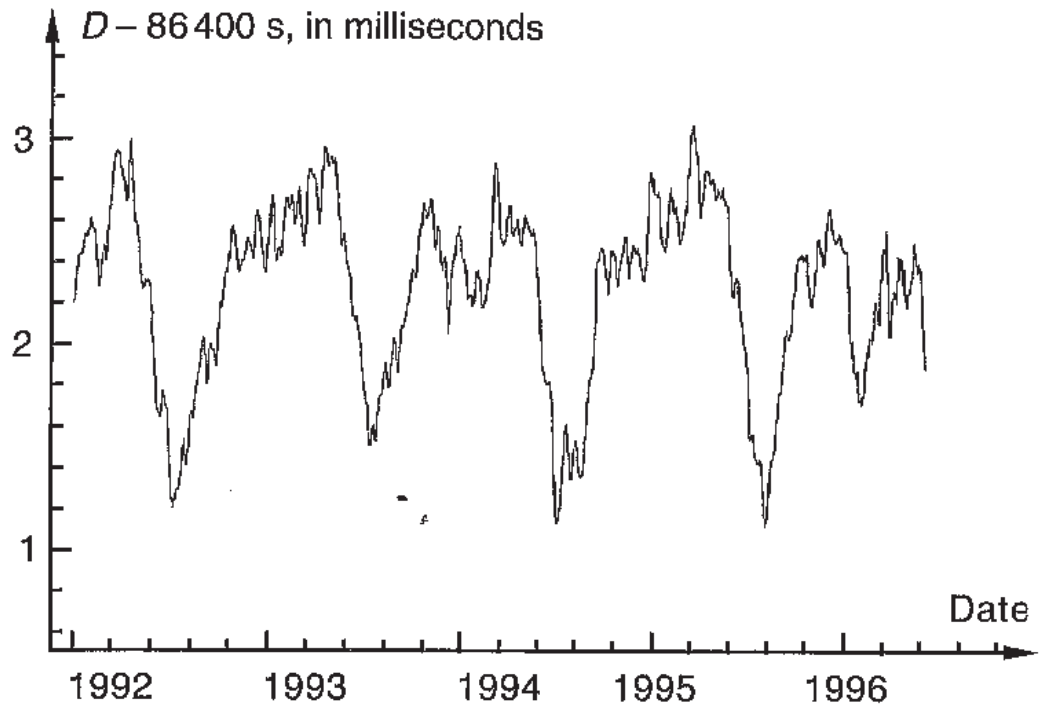


Fig. 8.5. Fluctuations in the length of day D in (atomic) milliseconds, after correcting for the effects of short term zonal tides. (Figure kindly communicated by the IERS.)

the object in question, which allows its position to be predicted. Of course, the existence of this frame in the radio domain of VLBI is not always helpful, practically speaking. But it has been extended to the visible by fixing the angle and angular velocity of the frame formed by 100 000 stars in the programme of the astrometric satellite HIPPARCOS [8.12, 8.13, 8.14, 8.15]. This was achieved by observing stars that radiate in both the radio and visible regions of the spectrum. The ephemerides of the planets and their satellites, as well as space probe orbitography, are crucial to exploration of the Solar System. In addition, astrophysicists find it useful to know the proper (angular) motions of the stars relative to a non-rotating frame.

A highly accurate, homogeneous terrestrial frame has only recently been established. In 1984, the Bureau international de l'heure undertook to publish the coordinates it had adopted for sites used to observe the Earth's rotation, thereby creating a first outline of what was to become the ITRF. National and regional geodetic networks, which may manifest differences of more than 100 m from one to another, are in the process of being fitted to the ITRF. This unification is all the more useful in that the Global Positioning System is based upon a system very close to the ITRF (to within a decimetre) and hence supplies coordinates in this frame. The centimetre accuracy of the ITRF provides a reference in which tectonic deformations can be studied, whether they be large scale motions associated with the tectonic plates, or localised motions in critical regions.

As regards the terrestrial rotation parameters, their importance in geophysics is clear from their unpredictability. Their variations do not always lead unambiguously to a geophysical model. Conversely, however, they give constraints that can lead to rejection of models whose predictions fall too wide of the mark. These parameters also have geometrical importance when transforming coordinates between the celestial and terrestrial systems. For example, VLBI observation of a space probe yields its direction referred to the Earth. In order to organise a rendezvous with another planet, corrections must be made to its path, and this requires knowledge of its position relative to the celestial frame. In this example, the Earth's rotation parameters must be known in real time, which is only possible by prediction. The example shows why good values for these parameters are needed at short notice if the uncertainties of prediction are to be minimised. The IERS has a Rapid Service (in the USA), responsible for predicting the Earth rotation parameters.

8.2 Ephemeris Time

8.2.1 Definition

The conceptual definition of Ephemeris Time ET was given in Chapter 4. It is a form of classical dynamical time. The unit and origin of ET are conventionally defined by adopting a numerical expression for the *geometric mean longitude* of the Sun established by Newcomb, viz.,

$$L = 279°41'48.04'' + 129\,602\,768.13''T + 1.089''T^2 \,, \tag{8.14}$$

where T is counted in Julian centuries of 36 525 *ephemeris days*. The origin of T is at the beginning of the year 1900, at the instant when the geometric mean longitude of the Sun took the value $279°41'48.04''$. This instant of time is dated 1900 January 0, 12 h ET exactly (decision taken by the IAU in 1958).

For those readers unfamiliar with astronomical terminology, let us review briefly. The longitude of the Sun is measured along the ecliptic from the vernal equinox. The latter is the intersection of the ecliptic and the equator and moves in space because of the motion of these two planes. Its motion in the ecliptic is traditionally described by two types of terms, namely, those represented by a polynomial function of time (*secular terms*), and those represented explicitly by periodic functions. The adjective 'mean' refers to the fact that periodic terms have been subtracted off. Likewise, the Sun's motion in the ecliptic, as seen from the Earth, involves periodic components due to the eccentricity of the Earth's orbit together with perturbations from the other planets and the Moon, and a (very small) secular component. Only the secular terms remain in the mean longitude.

Concerning the adjective 'geometric', this refers to the problem of *aberration*. We see the Sun in the position it occupied about 8 minutes earlier, the time required for its light to reach us. This is its *apparent position*. The geometric position is corrected for this delay.

Expression (8.14) contains the definition of the second of Ephemeris Time which is in fact based upon the speed in longitude at $T = 0$. The official definition given in Section 4.3.3 was adopted to bring out a fraction of a duration, but it appeals to Ephemeris Time itself, and this is not defined. The value of this fraction is found by calculating the length of the tropical year in seconds from (8.14), as L increases by 360°, putting $T = 0$ in the T^2 term.

8.2.2 Determining Ephemeris Time

Ephemeris Time was designed and used before atomic time had come into existence, or at least, before it had become widely accepted in the astronomical community. We must therefore place ourselves in this context.

Expression (8.14) can be viewed as the primary definition of ET. The corresponding determination consists in measuring the longitude of the Sun at an instant of time that the observer dates τ_C with his or her clock C. From the longitude measurement, ET is deduced using (8.14) and we thus obtain $ET - \tau_C$. As the uncertainties in UT1 were much smaller than those in ET, the result was given a universal character by specifying

$$\Delta T = ET - UT1 \text{ at observation date .} \tag{8.15}$$

The slowness in the variation and lack of precision in ΔT allowed a great deal of freedom in the choice of reference scale used to date the measure.

In order to measure its longitude, the Sun had to be situated relative to the stars on the sky background. Clearly, this could not be done directly. What was possible, however, was to date solar and stellar transits of the meridian in UT1 and use the Earth's angular velocity to deduce the difference between their right ascensions. Measuring their zenithal distances during these transits gave differences of declination needed to evaluate the longitude of the Sun. Many corrections were needed. One was for aberration, as we have seen, but there was also a correction for parallax, since the longitude in (8.14) is geocentric, and corrections for yet other astronomical phenomena. In fact, this primary determination of ET is extremely difficult and also very imprecise. If we assume that the likely uncertainty in L is $0.5''$, the corresponding uncertainty in ET from observation of the Sun is 10 s. If we hope to measure a time interval in ET by two observations of the Sun, with relative uncertainty of 10^{-8}, these observations must be made more than 50 years apart.

Other approaches had therefore to be explored. Astronomers who set up geocentric ephemerides, giving angular positions and distances of bodies in the Solar System, had to choose the unit and origin of their time parameter so that it satisfied the condition expressed by (8.14). Once this was done, all such ephemerides had ET as their argument. In principle, observation of any of these bodies gave access to ET, with a possible gain in precision. The fact nevertheless remained that this way of finding ET depended on a sort of calibration via solar observations.

The faster the orbital motion of the observed body, the greater is the precision in ET. In this respect, the Moon is the most suitable object. Unfortunately, the motion of the Moon is perturbed by the Earth's rotation (see Section 8.1.5a). Its mean geocentric angular velocity (or *mean motion*) is gradually slowed down in a way that can only be ascertained by observation. Variations in the Moon's mean motion could only be obtained if ET was first determined via observations of the Sun and planets. After this delicate calibration exercise, which required decades of tedious work, the Moon could then be used as a practical way of obtaining ET. The Moon was a kind of secondary clock that allowed greater precision in readings, but at the risk of introducing systematic errors that would increase with time, stemming from some error in its calibration.

To guarantee a unique realisation of ET, the IAU recommended the use of a specific lunar ephemeris and the notation ETj, where the suffix j indicates the one chosen. (From 1960 to 1984, $j = 0,\ 1,\ 2$ were used.)

The Moon's motion against the star background was measured by observing occultations of stars by the Moon, or else by using photographic methods. Large scale observation campaigns were organised at the end of the 1950s. Despite the considerable effort put in, these techniques still gave uncertainties of several hundredths of a second in annual averages of $\Delta Tj = ETj - UT1$. When TAI was set up, these determinations of ET came to an end, for it was simpler to adopt

$$ET - TAI = 32.184 \text{ s}\,, \tag{8.16}$$

a conventional constant value based on observation.

8.2.3 Difficulties and current solutions

Apart from the uncertainties mentioned in Section 8.2.2, large systematic errors affected ET. It was very sensitive to the way the celestial coordinate system was realised and also to the values adopted for various astronomical constants.

For example, catalogues giving positions and proper motions of stars, and realising a non-rotating geocentric frame (earlier forms of the frame (C) discussed in Section 8.1.1), had to be revised several times due to improved measurements and also because uncertainties in proper motions degrade accuracy as time goes by. Each revision caused ET to jump. To illustrate the part played by the astronomical constants, we may consider the aberration constant relating the mean longitude in (8.14) to the observed apparent longitude. The change in recommended value applied by the IAU in 1968 led to a jump of 0.6 s in ET.

In 1994, to overcome the difficulties inherent in defining ET by means of the solar longitude, the IAU simply recommended that the realisation of Terrestrial Time TT defined in Section 7.1.2 be extended backwards in time to the period before 1955, via the dynamics of the Solar System. Indeed, the TT scale unit has duration equal to that of the atomic second on the rotating geoid and the latter was defined in such a way as to agree with the best available determination of the ephemeris second. Moreover, the origin of TT was chosen to give continuity with ET. Hence, replacing ET by TT does not lead to any great discontinuity in the construction of the ephemerides.

In this modern application of the Ephemeris Time concept, which allows for relativistic dynamics, the scale unit is fixed by the present atomic definition of the second, and the origin by relation to International Atomic Time. Each ephemeris k nevertheless leads to a different approximation TT(k) of TT that must be clearly stated (IAU Resolution C7, 1994).

In Section 4.3.4, we mentioned the rather historical role played by ET. Figure 8.7 gives some idea of its frequency stability.

8.3 Pulsar Time

Pulsars were first observed in 1967 [8.16]. These objects send out radio pulses at regular intervals. It is thought that they are stars composed mainly of neutrons in a state of ultimate compression and emitting a beam of radiation in a fixed direction relative to the matter within them. The pulses of radiation received here on Earth are thus witness to their rotation.

All pulsars discovered between 1967 and 1982 have rotation periods of the order of one second. The frequency at which their pulses arrive on Earth is affected by changes in distance between the pulsar and the observer, and in particular, by the Earth's orbital motion. After correcting and referring to the atomic time scale, relative frequency instabilities for these pulses lie in the range 10^{-10} to 10^{-11}. However, in 1982, astronomers added another category of pulsars to their collection of celestial objects. These were the *millisecond pulsars*, with periods of a few milliseconds. The diameters of these objects are

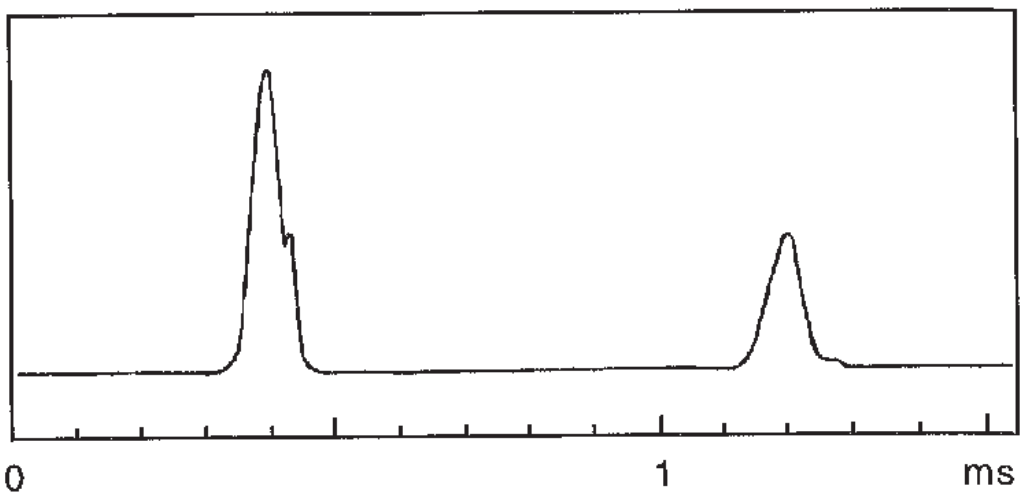


Fig. 8.6. Mean shape of pulses from the pulsar PSR1937+21. The presence of an intermediate pulse suggests that the magnetic axis of the pulsar is perpendicular to the axis of rotation.

estimated at about 20 km and their masses at around 1.4 times the mass of the Sun. The first was discovered by D.C. Backer and his team using a huge fixed radio antenna, 300 m in diameter, at Arecibo in Puerto Rico [8.17]. This pulsar PSR1937+21 has a period of 1.6 ms. The shape of its pulses is shown in Figure 8.6. It was soon realised that the arrival of pulses from PSR1937+21 could be dated relative to TAI with an uncertainty of a few tenths of a microsecond. In addition, once the gradual slowdown in its rotation had been taken into account, it had a long term frequency stability at least as good as that of the best atomic clocks running at the time (instabilities of 10^{-14} or less over one year). Several dozen other millisecond pulsars were subsequently discovered, some of them being components of multiple star systems. An international programme was set up to time pulse arrivals. Observations were made at frequencies of the order of 1 GHz, using large radio antennas. The power collected is very low, of the order of 10^{-16} W with the largest antennas.† This programme does not require continuous observation. The frequency stability of both the pulsars and TAI is such that observations may be interrupted for periods of a few weeks or months without risk of losing a single period.

Some have suggested that millisecond pulsars will bring time measurement back within the province of astronomy. However, this would be to forget the spin-down effect just mentioned, which rather limits the usefulness of measuring time by pulsar rotation. In the case of PSR1937+21, for example, the

† In this field, the unit of flux density used is the Jansky, where 1 Jy = 10^{-26} W Hz^{-1}m^{-2}. At typical observation frequencies, values of 10 to a few 100 mJy are obtained.

period P increases by 10.5×10^{-20} s per second, so that in relative values

$$\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}t} = 7 \times 10^{-17}\ \mathrm{s}^{-1}\ . \tag{8.17}$$

This relative spin-down is ten times greater than the same for the Earth. Although it is thought that $\mathrm{d}P/\mathrm{d}t$ is constant, its value cannot be known a priori with sufficient accuracy to rival the frequency accuracy of atomic clocks. There is a further problem. The rotational phase of the pulsar must be transferred to Earth and uncertainties enter as a result of the relative motion of observer and source. For example, observations depend on the position of the pulsar, because of the orbital motion of the Earth, and its position cannot be determined accurately enough. In addition, pulsars have their own proper motion, and possibly orbital motion if they are in binary systems. There may also be fluctuations in propagation times if the signal crosses ionised media in space. For these reasons, pulsar rotation cannot be made the basis for defining a time unit, and nor can it be used as a clock to realise the time scales, e.g., TCG, used in the theory.

As a consequence of the slowing rotation and instabilities introduced when evaluating variations in the propagation time, we must reconsider what is meant by the frequency instability of a pulsar. In the following, the *intrinsic instability* will be taken to mean the hypothetical instability after making ideal corrections for the constant $\mathrm{d}P/\mathrm{d}t$, as it would be observed in a frame moving with the pulsar. It is this intrinsic instability that is considered to be smaller than the frequency instability of atomic clocks.

We may then imagine the possibility of calibrating pulsar clocks using atomic clocks, by determining their frequency drift over a time interval of duration I. We might be able to use these to improve the frequency stability of TAI retrospectively for sampling periods shorter than I, by forming a mixed time scale [8.18]. Unfortunately, random uncertainties in pulse arrival times require I to be very long. For example, if atomic time has a frequency instability of 10^{-14} over one year, then with the help of the pulsar giving the most suitable pulses that can be dated to within 0.3 μs, we can only hope for better stability in the mixed scale when sampling times lie between one year and I. Recent advances with atomic clocks would mean extending I to several decades, which seems to spell the end for this idea. A *Pulsar Time* has nevertheless been envisaged. It would be constructed by means of an algorithm combining observations of a group of pulsars whose comparison with atomic time might in the long run lead to new discoveries [8.18].

A rather different contribution of millisecond pulsars to time measurement has also been studied. By measuring the Doppler effect, the timing of these

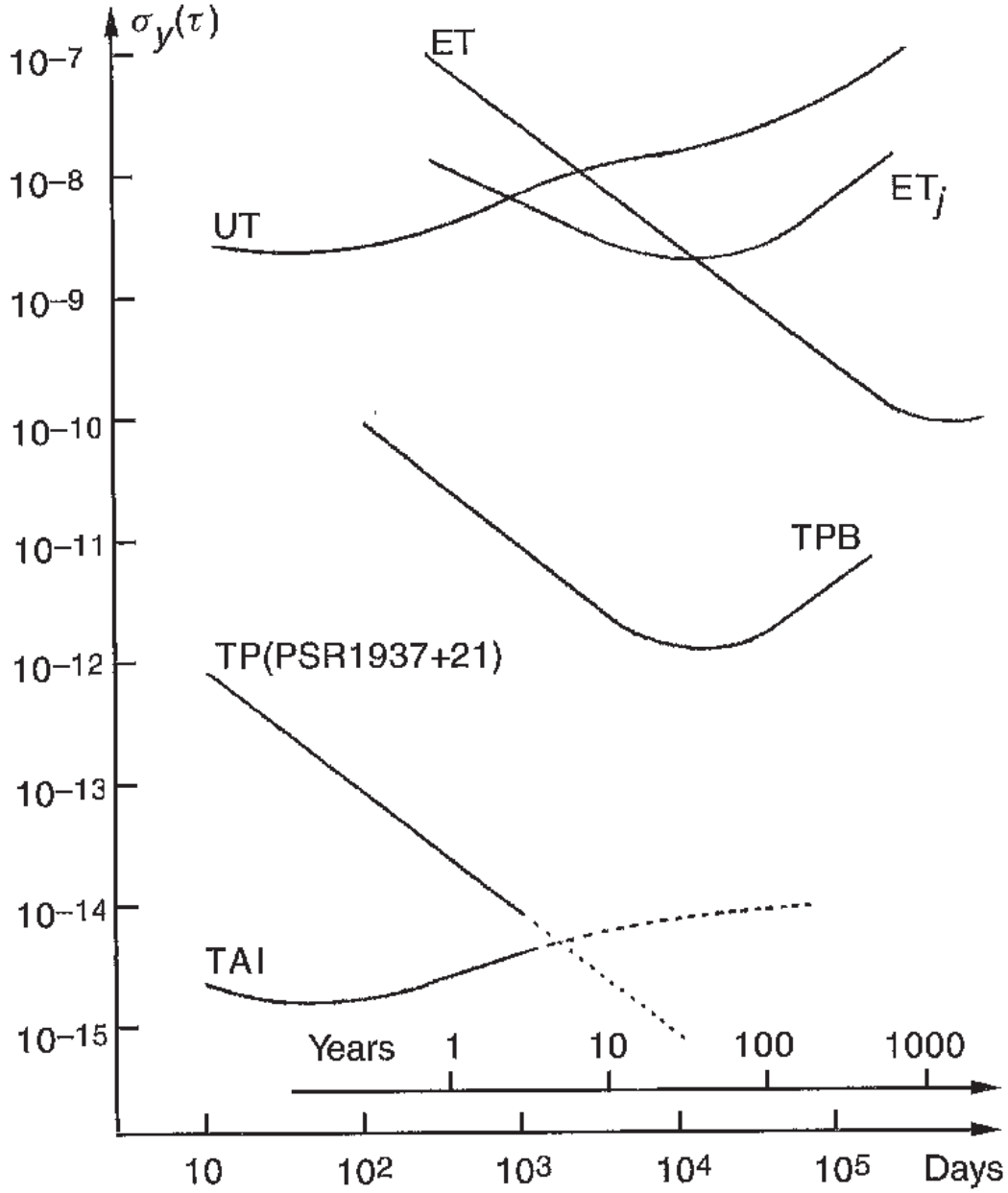


Fig. 8.7. Instability of astronomical time scales and TAI: ET and ETj are ephemeris time scales set up by observing the Sun and Moon, respectively; UT is Universal Time; TP (PSR1937+21) is time based on the rotation of the named pulsar; TBP is time based on the orbital motion of a binary pulsar (optimal case). Concerning TAI, properties for 1998 are projected into the future. This figure is based upon elements from [8.18] and [8.21].

pulsars provides the Earth's position on its orbit and hence leads to a new measure for Ephemeris Time. There is much greater precision in readings, compared with observations based upon the definition (8.14) of Ephemeris Time in terms of the Sun's longitude. Frequency stability is also improved for sampling periods less than six months. But in the long term, which is the most important, nothing is gained over traditional observation methods [8.18].

A third method for measuring time based on pulsars consists in monitoring the orbital motion of a pulsar with a binary companion, by observing the Doppler effect. The theory for such motions is exceedingly complex, because gravitational fields are very strong and effects of gravitational waves are significant. However, it can be developed. The question remains as to whether a dynamical time scale based on this motion can rival atomic time. In the study [8.18], the authors reach the following conclusion. The discovery, observation and interpretation of binary pulsars might have occurred before the emergence of atomic clocks (since quartz clocks would have been sufficient for this purpose), and in this case, binary pulsar time would have been a good candidate for replacing Ephemeris Time. However, this measure of time cannot take the place of atomic time, which has greater accuracy and stability.

It thus seems likely that millisecond pulsars will provide neither a unit of time nor the time we read on our clocks. Notwithstanding, up until now, instabilities in International Atomic Time have to some extent limited the quality of scientific results that can be obtained by timing pulsars. Indeed, this is what justifies the retrospective calculation of more accurate and stable atomic time scales than TAI, as discussed in Section 7.4.3. New frequency standards may render the imperfections in atomic time negligible, but only time will tell.

Pulsar timing is the most demanding application of atomic time scales, with regard to their accuracy and their long term stability (over six months or more). It is bringing new results in astrometry, Solar System dynamics and the theory of gravitation itself. The timing of binary pulsars has indirectly confirmed the existence of gravitational waves, predicted by general relativity, and it has thereby strengthened confidence in the theory [8.19, 8.20]. This research was rewarded by the Nobel Prize for Physics in 1993, attributed to R.A. Hulse and J.H. Taylor.

Frequency instabilities in times related to pulsars are illustrated in Figure 8.7. The figure gives a rather qualitative overview of the problem. For astronomical times, it gives only the likely orders of magnitude. For atomic time, it assumes that characteristics of standards will remain at the same level.

9

Ultraprecise time and frequency applications

Whether it concerns frequencies or times on clocks, time measurement is omnipresent today. An ordinary wristwatch used to organise our everyday activities is a precision instrument. Driven by a quartz crystal, its frequency instability over one day may be less than 10^{-7}. Some watches are even servo controlled by time signals which are themselves linked to the best caesium standards. We may thus wear an instrument whose frequency remains accurate to a few parts in 10^{-14}, provided that we average over millions of years. In our own houses, we are surrounded by time and frequency sources: time switches, computer clocks, radio receiver oscillators, cordless phones, and so on. Many systems and services depend on time: telecommunications, television, electricity distribution, positioning.

It would be an interesting exercise to choose a relative frequency inaccuracy, say 10^{-10}, and determine which activities would become impossible if no frequency standard were able to go beyond this limit. It is clear that the Global Positioning System and its host of applications could not exist. International Atomic Time would not exist either, since astronomy would have retained control over the world's time scales. Space geodesy would not be precise enough to make a contribution to geophysics. The same exercise could be carried out at different levels of accuracy. Unfortunately, we must admit that we are unable to provide exhaustive consideration of the different cases. We can only invite the reader to reflect upon the historical evolution of techniques, particularly those concerned with space and telecommunications, where time measurement is a key element.

As an example, we shall examine several applications of time measurement using the best standards, either the most accurate ones, or the most stable for various sampling periods, or again those providing a good compromise between accuracy and stability on the one hand, and reliability on the other. Indeed, in a great many applications, we must accept a slight degradation with

respect to the quality of laboratory standards if we are to obtain reliable and convenient instruments. For such applications, commercially manufactured rubidium and caesium clocks have been produced in their thousands, and even hydrogen masers are sometimes used.

What is the purpose of all these instruments? Their main use centres upon three different operations. The first is the measurement of time intervals and frequencies, for example, when measuring distances by radar or laser ranging. The second arises when systems are based on clock synchronism, as in the case of the Global Positioning System, or on equality of frequencies (syntonisation). The last occurs when a time scale is used to represent the time in some physical theory, for example when modelling the orbits of satellites and planets. However, this way of breaking time measurement down into different categories cannot serve as a basis for systematic presentation of the subject, for the categories tend to overlap. We shall therefore illustrate our discussion by selecting some practical and scientific themes which involve time measurement in characteristic ways.

9.1 Fundamental research

9.1.1 Metrology

The second is by far the best realised of the base units in the International System of Units (SI units). It is therefore quite natural to define the other base units in terms of the second using only the laws of physics and the fundamental constants they involve. By accepting the postulate that such constants do indeed exist (i.e., Einstein's principle of equivalence), all the accuracy in the realisation of the second is thereby, in principle, carried over to the other units. This was the step taken in 1983, when the definition of the metre was based on the speed of light. Of the seven base units in the International System (see Appendix 3), there are now only three that remain independent, viz., the second, the kilogram and the kelvin. For example the definition of the ampere is based upon the notion of force in [m kg s^{-2}] and distance in [m] between two conductors. This definition refers in the end to the second and the kilogram.

A significant advance would be made if the kilogram could be defined in terms of the second. The present definition based upon the mass of a particular object, a cylinder of platinum–iridium alloy kept at the BIPM, gives rise to a relative uncertainty barely less than 10^{-8}. It may be possible to achieve this via electromagnetic or electrostatic forces. All measures, apart from those appealing to thermodynamics, would then be related to the second.

This process whereby the measure of a quantity is transposed into a measure of time also applies to voltages using the Josephson effect, and to electrical resistances, via the quantum Hall effect. However, these quantum effects cannot serve as a basis for new definitions of the units used in electricity, at least, not for the moment, because the link with mechanical units would then be lost: the mechanical watt must remain equal to the electrical watt. For the time being, they are used to maintain the values of the volt and the ohm, with these values still being related to the definition of the ampere.

The Josephson effect occurs when two superconductors are separated by a very thin insulating barrier. When subjected to ultrahigh frequency radiation of frequency ν, this junction exhibits a stair-shaped current–voltage characteristic, levelling out at specific values V_n of the applied potential difference. These are given theoretically by the relation

$$V_n = n\frac{h}{2e}\nu \,, \tag{9.1}$$

where n is a whole number, h is Planck's constant and e is the electron charge.

The uncertainty in the measurement of the ratio $h/2e$ afforded by the above equation is smaller than the uncertainty arising when h and e are found by separate experimental determinations. In 1988, following an international campaign of measurements, a conventional value denoted K_J was adopted for $h/2e$. This is given by

$$K_J = 483\,597.9\ \text{GHz}\,\text{V}^{-1} \,. \tag{9.2}$$

This provides a stable reference for potential differences, tied to a frequency measurement.

The quantum Hall effect, also known as the von Klitzing effect, occurs when a thin semiconducting layer is cooled to about 1 K and subjected to a magnetic field of the order of 15 T, perpendicular to the layer. The Hall voltage V_t, which is orthogonal to the direction of the applied current I, behaves in such a way that the resistance $R = V_t/I$ is given theoretically by

$$R = \frac{1}{n}\frac{h}{e^2} \,, \tag{9.3}$$

where n is once again a whole number. Experimental determination of the quantum Hall resistance yields a value for h/e^2 that is more precise than can be obtained from separate determinations of h and e, and in 1988, a conventional value R_K was adopted for this ratio, viz.,

$$R_K = 25\,812.807\ \Omega \,. \tag{9.4}$$

Since the beginning of the 1990s, it has been possible to send electrons one

by one through a potential barrier. It has therefore become feasible to create an electric current whose intensity would be determined by the frequency of the signal controlling the opening of the barrier. It would then be possible to check the mutual consistency of methods for determining the electrical units, since the intensity of a current is also related to the quantity frequency through the Josephson and quantum Hall effects.

The conventional values of K_J and R_K implicitly define electrical units that do not strictly conform to official SI units. However, they have the advantage of providing a uniform representation of the volt and the ohm with far greater reproducibility than can be had from absolute determinations based upon the definition of the ampere.

9.1.2 Measuring physical constants

Some constants can be determined simply by measuring the frequency of a physical phenomenon, and their values can therefore be obtained with very low uncertainty.

The Rydberg constant is an example. It has been ascertained to within 7.7×10^{-12}. This result is the outcome of work carried out by teams at the Ecole normale supérieure in Paris (France) and the Max Planck Institute in Garching (Germany). Using methods of high resolution spectroscopy, they measured the frequency of several optical transitions in the hydrogen atom. The value obtained was [9.1]

$$R_\infty = 109\,737.315\,685\,50\ \text{cm}^{-1}\ . \tag{9.5}$$

The Landé factor g for the electron, close to 2, has been measured at the University of Washington in Seattle by confining a single electron in a Penning trap. This is produced by superposing a uniform magnetic field on an inhomogeneous electric field. The value of g is equal to twice the ratio of the electron's Larmor precession frequency to its cyclotron frequency. The value obtained is given by [9.2]

$$\frac{g}{2} = 1.001\,159\,652\,188 \tag{9.6}$$

with uncertainty of 4×10^{-12}.

Applying the same technique to a positron, the same value of g was obtained, to within 2×10^{-12}. This result is currently the most accurate confirmation available of the hypothesis that antiparticles have the same properties as their corresponding particles.

By comparing the cyclotron frequencies of protons and antiprotons in a Penning trap, it has been shown that their inertial masses are equal, to within 4×10^{-8} [9.3]. Using a similar technique, the ratio of the proton and electron masses has been measured to within 3×10^{-9} [9.4].

As can be seen from (9.6), the value of the electron Landé factor differs from 2 by only a tiny amount. The deviation from this value depends on quantum electrodynamic (QED) effects and can be expressed in terms of the fine structure constant α. This is one of the most important constants in atomic and quantum physics. Recent theoretical progress has made it possible to use the measurement of g to obtain [9.5]

$$1/\alpha = 137.035\,999\,44\,, \tag{9.7}$$

to within 4×10^{-9}.

9.1.3 Impact on atomic physics

(a) Checking linearity of quantum mechanics

The basic equation in quantum mechanics is the Schrödinger equation (6.27), which is linear. However, it has been shown theoretically that there may be a slight non-linearity [9.6]. The transition frequency between two energy levels would then depend on the population of these levels, making the transition slightly anharmonic. Experimental tests have been carried out, one with Be^+ ions in a radiofrequency trap [9.7], and the other with hydrogen atoms in a hydrogen maser [9.8]. In both cases, the transition frequency between the relevant hyperfine sublevels was found not to depend on their populations, up to experimental uncertainties. Since these uncertainties were extremely low, thanks to the very low frequency instability of the devices used for the tests, we may conclude that a non-linear generalisation of the Schrödinger equation is unnecessary.

(b) Understanding properties of atoms and molecules

The contribution here is important and varied and a detailed description would go beyond the scope of this book.

Techniques of high resolution spectroscopy developed and used in atomic or molecular frequency standards have made it possible to measure a great many transition frequencies with very low uncertainties. Examples are the hyperfine transition frequencies of hydrogen and its isotopes, rubidium isotopes of mass 85 and 87, and many ions with a single valence electron (Section 6.7). Likewise, atomic and molecular transition frequencies in the infrared and the visible are now known with uncertainties as low as a few times 10^{-12}.

These determinations, together with investigations of electric and magnetic field effects, have greatly advanced our knowledge of atomic constants, such as the nuclear and electronic gyromagnetic ratios, and interatomic couplings in molecules.

A better understanding of the interatomic potential has been obtained by measuring the spin-exchange collision cross section and spin-exchange frequency shift cross section of the hydrogen atom in the range 1–373 K. Studies are underway to compare theoretical and experimental values of these cross sections for caesium and rubidium atoms with low relative velocities. These studies are motivated by the use of laser-cooled atoms in clocks.

An unexpected contribution to the shift in the hyperfine transition frequency of rubidium atoms has been observed when they are subjected to a strong magnetic field of around 5 T [9.9]. The result implies that the magnetic interaction energy between the nucleus and the electron cloud is not sufficiently well known and that further terms must be evaluated [9.10]. These might well affect the value of the frequency shift due to the second order Zeeman effect, a value that must be taken into account in atomic time and frequency standards. This question could be investigated in experiments with a caesium beam frequency standard [9.11]. This standard would be unusual in that the transition

$$F = 3\,, \quad m_F = -1 \quad \longleftrightarrow \quad F = 4\,, \quad m_F = -1$$

is used as the reference transition. Its frequency has an extremum for an applied magnetic field close to 82 mT. Precise measurement of the transition frequency, close to 8.901 GHz, should provide a way of checking that the Breit–Rabi formula (6.4) contains all the terms needed to calculate the Zeeman frequency shift.

9.1.4 Spacetime structure and gravity

Most tests for models of spacetime structure and gravitation involve time measurements. An exhaustive discussion of the significance of these tests can be found in [9.12, 9.13]. We will consider a few examples of experiments in which atomic time standards play a critical role. Certain tests are based directly upon the stability or accuracy of time standards [9.14]. In other cases, these standards are essential auxiliary tools.

(a) Direct use of atomic time standards

Several clock transportation experiments have been organised to check the predictions of general relativity. In particular, in 1972, Hafele and Keating of the US Naval Observatory arranged for transportation of four caesium clocks

around the world on commercial flights. In the westward journey, they observed a gain of 237 ± 7 ns, and in the eastward journey, a loss of 59 ± 10 ns. Theoretical predictions gave a gain of 275 ± 21 ns and a loss of 40 ± 23 ns, respectively, tainted by uncertainties arising from the poorly identified trajectories followed by the clocks [9.15]. This experiment was considered as a check on frequency shifts due to gravity, but it was not very informative, because Pound, Rebka and Snider had already made the same check to better accuracy using the Mössbauer effect [9.16, 9.17]. However, along with several analogous experiments, it was a convincing argument for treating time measurement within a relativistic framework.

A further measurement of the gravitational frequency shift was made by Vessot and Levine in 1976 [9.18]. A hydrogen maser was launched in a probe, up to an altitude of 10 000 km. During the flight, its phase was continually compared with the phase of a ground-based hydrogen maser, using a three-way link to eliminate the Doppler effect. The observed frequency shift was found to agree with the predictions of general relativity, within the limits of experimental uncertainty, viz., a relative uncertainty of 7×10^{-5}. More accurate experiments are planned along the same lines, either in satellites with highly elliptical orbits around the Earth, or on a space probe flying towards the Sun. These experiments are interpreted as tests of the Einstein Equivalence Principle (EEP) [9.12].

By comparing a caesium clock, based upon a hyperfine structure transition, with a magnesium clock, based upon a fine structure transition, operating side by side, Godone *et al.* [9.19] were able to place limits on the time evolution of the product $g_p(m_e/m_p)$, where g_p is the gyromagnetic ratio of the proton and m_e, m_p are the electron and proton masses, respectively. This (relative) limit is equal to 5.4×10^{-13} per year. Combining with astrophysical data, these authors have shown that the relative time variation of the fine structure constant α is less than 2.7×10^{-13} per year. During the same experiment, the clocks underwent an annual variation in the gravitational potential of the Sun, due to the eccentricity of the Earth's orbit. If EEP were violated, an annual variation would have been found in the frequency difference between the two clocks. This was not the case [9.20]. In more recent work, comparison between a hydrogen maser and a trapped $^{199}Hg^+$ ion frequency standard gives the upper limit [9.21]

$$\frac{1}{\alpha}\left|\frac{d\alpha}{dt}\right| < 3.4 \times 10^{-14} \text{ per year} .$$

Furthermore, taking advantage of the remarkable frequency stability of caesium and rubidium fountains, measurements spread over an interval of 24

months reduced the upper limit for the possible variation of α. We now have [6.23]

$$\frac{1}{\alpha}\left|\frac{\mathrm{d}\alpha}{\mathrm{d}t}\right| < 3.1 \times 10^{-15} \text{ per year}.$$

These last two results are based upon the fact that the theoretical expression for the hyperfine transition of ground state alkali atoms (such as H, Rb, Cs, Hg^+) contains correction terms depending on the product αZ, where Z is the atomic number of the element. Any time variation in α would induce different changes in the transition frequency under consideration.

One postulate of relativity theories is the constancy of the speed of light c when measured locally. Let us note in passing that this constancy can only be checked experimentally if we have independent standards for time and length. For this purpose, we require material length standards, such as rulers, interferometer arms or a geodetic base. The constancy of c is generally checked by a two-way propagation. This was the case in the famous Michelson–Morley experiment [9.22], which was unable to detect any effect due to the motion of the observer relative to a hypothetical ether. However, we should also check that the outward velocity is equal to the return velocity, or more generally, that the velocity is indeed isotropic. This was done at the Jet Propulsion Laboratory (USA) in 1990, using the frequency stability of two hydrogen masers placed 21 km apart, compared by one-way transmissions in optical fibres [9.23]. Whilst the baseline of the two masers was displaced by the Earth's rotation, no anisotropy was revealed, within the limits of experimental uncertainty, so that $|\Delta c|/c < 3.5 \times 10^{-7}$. It has been proposed to repeat this experiment using a spaceborne maser [9.24]. This should reduce the relative uncertainty to 1.5×10^{-9}. In the same area of investigation, long term comparison of clocks on board GPS satellites with clocks in ground stations has led to the upper bound $|\Delta c|/c < 4.9 \times 10^{-9}$ when all directions are considered and $|\Delta c|/c < 1.6 \times 10^{-9}$ for the equatorial component [9.25].

(b) Indirect use of atomic time standards

In other tests of relativity, the quality of time measurements is exploited for the purposes of highly accurate positioning. Radar telemetry has thus been applied to measure the advance of the perihelion of Mercury more precisely than was possible by optical astrometry using angle determinations [9.26]. The relative inaccuracy in this historic test of Einstein's theory has been reduced to 10^{-3}. Laser ranging on lunar reflectors made it possible to seek a possible violation of the weak equivalence principle for massive bodies (equivalence of inertial and gravitational masses). This putative violation goes by the name

of the *Nordtvedt effect* [9.27]. Other research has been carried out on possible time variations of the gravitational constant, based upon precise measurements of the distances of artificial satellites placed in orbit around Mercury.

As discussed in Section 8.1.4a, angular measurements can be made with great accuracy using VLBI. This method has been used to improve another classic test of general relativity, the deflection of light by matter. It has been possible to measure the bending of quasar light by the Sun, and even by Jupiter [9.28].

All these experiments require detailed models of orbital motions. To this end, the atomic time scale TAI is used to represent the time parameters of the theory, as defined in Section 7.1. Since observations are made from the rotating Earth, the parameters describing this rotation are also needed to very great accuracy. However, we have seen that these parameters are determined on the basis of time and frequency techniques. It is quite clear that time metrology is implicated in such work in an extremely complex manner.

One new form of test involves measuring the time delay of signals from a celestial object when they pass close by the Sun (the *Shapiro time delay*). Time enters through positioning (VLBI), the orbital model, stability of Earth-based and spaceborne clocks and by the method of radar ranging. This method has been applied to signals from a millisecond pulsar.

We should also record here research involving the timing of millisecond pulsars, discussed in Section 8.3.

(c) Ultimate aims of these tests

Since the publication of Einstein's general relativity, many other gravitational theories have been devised. In most of these theories, gravitation is viewed as in general relativity as a manifestation of the structure of spacetime itself, described mathematically by a metric. Yet other theories assume the existence of fields in addition to the metric, or even replacing it.

In weak gravitational fields and for small speeds relative to the speed of light, post-Newtonian metrics are used, e.g., (3.14) at a low level of approximation. Following work by Nordtvedt [9.29] and Will [9.30], the various competing theories have been encoded by introducing ten parameters known as the PPN parameters, where PPN stands for *parametrised post-Newtonian* formalism. These parameters are used to represent the post-Newtonian metrics in a single expression, but they take different values in the different theories. Observation then leads to estimates of their values, or the value of one of their combinations. For example, we have seen that the relative gravitational frequency shift is $\Delta U/c^2$ according to general relativity. Its PPN expression is $(1 + \alpha)\Delta U/c^2$. The parameter α (not to be confused with the fine structure

constant) is then a measurable quantity, which general relativity predicts to be zero. Vessot and Levine's experiment shows that $|\alpha| < 2 \times 10^{-4}$.

Up until now, tests carried out in the Solar System, for weak fields and low velocities, have led to PPN parameters agreeing with general relativity, within the limits of experimental uncertainty. This only means that we have not yet been able to discriminate between theories whose consequences may differ when we model the Universe and its evolution. This is why tests must involve strong gravitational fields, such as those in the vicinity of binary pulsars. Likewise, observation and modelling of astronomical objects like black holes are of key importance in cosmology.

Cosmology is indeed the ultimate aim of all these investigations. Here is one of the great human dreams, and are not such dreams the glory of humankind? In the same spirit, the search for a single unified theory covering all macroscopic phenomena as well as the world of elementary particle physics has not yet borne its fruits. A better understanding of the structure of spacetime and gravity would be a significant advance.

For those who are not convinced of this claim, or for those who invest their time and money in the hope of a more immediate return, we may add that the relativistic theoretical framework has become an absolute necessity for applications. An example is ultraprecise positioning. In addition to its obvious use in transportation, it is crucial for measuring large scale deformations of the Earth's crust (relevant to earthquake prediction) and monitoring sea levels around the world as the climate evolves. A sound experimental basis for the theory is therefore of the utmost importance.

9.2 Positioning, geodesy and navigation

9.2.1 Principles of time and frequency methods

The idea of determining distances by measuring the two-way propagation time of a signal is widely used in astronomy and geodesy, using laser ranging and radar, not to mention many applications in other fields based on the same principle. A three-dimensional positioning method is then possible, provided that three targets are available and the observer is not situated in the same plane. This is the principle governing ultraprecise geodesy by laser ranging artificial satellites.

In satellite positioning systems, it is more convenient for an observer to use a one-way transmission from the satellite. This also requires much less powerful emissions. However, the inevitable absence of synchronisation between transmitter and receiver must somehow be compensated. Two solutions have

been used for worldwide systems, and also for global geodesy. One is based on Doppler shifting of frequencies transmitted by the satellite. The other uses the *pseudo-range* calculated from the difference between the reception time as measured by the receiver clock and the transmission time as measured by the satellite clock, thereby including the synchronisation error.

In order to give a simple description of the Doppler method, let us assume that the frequency ν_0 of the transmitter is known to the observer and that the latter is fixed relative to the Earth's surface. We shall not consider relativistic effects here, although in practice they should be taken into account. The observer measures the received frequency $\nu_{0\mathrm{R}}$ and calculates the difference $\Delta\nu = \nu_{0\mathrm{R}} - \nu_0$. This difference is integrated over a time interval $\Delta t = t_2 - t_1$. We then obtain the number n of periods $P = \nu_0^{-1}$ lost (when $n < 0$) or gained (when $n > 0$) at reception, depending on whether the satellite is moving away or approaching, respectively. The quantity cnP is just the difference between the distances to the satellite at dates t_2 and t_1. If the positions of the satellite are known at these dates, in a frame fixed relative to the Earth, the observer's position must lie on a well-defined hyperboloid in this frame. As the satellite continues its orbit, measurements are repeated and the hyperboloids produced in this way have only one common point, which must be the position of the observer. However, in practice, observation of a single satellite can only lead to satisfactory precision if the altitude of the observer is known. Of course, this description has been simplified. Other factors must be included, such as a possible frequency offset, refraction, the observer's own motion, and so on. However, the basic features of the system are clear enough. Measurements involving a single satellite give the 'fix', but not in an instantaneous way, since the satellite must be observed over a portion of its orbit. The whole set of directions is obtained all the more quickly as the satellite orbit is lower. Global coverage of the Earth by low-orbiting satellites thus requires polar orbits.

The pseudo-range method involves several synchronised spaceborne or Earth-based transmitters whose positions are known in the terrestrial frame. When two transmitters simultaneously emit signals, the observer measures the difference between reception times and thereby deduces the difference in the distances to these transmitters. The observer is then located on a known hyperboloid. With four transmitters, the position is instantaneously determined. Geometrical constraints require the transmitters to be spaceborne. However, for air and sea navigation, when the altitude is known, Earth-based two-dimensional positioning systems are used. In both cases, redundant information from extra transmitters provides added precision and safety. Apart from the position, a time comparison between the observer's clock and the time base of the system

is obtained as a kind of bonus. As we saw in Chapter 7, this by-product can be of great importance for the purposes of time dissemination.

The close relationship between the Doppler method and the pseudo-ranging method is worth noting, even though, in practice, this relationship may be concealed by the way data is treated mathematically. Such methods, applied with the help of artificial satellites, require a precise description of satellite orbits in the terrestrial frame. This is achieved by setting up a network of tracking stations at known positions. Measurements are immediately transmitted to a control centre where they are processed in the shortest possible time, so that orbital models sent back to the satellites are kept as up-to-date as possible, ready for transmission to users.*

We have already mentioned the Global Positioning System (GPS), based upon pseudo-range measurements, because it can be used to compare clocks, to disseminate time (Chapters 5 and 7) and to measure the Earth's rotation (Chapter 8). There has been an almost explosive increase in the number of users of GPS in its normal positioning mode. In the next section, we hope to satisfy the reader's curiosity by describing this system in a little more detail, although many specialised works are already devoted to the subject. In particular, reference [9.31] explains how it works and describes its various modes of use in geodesy. We shall then go on to outline some other systems.

9.2.2 The Global Positioning System

(a) The satellites and their messages

The Navstar system (Navigation Satellite Time and Ranging), better known today by the acronym GPS, stems from studies which began in 1965. Deployment was not completed until 1995, a development period that gives some idea of the magnitude of the undertaking. GPS results from a merger between US Navy and US Air Force projects (Timation and 621B, respectively).

The system uses twenty-four satellites in quasi-circular orbits at an altitude of about 20 000 km, distributed in six orbital planes at intervals of 60° longitude and inclined at 55° to the equator. The orbital period of these satellites is half the Earth's sidereal period. (The latter is 23 h 56 m and corresponds to an increase of 360° in the angle A_S defined in Section 8.1.2.) Hence, a fixed observer on the Earth sees the same satellite following the same path across the sky, in azimuth and elevation, every 23 h 56 m. At least four satellites are always visible from any point on the Earth's surface (apart from a few very shortlived exceptions) in a configuration suitable for obtaining the observer's position.

The first eleven satellites in the system, known as block I, carried rubidium clocks for the very first, then caesium clocks, as well as reserve quartz clocks as a security measure. Block II satellites, launched from 1985, are each equipped with two caesium clocks and two rubidium clocks, backed up by quartz clocks. They have a lifetime of about eight years.

The basic principle behind GPS is synchronism of on-board clocks. In practice, the clocks are not physically synchronised, but the satellites transmit data which can be used to calculate the offset between their readings and the time base of the system, viz., GPS Time, with uncertainties of a few nanoseconds.

GPS signals are transmitted at two carrier frequencies common to all satellites in the system. These are L1 at 1575.42 MHz and L2 at 1227.60 MHz. The information they transmit, called the message, is carried by 180° phase reversals in the carrier. It consists of the following components:

- a pseudo-random code C/A,
- a pseudo-random code P,
- a message carrying the data.

The pseudo-random codes are different for each satellite and mutually orthogonal. The autocorrelation function of each code is almost zero, except for zero delay. Correlation functions between satellite codes are almost zero. This explains how receivers distinguish the different satellites, generating the code for each one, and how they measure the reception time of the signal from each satellite (maximal correlation) using their internal clock.

The C/A code (*Coarse Acquisition* or *Clear Access*) is only transmitted at frequency L1, phase shifted by 90°. It repeats with period 1 ms and comprises a 1023 bit sequence.

The P code (*Precise* or *Protected*) is transmitted at both L1 and L2. Its period of repetition is 267 days, but one week segments are transmitted and changed every week. It can be encrypted to exclude unauthorised users.

The data message is transmitted at the low modulation speed of 50 bits s^{-1}. It is made up of *frames* of 1500 bits (lasting 30 s), each divided into five *subframes*. The content of subframes 1 to 3 repeats from frame to frame (unless this context is refreshed) and supplies information about the transmitting satellite, i.e., its health, orbital elements, and coefficients for calculating corrections to the clock currently in use. Subframes 4 and 5 provide information concerning the system as a whole, i.e., models for calculating ionospheric delays, offset between UTC and GPS Time, rough ephemerides and state of health for the other satellites in the system, and information about codes. Since all this information is too bulky to fit into two subframes, it is spread over subframes 4 and 5 of twenty-five consecutive frames. The whole set of information can thus only

be obtained by tracking the satellite for a total of 12.5 minutes. This explains why tracking extends over 13 minute periods when very precise time measurement applications are at stake (see Section 5.6.3). At each of the frequencies L1 and L2, the transmitted signal is the product of modulations corresponding to the code and the message. Note that during the transmission of one message bit, i.e., 20 ms, the C/A code repeats twenty times.

(b) Normal operation of the system

By 'normal operation', we understand simple real-time positioning applications using ordinary GPS receivers. We shall not discuss the many other modes of use for these receivers.

The determination of satellite orbits and calculation of clock corrections carried out at the master control station are based on tracking by five control stations. The geodetic coordinates of these stations are known with uncertainties as low as a few tens of centimetres relative to a reference system called WGS84. Every attempt is being made to bring the WGS84 system into line with the IERS frame known as ITRF, mentioned in Section 8.1.6. Deviations are now of the order of one decimetre. Orbital parameters are thus calculated and transmitted in the WGS84 system, in practice ITRF, and the same can therefore be said of the observer's coordinates.†

GPS Time is aligned on UTC (USNO) to within 100 ns, modulo 1 s. The latter follows UTC to within 100 ns.

Orbits are described by means of six parameters similar to the Keplerian parameters, viz., semi-major axis, eccentricity, inclination to the equator, right ascension of the ascending node, argument of the perigee, and mean anomaly. However, the orbit of an artificial satellite is perturbed by irregularities in the Earth's gravitational field, solar radiation pressure, the lunisolar tidal potential and the residual atmosphere. According to the approach usually adopted by astronomers, orbital parameters are thus regarded as variables. Using their values at a given instant, we obtain an orbit very close to the true orbit over a range of neighbouring times. Their rates of change are broadcast and parameters should be changed every hour or so.

Position uncertainties for a moving observer may be as low as about 10 m when the P code is used at both frequencies. With the C/A code, they are around 30 m. When *Selective Availability* (SA) was applied, until May 2000, uncertainties were about 100 m for unauthorised users.

† National cartography and marine charts have been established in regional systems, using traditional methods of geodesy. It is planned to transform them to ITRF, although this has not yet been achieved. Discrepancies with WGS84 and ITRF may reach several hundred metres. This must be taken into account and some receivers make the necessary corrections.

(c) Other applications

GPS can be used in differential mode to gain precision in a local positioning application. The position of the moving observer is referred to accurately known positions of nearby fixed stations simultaneously tracking the same satellites. Some errors are significantly reduced, including errors due to the ephemerides, satellite clock noise, refraction and SA when it was imposed. Uncertainties can be reduced to a few metres in real time. Differential positioning is also used to determine relative positions of timing receivers and thereby improve clock comparisons. Since these receivers are fixed, time averages can be taken and uncertainties in relative coordinates are of decimetre order [9.32].

For high precision applications, such as centimetre geodesy and study of the Earth's rotation, the phase of the carrier waves is followed. Modulation is removed by multiplying the received signal by itself. Since modulation is achieved by phase reversal, a pure double frequency is obtained. In this field, studies are coordinated by an international service known as the *International GPS Service for Geodynamics* (IGS). This service establishes precise ephemerides for GPS satellites, available one or two weeks after the event. They are useful, in particular, for reducing uncertainties in time comparisons when producing TAI.

(d) Civilian complement to GPS

As far as civilian applications are concerned, GPS has been criticised for its lack of precision due to SA and also for the absence of real time warnings when one of its satellites fails. Since the number of satellites visible only just exceeds the minimum required number of four, such failure may lead to incorrect positioning with disastrous consequences. The question then arises as to whether a civilian system should be set up in parallel.

The complexity of GPS is partly due to its military role. A system operating along the same lines has been designed without satellite-borne clocks. The time provided by a single master clock on the ground is transmitted to satellite transponders. However, this kind of system does not satisfy the need of the armed forces for a completely self-sufficient mode of operation, i.e., one in which no data need be injected by ground stations. Such autonomy, although its degrades quality, may be vital in time of conflict.

For the moment, two lines of development are under investigation. The first involves combined use of GPS and its Russian counterpart GLONASS, to be discussed shortly. The number of satellites available may then rise to forty-eight, with a dozen or so simultaneously in view. The resulting redundancy means that an offending satellite can be detected and eliminated. Mixed

GPS/GLONASS receivers are already available. The other development involves setting up a civilian complement to GPS, in which transmitters would be carried by geostationary satellites and transmit with GPS codes [9.33].

9.2.3 Other positioning systems

GLONASS was developed by the USSR, and subsequently by Russia, in parallel with GPS. It operates according to the same principles, although the satellites follow slightly different orbits at an altitude of about 20 000 km. A fixed observer also sees a satellite move along the same apparent trajectory every 23 h 56 m, but this satellite is a different one each day, with a one week cycle. Satellites are recognised by a range of frequencies. Orbital models are transmitted in the form of coordinates. In addition, the time base known as GLONASS Time is different, and not so closely synchronised with UTC. The geodetic frame is also different. This makes the construction of mixed GPS/GLONASS receivers a complicated task, although not impossible.

Before GPS and GLONASS came on the scene, the Doppler system Navsat was widely used for maritime navigation from 1964. The six satellites in polar orbits at altitude 1100 km transmitted frequencies 150 MHz and 400 MHz, and also synchronisation markers. Messages providing orbital models were transmitted by phase modulation.

Another Doppler system known as DORIS was designed by the French Centre national d'études spatiales (CNES) for high precision positioning [9.34]. It uses the opposite approach to Navsat in the sense that frequencies are transmitted by the beacons to be localised and received by satellites. Both beacons and satellites are equipped with ultra-stable quartz oscillators. Frequencies are close to 400 MHz and 2 GHz. The frequency instability of beacon and satellite oscillators has to be less than 5×10^{-13} over the counting period (the interval Δt defined in Section 9.2.1) and 10^{-12} throughout the pass (e.g., for a period of 1000 s). The on-board frequency is determined by comparing with a ground-based master clock adjusted to TAI. The uncertainty must be less than 10 μs. A basic network of forty to fifty fixed beacons on Earth serves to determine satellite orbits. The system was tried out on the Earth observation satellite Spot 2. It was subsequently adopted as one of the positioning systems for the TOPEX/POSEIDON oceanographic satellite, to be discussed below. In the latter case, it was shown that the uncertainty in the radial component of the satellite position is around 3 cm. For the two other components, the uncertainty is about 20 cm. DORIS also provides the beacon positions, with uncertainty as low as 2 cm. This precision allowed DORIS to feature amongst the panoply of

techniques used at the IERS for constructing ITRF and measuring the Earth's rotation.

The PRARE system (*Precise Range and Range Rate Equipment*) was originally designed in Germany for precise determination of artificial satellite orbits [9.35]. It also includes a satellite-borne space segment, a network of tracking stations and a processing centre. It combines pseudo-ranging and Doppler measurements on signals transmitted by the satellite at frequencies 2 GHz and 8 GHz. This system is used by the ESA's Earth Reconnaissance Satellites (ERS) in conjunction with GPS. It has the same order of precision as DORIS.

9.3 Very Long Baseline Interferometry (VLBI)

In Section 8.1.4, we discussed the part played by Very Long Baseline Interferometry (VLBI) in the measurement of Universal Time. In its applications to the study of the Earth's rotation and the construction of celestial and terrestrial reference frames, angular separations between sources should be as large and varied as possible, in order to separate the various quantities to be determined. We have already seen that this type of measurement takes advantage of the frequency stability of hydrogen masers.

VLBI was originally devised to study the structure of celestial sources with small angular dimensions emitting in the radio region of the spectrum. Resolution is improved when reception frequency increases and antennas are more widely spaced. At centimetre and decimetre wavelengths and using Earth-based antennas, angular resolutions of $0.001''$ are achieved. Reception in the millimetre wavelengths, still under development, reaches resolutions of $0.0001''$. VLBI projects involving one Earth-based and one spaceborne antenna would obtain resolutions down to a few tens of microseconds of arc [9.36].

VLBI is also used in differential mode to locate sources on the sky relative to the directions of compact extragalactic sources forming the primary reference frame. In this case nearby sources are the target and typical angular separations to be measured are of a few degrees.

This technique can be used in different ways to extend the celestial reference frame ICRF to the optical domain (see Section 3.3.2b). In particular, the space astrometry mission HIPPARCOS provided a catalogue of angular positions and proper motions of some 100 000 stars with an internal consistency at the level of $0.001''$ and $0.001''$/yr, respectively, but was unable to give an initial setting for the rotation and rotation rate of the whole catalogue [9.37]. The HIPPARCOS catalogue had to be aligned with the ICRF. One technique used for this purpose was VLBI observation of a few dozen radio stars belonging to the HIPPARCOS catalogue. By combining the various methods, it is estimated

that the fit is achieved to within 0.0005″ at the mean date of the HIPPARCOS observations (1991). However, it deteriorates at a rate of about 0.000 25″ per year through uncertainties in proper motions of the stars [9.38].

Differential VLBI is often used to track space probes, in order to situate them relative to ICRF, and also to locate artificial satellites placed in orbit around other planets.

9.4 The TOPEX/POSEIDON mission

As an example, we shall discuss the many time and frequency aspects of the TOPEX/POSEIDON mission.

Measurements of the vertical distance between an artificial satellite and the surface of the sea by means of an on-board radar altimeter were first attempted between 1975 and 1978 with the American Geos 3 satellite. They were continued using Seasat in 1978. Although the satellite failed after a few months, it had already proven that these measurements were useful for studying the oceans and their interaction with the atmosphere, climatology and internal geophysics. Geosat confirmed this appraisal over 1985–1989. More ambitious programmes were subsequently devised, with the European Space Agency's ERS and the TOPEX/POSEIDON collaboration between NASA and CNES (France). The satellite ERS 1 was launched in June 1991 and followed by ERS 2 in 1995. TOPEX/POSEIDON was launched on 10 August 1992.

The radar beam intersects a sufficiently large region of the sea surface to average out the effect of the waves. Altimetric measurements involve the various elements shown in Figure 9.1.

- The *reference ellipsoid* is a mathematical surface adopted by convention which approximately follows the sea surface. This removes the need to refer to the geocentre.
- The *geoid* is a surface of constant gravitational potential, after correcting for periodic terms introduced by the tidal potential. It would coincide with the mean sea surface if the waters of the oceans were homogeneous and at rest.
- The *real sea surface* is defined as the surface when wave effects have been averaged out.
- The *satellite orbit* is established in a geocentric frame. Its altitude is calculated above the reference ellipsoid.

Altimetric measurement gives the altitude of the satellite above the real sea surface. Knowing the satellite orbit, we then obtain the *height of the sea surface* above the reference ellipsoid.

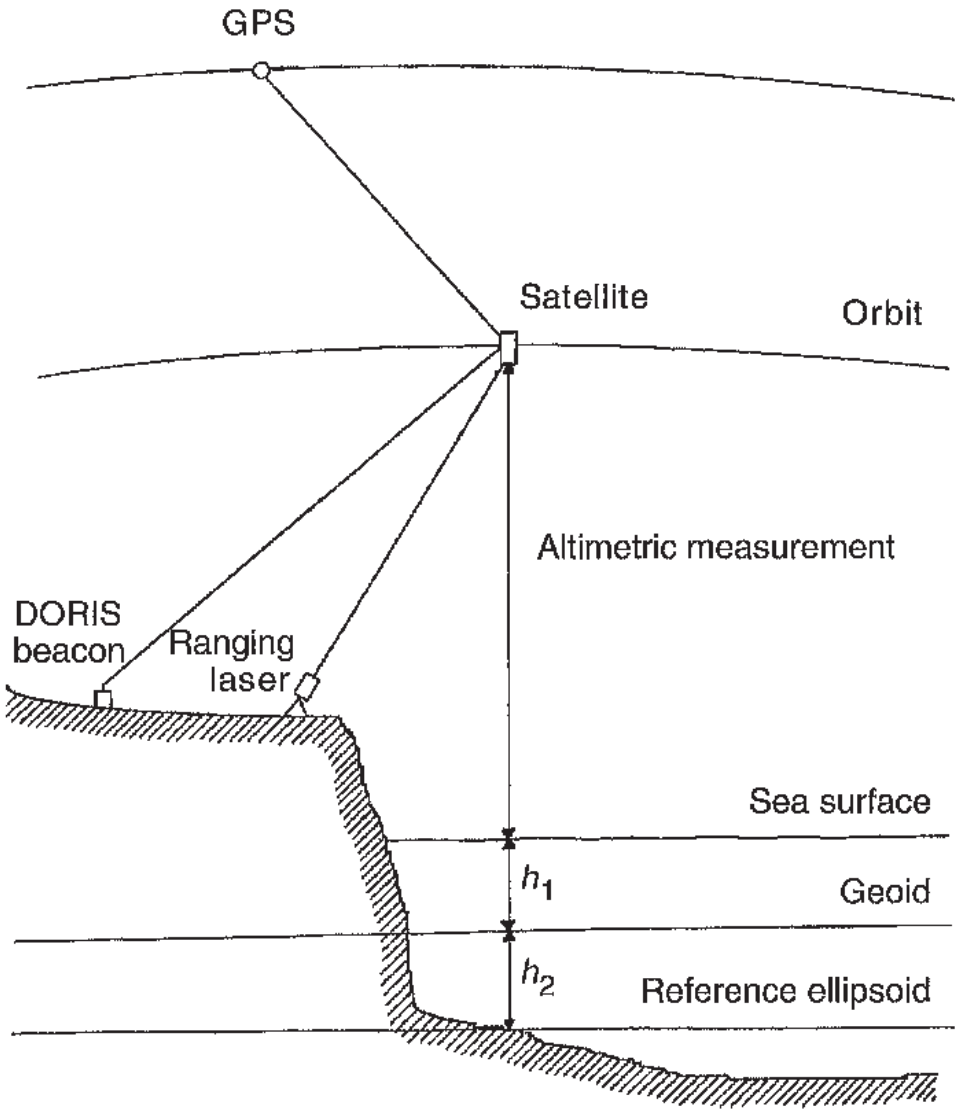


Fig. 9.1. Geometrical elements of altimetry above the oceans.

Two elements of great scientific importance are the height h_2 of the geoid above the reference ellipsoid, which may be positive or negative, and attains values of a hundred metres or so, and the height h_1 of the sea surface above the geoid, which may also be positive or negative, and attains values of around one metre. We note immediately that the measurement yields the sum of these two elements. This ambiguity is removed by their variability. Over several years, the geoid can be taken as invariable and the oceanic variation is obtained. If we assume that oceanic effects have zero time average, then the mean surface represents the spatial irregularities of the geoid. In the very long term, however, difficulties may remain here. For example, the melting of the polar ice caps affects both sea levels and the shape of the geoid.

The form of the radar echo is affected by conditions at sea. By analysing it, information is obtained about the waves, including their height and direction, quite apart from geometrical data concerning altitudes.

Sea levels are affected by temperature differences, salinity, the atmosphere and dynamical effects such as those due to centrifugal and Coriolis forces. It is easy to see therefore that their study can reveal currents, eddies, and heat energy transfers generating climatic change. However, such effects amount to no more than a few decimetres. In order to study them, they must be measured with uncertainties at the centimetre level. It is interesting to list the scientific themes mentioned in the call for research proposals issued by the CNES when TOPEX/POSEIDON was still under preparation [9.39]:

- to conduct studies of the permanent and variable circulation of the ocean, and its interaction with the atmosphere, on global and basin scales, in order to contribute to our understanding of the climatic role played by the oceans;
- to develop an improved understanding of oceanic circulation on regional scales over periods of a few months;
- to obtain an improved understanding of heat transfer in the ocean;
- to conduct studies of tidal models;
- to study the statistics of ocean variability;
- to improve knowledge of the gravity field and associated geoid;
- to conduct studies of geophysics;
- to study the physics of waves.

The TOPEX/POSEIDON satellite is in circular orbit at an altitude of about 1300 km, inclined at 66° to the equator. It does not therefore cover the polar regions. The track of its orbit, that is, the vertical projection of the orbit on the Earth's surface, repeats every 10 days and the distance between tracks is 316 km at the equator, decreasing at higher latitudes. The mission, initially planned until 1997, was still running in 2000. Geophysical assessment of the mission and scientific results clearly show how successful it has been. This is attested by the 1200 pages devoted to the subject in Vols. 99 and 100 of the *Journal of Geophysical Research* for 1994 and 1995, and many other scientific papers published since then. In particular, uncertainty in the height of the sea surface above the ellipsoid is around 5 cm for a single pass. This includes uncertainties in the orbit and the altimetric measurement in almost equal parts. By taking a monthly average, the uncertainty can be reduced to 1 cm. It has also been shown that the variation in the mean sea level can be measured to within 1 mm yr^{-1}. This achievement, combined with the success of ERS 1 and 2, has encouraged projects to launch further altimetric satellites. (ERS 1 and 2 are complementary because they have different tracks and repeat in different ways.) Such projects would study long term effects, including long term variations in sea level, as well as providing continuous monitoring (Jason project).

Leaving aside the scientific results, which are not our main concern here, we shall now discuss the role played by time and frequency in the TOPEX/POSEIDON mission.

The altimeter is nothing other than a radar, measuring the time of flight of a signal. The inaccuracy of the altimeter clock must be less than 10^{-10} if its contribution is to be negligible. Its frequency instability over the time of flight, roughly 10 ms, must also be less than this same value. Measurement dates must be expressed in TAI to the nearest 10 μs.

Time measurements are critical in determining the satellite orbit. The precision required for TOPEX/POSEIDON (an uncertainty of a few centimetres in the radial distance) went far beyond anything that had previously been achieved. Three main tracking systems were used, viz., laser ranging via retroreflectors on the satellite, an on-board GPS receiver, and the DORIS system. These time/frequency systems have been discussed in Section 9.2. Knowledge of the Earth's gravitational field is required to model the orbit. The model is based on trackings of a great many satellites in very different orbits. Such trackings also involve time/frequency positioning techniques. Moreover, in the present case, the model was improved by tracking TOPEX/POSEIDON itself. Such analyses require precise knowledge of Earth rotation parameters, because the orbit is described in a non-rotating system, whereas the satellite position must be given in the reference system that accompanies the Earth in its rotation. This rotation is now monitored solely by means of time/frequency techniques (see Section 8.1.4). The rotation parameters can be determined when DORIS tracks TOPEX/POSEIDON, but it is useful to combine the results with those stemming from other techniques. Among the latter, VLBI requires the most stable clocks that can be built.

Finally, TAI is involved in various ways. It is the base for synchronisation (to within 1 μs in tracking stations), the reference for frequency accuracy, and the time argument in dynamical theories used for orbital calculations.

9.5 Telecommunications

Telecommunications networks have a tree structure in which the ends of the branches are user terminals. Several hierarchical levels contain nodes at which messages are switched from one channel to another. In the analogue transmission mode, several messages are sent together through the same communication channel using single sideband amplitude modulation without transmission of the carrier. In order to correctly reconstitute the message, the frequencies of oscillators equipping modulation/demodulation centres must be tuned to the nearest 10^{-8} [9.40].

Digital voice transmission represents a significant development. Voice frequencies, between about 300 Hz and 3600 Hz, are sampled at a rate of 8000 times per second and coded in 8 bit words. It is then possible to blend messages together and transmit them simultaneously through the same channel, up to thirty at a time, in a process called multiplexing. In order to demultiplex, a frequency reference is needed at reception. However, a quite substantial shift can be tolerated between multiplexing and demultiplexing frequencies, since a time shift is not audible.

For the transmission of numerical data in which information loss due to time shifts cannot be tolerated, a need arises for good syntonisation, i.e., good frequency tuning, over the whole network. This problem has led to the current development of the *Synchronous Digital Hierarchy* (SDH). Here, frequencies at nodes and terminals must be tuned to within 10^{-11} on a daily average. In addition, the International Telecommunication Union stipulates that this frequency should be fixed relative to UTC frequency, and hence TAI frequency, to within 10^{-11} in normalised value [9.41].

Within the context of the present work, the needs of telecommunications may appear modest. More sophisticated solutions are ruled out by the quest for lower costs. Nevertheless, caesium clocks are a necessity at the master node of the network. Nodes at other levels and user terminals are equipped with quartz oscillators servo controlled by the network itself. GPS can contribute to the control of frequencies at the master node or at high level nodes.

Despite its name, SDH is based upon syntonisation. The use of synchronisation in the proper sense has sometimes been advocated for the advantages it would bring. There has even been some talk of a future need for real-time access to UTC to within 100 ns for telecommunications purposes [9.42]. This is no doubt a project for the rather distant future, about which we have been unable to obtain further details.

Appendix 1

Acronyms for time laboratories

The following is a list of acronyms for laboratories k maintaining an approximation UTC(k) to UTC, or an independent local atomic time scale TA(k), together with their name and location.

AMC	Alternate Master Clock station, Colorado Springs, CO, USA
AOS	Astronomiczne Obserwatorium Szerokosciowe, Borowiec, Poland
APL	Applied Physics Laboratory, Laurel, MA, USA
AUS	Consortium of Australian laboratories
BEV	Bundesamt für Eich- und Vermessungswesen, Vienna, Austria
BIRM	Beijing Institute of Radio Metrology and Measurement, Beijing, People's Republic of China
CAO	Cagliari Astronomical Observatory, Cagliari, Italy
CH	Consortium of Swiss laboratories
CNM	Centro Nacional de Metrologia, Mexico City, Mexico
CRL	Communications Research Laboratory, Tokyo, Japan
CSAO	Shaanxi Astronomical Observatory, Lintong, People's Republic of China
CSIR	Council for Scientific and Industrial Research, Pretoria, South Africa
DLR	Deutsche Forschungsanstalt für Luft- und Raumfahrt, Oberpfaffenhofen, Germany
DTAG	Deutsche Telekom AG, Darmstadt, Germany
F	Commission Nationale de l'Heure, Paris, France
GUM	Glówny Urzad Miar, Central Office of Measures, Warsaw, Poland

IEN	Istituto Elettrotecnico Nazionale Galileo Ferraris, Turin, Italy
IFAG	Bundesamt für Kartographie und Geodaesie Fundamental-station, Wettzell, Germany
IGMA	Instituto Geografico Militar, Buenos Aires, Argentina
INPL	National Physical Laboratory, Jerusalem, Israel
IPQ	Institute Portugues da Qualidade, Monte de Caparica, Portugal
JATC	Joint Atomic Time Commission, Lintong, People's Republic of China
KRIS	Korea Research Institute of Standards and Science, Taejon, Republic of Korea
LDS	The University of Leeds, Leeds, United Kingdom
MSL	Measurement Standards Laboratory, Lower Hutt, New Zealand
NAO	National Astronomical Observatory, Misuzawa, Japan
NIM	National Institute of Metrology, Beijing, People's Republic of China
NIST	National Institute of Standards and Technology, Boulder, CO, USA
NPL	National Physical Laboratory, Teddington, United Kingdom
NPLI	National Physical Laboratory, New Delhi, India
NRC	National Research Council, Ottawa, Canada
NRLM	National Research Laboratory of Metrology, Tsukuba, Japan
OMH	Orszagos Mérésügyi Hivatal, Budapest, Hungary
ONBA	Observatorio Naval, Buenos Aires, Argentina
ONRJ	Observatorio Nacional, Rio de Janeiro, Brazil
OP	Observatoire de Paris, Paris, France
ORB	Observatoire Royal de Belgique, Brussels, Belgium
PTB	Physikalisch-Technische Bundesanstalt, Braunschweig, Germany
ROA	Real Instituto y Observatorio de la Armada, San Fernando, Spain
SCL	Standards and Calibration Laboratory, Hong Kong, People's Republic of China
SO	Shanghai Observatory, Shanghai, People's Republic of China
SP	Swedish National Testing and Research Institute, Boras, Sweden
SU	Institute of Metrology for Time and Space (IMPV), NPO 'VNIIFTRI', Mendeleevo, Moscow Region, Russia
TL	Telecommunication Laboratories, Chung-Li, Taiwan

TP	Institute of Radio Engineering and Electronics, Academy of Sciences, Czech Republic
UME	Ulusai Metroloji Enstitüsü, Marmara Research Centre, National Metrology Institute, Gebze-Kocaeli, Turkey
USNO	US Naval Observatory, Washington, DC, USA
VSL	Van Swiden Laboratorium, Delft, Netherlands

Appendix 2
Abbreviations

Time scales

Abbreviations in bold type are official in all languages.

EAL	Free (unsteered) atomic time scale
ET	Ephemeris Time
GMST	Greenwich Mean Sidereal Time
GMT	Greenwich Mean Time
TA(*k*)	Atomic time established by laboratory *k*
TAI	International Atomic Time
TCB	Barycentric Coordinate Time
TCG	Geocentric Coordinate Time
TDB	Barycentric Dynamical Time
TDT	Terrestrial Dynamical Time (became **TT** in 1991)
TT	Terrestrial Time
UT	Universal Time
UTC	Coordinated Universal Time
UT1	Universal Time 1 (forms 0 and 2 also exist)

Miscellaneous

BIH	Bureau international de l'heure
BIPM	Bureau international des poids et mesures (International Bureau of Weights and Measures)
CCDS	Consultative Committee for the Definition of the Second (became CCTF in 1997)
CCIR	International Radio Consultative Committee
CCTF	Consultative Committee for Time and Frequency

CGPM	Conférence générale des poids et mesures (General Conference on Weights and Measures)
CIPM	Comité international des poids et mesures (International Committee for Weights and Measures)
CNES	Centre national d'études spatiales, France
GLONASS	Global Navigation Satellite System
GPS	Global Positioning System
IAU	International Astronomical Union
ICRF	International Celestial Reference Frame
IERS	International Earth Rotation Service
ITRF	International Terrestrial Reference Frame
ITU	International Telecommunication Union
IUGG	International Union of Geodesy and Geophysics
JD	Julian Date
LHA	Laboratoire de l'horloge atomique, Orsay, France
LPTF	Laboratoire primaire du temps et des fréquences, Paris
MJD	Modified Julian Date
NASA	National Aeronautics and Space Administration
SI	International System of Units
TWSTFT	Two-Way Satellite Time and Frequency Transfer
URSI	International Union of Radio Science
VLBI	Very Long Baseline Interferometry
VLF	Very Low Frequency
WGS	World Geodetic System

Appendix 3

Definitions of base units in the SI system

The metre [m]

The metre is the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second. (17th General Conference on Weights and Measures, 1983.)

The kilogram [kg]

The General Conference sanctions the prototype of the kilogram adopted by the International Committee. This prototype shall henceforth be considered to be the unit of mass. (1st CGPM, 1889.)

The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram. (3rd CGPM, 1901.)

The second [s]

The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atom. (13th CGPM, 1967.)

The ampere [A]

The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 metre apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per metre of length. (9th CGPM, 1948.)

The kelvin [K]

The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water. (13th CGPM, 1967.)

The mole [mol]

(1) The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12; its symbol is 'mol'.
(2) When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles. (14th CGPM, 1971.)

The candela [cd]

The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of 1/683 watt per steradian. (16th CGPM, 1979.)

Appendix 4

International services

Time

Bureau international des poids et mesures, Time Section,
Pavillon de Breteuil, F-92312 Sèvres Cedex (France).

Telephone	BIPM Switchboard	+33 1 45 07 70 70
Fax	BIPM General	+33 1 45 34 20 21
	BIPM Time Section	+33 1 45 07 70 59
Internet	`http://www.bipm.fr`	
	or anonymous ftp	62.161.69.5 (subdirectory TAI)
E-mail	`tai@bipm.fr`	

Rotation of the Earth (Universal Time)

International Earth Rotation Service (IERS).

Central Bureau

Bundesamt fuer Kartographie und Geodaesie (BKG),
Richard Strauss Allee 11, D-60598 Frankfurt-am-Main (Germany).

Telephone	+49 69 6333 273
Fax	+49 69 6333 425
E-mail	`richter@ifag.de`

Product Centre on Earth Orientation

Observatoire de Paris, Département d'astronomie fondamentale,
61, avenue de l'Observatoire, 75014 Paris (France).

Internet	`http://hpiers.obspm.fr`	
	or anonymous ftp `hpiers.obspm.fr`	145.238.100.28
E-mail	`gambis@obspm.fr`	

Rapid Service and Predictions of Earth Rotation Parameters

US Naval Observatory, Earth Orientation Department,
3450 Massachusetts Avenue NW, Washington DC 20392-5420, USA.

Internet	`http://maia.usno.navy.mil`	
	or anonymous ftp `maia.usno.navy.mil`	192.5.41.22
E-mail	`jimr@maia.usno.navy.mil`	

References

Chapter 2

[2.1] H. Poincaré (1906). *La valeur de la science* (Flammarion, Paris).

Chapter 3

[3.1] C.M. Will (1993). *Theory and experiment in gravitational physics* (revised edn.) (Cambridge University Press, Cambridge).

[3.2] J.L. Anderson (1967). *Principles of relativity physics* (Academic Press, New York).

[3.3] B. Guinot (1997). Application of general relativity to metrology, *Metrologia* **34**, 261–290.

[3.4] C.W. Misner, K.S. Thorne, J.A. Wheeler (1973). *Gravitation* (Freeman, New York, San Francisco, London).

[3.5] International Astronomical Union (1991). *IAU transactions*, Vol. XXI B, ed. J. Bergeron (Kluwer, Dordrecht, Boston, London).

[3.6] V.A. Brumberg (1991). *Essential relativistic celestial mechanics* (Adam Hilger, Bristol, Philadelphia, New York).

[3.7] International Earth Rotation Service (1996). *Conventions*, Technical note 21, ed. D.D. McCarthy (Observatoire de Paris).

[3.8] S.A. Klioner (1992). The problem of clock synchronisation: a relativistic approach, *Celestial mechanics and dynamical astronomy* **83**, 81–109.

[3.9] P. Wolf, G. Petit (1996). Relativistic theory for clock syntonization and the realization of geocentric coordinate times, *Astron. Astrophys.* **304**, 653–661.

[3.10] P. Willis (1991). IAG Newsletter, *Bulletin géodésique* **65**, 189–208.

Chapter 4

[4.1] G. Bigourdan (1914). Le jour et ses divisions, les fuseaux horaires et l'Association internationale de l'heure, *Annuaire du Bureau des longitudes, notice B.*

[4.2] October 1884. International conference held in Washington for the adoption of a single prime meridian and a universal time, *Procès-verbaux des Séances* (Gibson Bros., Printers and Bookbinders, Washington).

[4.3] J. Renaud (1917). L'avance de l'heure légale pendant l'été 1916, *Annuaire du Bureau des longitudes, notice B.*

[4.4] A. Lambert (1940). Le Bureau international de l'heure, son rôle, son fonctionnement, *Annuaire du Bureau des longitudes.*

[4.5] B. Guinot (1992). Le Bureau international de l'heure de 1911 à 1964: Le temps astronomique et la naissance du temps atomique, in *Le Bureau international de l'heure, 75 ans au service de l'heure universelle*, 31–43 (Observatoire de Paris, Bureau des longitudes).

[4.6] P. Pâquet (1992). Le BIH de 1964 à 1987. Temps atomique et géodynamique globale, évolution et réorganisation, in *Le Bureau international de l'heure, 75 ans au service de l'heure universelle*, 45–57 (Observatoire de Paris, Bureau des longitudes).

[4.7] D. Howse (1980). *Greenwich time and the discovery of the longitude* (Oxford University Press, Oxford, New York, Toronto, Melbourne).

[4.8] W. de Sitter (1927). On the secular acceleration and fluctuation of the longitude of the Moon, Sun, Mercury and Venus, *Bull. Astron. Inst. Neth.* **iv**, 21–38.

[4.9] A. Danjon (1929). Le temps, sa définition pratique, sa mesure, *L'astronomie* **43**, 13–22.

[4.10] G.M. Clemence (1950). On revising the official system of astronomical constants, *Constantes fondamentales de l'astronomie*, 9–28 (CNRS, Paris).

[4.11] A. Scheibe, U. Adelsberger (1936). Nachweiss von Schwankungen der astronomischer Tageslänge mittels Quartzuhren, *Phys. Zeitschrift* **37**, 38.

[4.12] N. Stoyko (1937). Sur la périodicité dans l'irrégularité de la rotation de la Terre, *C. R. Acad. Sci. Paris* **250**, 79.

[4.13] E.W. Woolard, G.M. Clemence (1966). *Spherical astronomy* (Academic Press, New York, London).

[4.14] P. Forman (1985). Atomichron: The atomic clock from concept to commercial product, *Proc. IEEE* **73**, 1181–1204.

[4.15] R.E. Beehler (1967). A historical review of atomic frequency standards, *Proc. IEEE* **55**, 792–805.

[4.16] R.L. Sydnor (1982). Distinguished scientist panel discussion, *Proc. 14th Annual Precise Time and Time Interval Applications and Planning Meeting*, 631–657 (Greenbelt, USA).

[4.17] N.F. Ramsey (1983). History of atomic clocks, *J. of Res. National Bureau of Standards* **88**, 301–320.

[4.18] L. Essen, J.V.L. Parry (1957). The caesium resonator as a standard of frequency and time, *Phil. Trans. R. Soc. London Ser. A* **250**, 45–69.

[4.19] P. Forman (1982). Atomic clocks. Preview of an exhibit at the Smithsonian, *Proc. 36th Ann. Freq. Contr. Symp.*, 220–222 (Philadelphia, USA).

[4.20] W. Markowitz, R.G. Hall, L. Essen, J.V.L. Parry (1958). Frequency of cesium in terms of ephemeris time, *Phys. Rev. Lett.* **1**, 105–107.

[4.21] B. Guinot (1967). Formation d'une échelle moyenne de temps atomique, *Bull. Astron., Série 3* **2**, 449–464.

[4.22] J.McA. Steele, W. Markowitz, C.A. Lidback (1964). Telstar time synchronization, *IEEE Trans. Instrum. Meas.*, Vol. IM-**13**, No. 4, 164–170.

[4.23] G. Becker (1967). La nouvelle définition de la seconde et les théories relativistes, *Com. Cons. Définition de la Seconde, 4e session*, S27–S28.

[4.24] B. Guinot (1997). Application of general relativity to metrology, *Metrologia* **34**, 261–290.

[4.25] D.H. Sadler (1978). Mean solar time on the meridian of Greenwich, *Q. J. R. Astron. Soc.* **19**, 290–309.

Chapter 5

[5.1] ISO (1993). *International Vocabulary of Basic and General Terms in Metrology*, 2nd edn. (International Organization for Standardization, Geneva).

[5.2] IEEE (1988). *IEEE Standard definitions of physical quantities for fundamental frequency and time metrology* (Institute of Electrical and Electronic Engineers, Standard no. 1139-1988, New York).

[5.3] CCIR (1992). Measurements of frequency and time (phase), in *International Telecommunications Union: Recommendation 538-2 of the International Radio Consultative Committee* (CCIR, Geneva).

[5.4] IEEE (1984). *Frequency stability: fundamentals and measurement*, ed. V.F. Kroupa (IEEE Press, New York).

[5.5] NIST (1990). *Characterization of clocks and oscillators*, eds. D.B. Sullivan, D.W. Allan, D.A. Howe and F.L. Walls (NIST Technical Note 1337).

[5.6] P. Kartaschoff (1978). *Frequency and time* (Academic Press, London, New York, San Francisco).

[5.7] J.A. Barnes, A.R. Chi, L.S. Cutler, D.J. Healey, D.B. Leeson, T.E. McGunigal, J.A. Mullen Jr., W.L. Smith, R.L. Sydnor, R.F.C. Vessot, G.M.R. Winkler (1971). Characterization of frequency stability, *IEEE Trans. Instrum. Meas.* **20**, 105–120.

[5.8] P. Lesage, C. Audoin (1979). Characterization and measurement of time and frequency stability, *Radio Sci.* **14**, 521–539.

[5.9] D.W. Allan (1987). Time and frequency (time domain) characterization, estimation, and prediction of precision clocks and oscillators, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **34**, 647–654.

[5.10] D.A. Howe, D.W. Allan, J.A. Barnes (1981). Properties of signal sources and measurement methods, *Proc. 35th Annual Frequency Control Symposium*, A1–A47 (Philadelphia, USA).

[5.11] C.A. Greenhall (1991). Recipes for degrees of freedom of frequency stability estimators, *IEEE Trans. Instrum. Meas.* **40**, 994–999.

[5.12] L.-G. Bernier (1988). Linear prediction of the non-stationary clock error function, *Proc. 2nd European Frequency and Time Forum*, 125–137 (Neuchâtel, Switzerland).

[5.13] D.W. Allan (1991). Time and frequency metrology: current status and future considerations, *Proc. 5th European Frequency and Time Forum*, 1–9 (Besançon, France).

[5.14] F. Boquet (1913). Sur l'organisation du service de l'heure à l'Observatoire de Paris, *Bull. Soc. Astronomique de France*, June issue.

[5.15] E.A. Gerber, A. Ballato (eds.) (1985). *Precision frequency control*, 2 volumes (Academic Press, Orlando).

[5.16] F.L. Walls, J.-J. Gagnepain (1992). Environmental sensitivities of quartz crystal oscillators, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **39**, 241–249.

[5.17] R.-J. Besson (1977). A new electrodeless resonator design, *Proc. 31st Annual Frequency Control Symposium*, 147–152 (Atlantic City, USA).

[5.18] B. Guinot (1997). Application of general relativity to metrology, *Metrologia* **34**, 261–290.

[5.19] G. Petit, P. Wolf (1994). Relativistic theory for picosecond time transfer in the vicinity of the Earth, *Astron. Astrophys.* **286**, 971–977.

[5.20] P. Wolf, G. Petit (1995). Relativistic theory for clock syntonization and the realization of geocentric coordinate times, *Astron. Astrophys.* **304**, 653–661.

[5.21] P. Parcelier, G. Freon (1976). Comparaison d'horloges atomiques par réception de signaux de télévision, *Bull. Information Bur. Nat. de Métrologie* **23**, 32–39.

[5.22] D.W. Allan, C. Thomas (1994). Technical directives for standardization of GPS time receiver software, *Metrologia* **31**, 69–79.
[5.23] M. Brunet (1979). Synchronization of atomic clocks through the SYMPHONIE satellite, *Radio Sci.* **14**, 721–730.
[5.24] B.S. Dudnik, B.L. Kashcheev, Y.A. Koval, S.F. Semenov, M.D. Sopelnikov, A.A. Tkachuk (1986). New equipment for comparison of time and frequency standards using meteorite trails, *Izm. Teknika* **4**.
[5.25] P. Friedelance (1994). L'expérience LASSO, Doctoral thesis, University of Paris VI.

Chapter 6

[6.1] P. Kartaschoff (1978). *Frequency and time* (Academic Press, London).
[6.2] H. Hellwig (1985). Microwave frequency and time standards, in *Precision frequency control*, 113–176, eds. E.A. Gerber, A. Ballato (Academic Press, Orlando).
[6.3] C. Audoin (1991). Les étalons atomiques de fréquence, in *La mesure de la fréquence des oscillateurs*, 217–276 (Chronos, Masson, Paris).
[6.4] J. Vanier, C. Audoin (1989). *The quantum physics of atomic frequency standards*, 2 vols. (IOP Publishing, Bristol and Philadelphia).
[6.5] B. Cagnac, J.C. Pebay-Payroula (1983). *Physique atomique*, 2 vols. (Dunod Université, Paris).
[6.6] N.F. Ramsey (1956). *Molecular beams* (Oxford University Press, London).
[6.7] A. Kastler (1950). Quelques suggestions concernant la production optique et la détection optique d'une inégalité de population des niveaux de quantification spatiale des atomes. Applications à l'expérience de Stern et Gerlach et à la résonance magnétique, *J. Phys. Radium* **11**, 255–265.
[6.8] R.H. Dicke (1953). The effect of collisions upon the Doppler width of spectral lines, *Phys. Rev.* **89**, 472–473.
[6.9] A. de Marchi (1987). Understanding environmental sensitivity and ageing of cesium beam frequency standards, *Proc. 1st European Frequency and Time Forum*, 288–293 (Besançon, France).
[6.10] L. Essen, J.V.L. Parry (1957). The caesium resonator as a standard of frequency and time, *Phil. Trans. R. Soc. London Ser. A* **250**, 45–69.
[6.11] A. Bauch, K. Dorenwendt, B. Fisher, T. Heindorff, E.K. Müller, R. Schröder (1987). CS2: The PTB's new primary clock, *IEEE Trans. Instrum. Meas.* **36**, 613–616.
[6.12] K. Dorenwendt, B. Fisher, T. Heindorff (1990). The PTB's primary time and frequency standards – performance and uncertainty, *Physica* **41**, 712–716.
[6.13] A. Bauch, T. Heindorff, R. Schröder, B. Fisher (1996). The PTB primary clock CS3: Type B evaluation of its standard uncertainty, *Metrologia* **33**, 249–259.
[6.14] A. Bauch, B. Fisher, T. Heindorff, R. Schröder (1998). Performance of the PTB reconstructed primary clock and an estimate of its current uncertainty, *Metrologia* **35**, 829–845.
[6.15] M. Arditi, J.L. Picqué (1980). A cesium beam atomic clock using laser optical pumping. Preliminary tests, *J. Phys. Lett.* **41**, L379–L381.
[6.16] L.L. Lewis, M. Feldman (1981). Optical pumping by lasers in atomic frequency standards, *Proc. Ann. Freq. Control Symposium* 612–624 (Philadelphia, USA).
[6.17] P. Cérez, G. Théobald, V. Giordano, N. Dimarcq, M. de Labachelerie (1991).

Laser diode optically pumped cesium beam frequency standards investigations at L.H.A., *IEEE Trans. Instrum. Meas.* **40**, 137–141.

[6.18] A. Makdissi, J.-P. Berthet, E. de Clercq (2000). Phase and light shift determination in an optically pumped cesium beam frequency standard, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **47**, 461–465.

[6.19] W.D. Lee, R.E. Drullinger, J.H. Shirley, C. Nelson, D.A. Jennings, L.O. Mullen, F.L. Walls, T.E. Parker, A. Hasegawa, K. Fukuda, N. Kotake, M. Kajita, T. Morikawa (1999). Accuracy evaluations and frequency comparison of NIST 7 and CRL 01, *Proc. joint meeting European Frequency and Time Forum and IEEE Int. Frequency Control Symposium* 62–65 (Besançon, France).

[6.20] C. Salomon, J. Dalibard, W. Philips, A. Clairon, S. Guellati (1990). Laser cooling of cesium atoms below 3 μK, *Europhys. Lett.* **12**, 683–688.

[6.21] A. Clairon, P. Laurent, G. Santarelli, S. Ghezali, S.N. Lea, M. Bouhara (1995). A cesium fountain frequency standard: preliminary measurement, *IEEE Trans. Instrum. Meas.* **44**, 128–131.

[6.22] A. Clairon, S. Ghezali, G. Santarelli, P. Laurent, E. Simon, S.N. Lea, M. Bouhara, S. Weyers, K. Szymaniec (1996). The LPTF preliminary accuracy evaluation of cesium fountain frequency standard, *Proc. 10th European Frequency and Time Forum* 218–223 (Brighton, United Kingdom).

[6.23] A. Clairon (2000). Private communication. Cold atom clocks, to be published in *Proc. ICAP 2000* (Florence, Italy).

[6.24] N.F. Ramsey (1990). Experiments with separated oscillatory fields and hydrogen masers, Nobel lecture, *Rev. Mod. Phys.* **62**, 541–551.

[6.25] C. Cohen-Tannoudji, B. Diu, F. Laloë (1973). *Mécanique quantique*, 2 volumes (Herman, Paris).

[6.26] C. Audoin, N. Dimarcq, V. Giordano, J. Viennet (1992). Physical origin of the frequency shifts in cesium beam frequency standards, *IEEE Trans. Instrum. Meas.* **39**, 412–421.

[6.27] J. Rabian, P. Rochat (1988). Full-digital processing in a new commercial cesium standard, *Proc. 2nd European Frequency and Time Forum* 461–468 (Neuchâtel, Switzerland).

[6.28] L.S. Cutler, R.P. Giffard (1992). Architecture and algorithms for new cesium beam frequency standard electronics, *Proc. IEEE Frequency Control Symposium* 127–133 (Hershey, USA).

[6.29] C. Audoin, F. Hamouda, L. Chassagne, R. Barillet (1999). Controlling the microwave amplitude in optically pumped cesium beam frequency standards, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **46**, 407–413.

[6.30] A. de Marchi, G.D. Rovera, A. Premoli (1984). Pulling by neighbouring transitions and its effect on the performances of cesium beam frequency standards, *Metrologia* **20**, 37–47.

[6.31] L.S. Cutler, C.A. Flory, R.P. Giffard, A. de Marchi (1991). Frequency pulling by hyperfine σ transitions in cesium beam atomic frequency standards, *J. Appl. Phys.* **69**, 2780–2792.

[6.32] *Guide to the expression of uncertainty in measurement*, International Organization for Standardization (Geneva, Switzerland).

[6.33] C. Audoin (1992). Caesium beam frequency standards: classical and optically pumped, *Metrologia* **29**, 113–134.

[6.34] G. Théobald, N. Dimarcq, V. Giordano, P. Cérez (1989). Ground state Zeeman coherence effects in an optically pumped cesium beam, *Opt. Comm.* **71**, 256–261.

[6.35] G.D. Rovera, E. de Clercq, A. Clairon (1994). An analysis of major frequency shifts in the LPTF optically pumped primary frequency standard, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **41**, 245–249.

[6.36] N. Dimarcq, V. Giordano, C. Jacques, G. Théobald, P. Cérez (1991). Experimental study of two-laser optical pumping in a cesium clock, *Proc. 5th European Frequency and Time Forum* 187–189 (Besançon, France).

[6.37] A. Makdissi, J.-P. Berthet, E. de Clercq (1997). Improvement of the short term stability of the LPTF cesium beam frequency standard, *Proc. 11th European Frequency and Time Forum* 564–566 (Neuchâtel, Switzerland).

[6.38] F. Hamouda, G. Théobald, P. Cérez, C. Audoin (2000). Analysis tools for the accurate evaluation of a small frequency standard, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **47**, 449–455.

[6.39] S. Oshima, Y. Nakadan, T. Ikegami, Y. Koga (1991). Light shifts in an optically pumped Cs beam frequency standard, *IEEE Trans. Instrum. Meas.* **40**, 1003–1007.

[6.40] D. Yang, J. Chen (1997). Light frequency shift in small optically pumped cesium beam frequency standards, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **47**, 449–455.

[6.41] P. Forman (1985). Atomichron: The atomic clock from concept to commercial product, *Proc. IEEE* **73**, 1181–1204.

[6.42] A. Ashkin (1970). Acceleration and trapping of particles by radiation pressure, *Phys. Rev. Lett.* **24**, 156–159.

[6.43] T. Hänsch, A. Schawlow (1975). Cooling of gases by laser radiation, *Opt. Comm.* **13**, 68–69.

[6.44] D. Wineland, H. Dehmelt (1975). Proposed $10^{14}\Delta\nu < \nu$ laser fluorescence spectroscopy on Tl^+ mono-ion oscillator III, *Bull. Amer. Phys. Soc.* **20**, 637.

[6.45] C. Cohen-Tannoudji, W.D. Phillips (1990). New mechanisms for laser cooling, *Phys. Today* **43**, 33–40.

[6.46] A. Aspect, J. Dalibard (1994). Le refroidissement des atomes par laser, *La Recherche* **25**, 30–37.

[6.47] S. Chu (1998). The manipulation of neutral particles, Nobel lecture, *Rev. Mod. Phys.* **70**, 685–706.

[6.48] C. Cohen-Tannoudji (1998). Manipulating atoms with photons, Nobel lecture, *Rev. Mod. Phys.* **70**, 707–719.

[6.49] W.D. Phillips (1998). Laser cooling and trapping of neutral atoms, Nobel lecture, *Rev. Mod. Phys.* **70**, 721–741.

[6.50] A.N. Luiten, A.G. Mann, M.E. Costa, D.G. Blair (1995). Power stabilized exceptionally high stability cryogenic sapphire resonator oscillator, *IEEE Trans. Instrum. Meas.* **44**, 132–135.

[6.51] E. Tiesinga, B.J. Verhaar, H.T.C. Stoof, D. van Bragt (1992). Spin-exchange frequency shift in a cesium atomic fountain, *Phys. Rev. A* **45**, R2671–R2673.

[6.52] K. Gibble, S. Chu (1993). Laser cooled Cs frequency standard and a measurement of the frequency shift due to ultracold collisions, *Phys. Rev. Lett.* **70**, 1771–1774.

[6.53] P. Laurent, P. Lemonde, E. Simon, G. Santarelli, A. Clairon, N. Dimarcq, P. Petit, C. Audoin, C. Salomon (1998). A cold atom clock in absence of gravity, *Eur. Phys. J. D* **3**, 201–204.

[6.54] A. Jornod, L.-G. Bernier, H. Schweda, G. Busca (1998). The SHM compact hydrogen maser for space applications. Report on the PEM physics package design verification, *Proc. 12th European Frequency and Time Forum* 121–125 (Warsaw, Poland).

[6.55] C. Salomon, C. Veillet (1998). ACES: Atomic clock ensemble in space, *Proc. 1st symposium on utilisation of the International Space Station* **SP-385**, 295–297 (Darmstadt, Germany).

[6.56] H.G. Robinson, S.R. Jefferts, D.B. Sullivan, L.W. Hollberg, N. Ashby, T. Heavner, J.H. Shirley, F.L. Walls, R.E. Drullinger (1998). Design studies for laser-cooled space clock, *Proc. IEEE Int. Freq. Control Symposium* 37–40 (Pasadena, USA).

[6.57] K. Gibble (1998). Laser-cooled microgravity clocks, *Proc. IEEE Int. Freq. Control Symposium* 41–45 (Pasadena, USA).

[6.58] B. Kokkelmans, B. Verhaar, K. Gibble, D. Heinzen (1997). Predictions for laser-cooled Rb clocks, *Phys. Rev. A* **56**, R4389–R4392.

[6.59] C. Fertig, K. Gibble (2000). Measurement and cancellation of the cold collision shift in an ^{87}Rb fountain clock, *Phys. Rev. Lett.* **85**, 1622–1625.

[6.60] Y. Sortais, S. Bize, C. Nicolas, A. Clairon (2000). Cold collision frequency shifts in an ^{87}Rb atomic fountain, *Phys. Rev. Lett.* **85**, 3117–3120.

[6.61] S. Bize, Y. Sortais, M.S. Santos, C. Mandache, A. Clairon, C. Salomon (1999). High-accuracy measurement of the ^{87}Rb ground-state hyperfine splitting in an atomic fountain, *Europhys. Lett.* **45**, 558–564.

[6.62] J.P. Gordon, H.J. Zeiger, C.H. Townes (1994). The maser: New type of microwave amplifier, frequency standard and spectrometer, *Phys. Rev.* **99**, 1264–1274.

[6.63] N.G. Basov, A.M. Prokhorov (1955). The theory of a molecular oscillator and a molecular power amplifier, *Discuss. Faraday Soc.* **19**, 96–99.

[6.64] D. Kleppner, H.M. Goldenberg, N.F. Ramsey (1962). Theory of the hydrogen maser, *Phys. Rev.* **126**, 603–615.

[6.65] D. Kleppner, H.C. Berg, S.B. Crampton, N.F. Ramsey, R.F.C. Vessot, H.E. Peters, J. Vanier (1965). Hydrogen maser principles and techniques, *Phys. Rev.* **138**, A972–A983.

[6.66] N. Koshelyaevsky (1991). H-masers in the USSR, *Proc. 5th European Time and Frequency Forum* 415–434 (Besançon, France).

[6.67] R.F.C. Vessot, M.W. Levine, L. Mueller, M. Baker (1967). The design of an atomic hydrogen maser system for satellite experiments, *Proc. 21st Annual Frequency Control Symposium* 512–542 (Atlantic City, USA).

[6.68] H.E. Peters, T.E. McGunigal, E.H. Johnson (1968). Hydrogen standard work at Goddard Space Flight Center, *Proc. 22nd Annual Frequency Control Symposium* 464–492 (Atlantic City, USA).

[6.69] H.E. Peters, B.H. Owings, T. Oakley, L. Beno (1987). Hydrogen maser for radioastronomy, *Proc. 41st Annual Frequency Control Symposium* 75–81 (Philadelphia, USA).

[6.70] B.A. Gaygerov, F.S. Rusin, V.P. Sysoev (1991). Portable atomic clock on the basis of an active hydrogen maser, sapphire, *Proc. 5th European Time and Frequency Forum* 293–295 (Besançon, France).

[6.71] H.G. Andresen, E. Pannaci (1966). Servo controlled hydrogen maser cavity tuning, *Proc. 20th Annual Frequency Control Symposium* 402–415 (Atlantic City, USA).

[6.72] N.A. Demidov, E.M. Ezhov, B.A. Sakharov, B.A. Uljanov, A. Bauch, B. Fischer (1992). Investigation of the frequency instability of CH1-75 hydrogen masers, *Proc. 6th European Time and Frequency Forum* 409–414 (Noordwijk, Holland).

[6.73] C. Audoin (1981). Fast cavity auto-tuning system for hydrogen maser, *Rev. Phys. A* **16**, 125–130 and **17**, 273.

[6.74] N. Yahyabey, R. Barillet, J. Viennet, C. Audoin (1987). Accord automatique de la cavité résonnante d'horloges à hydrogène: résultats préliminaires, *Proc. 1st European Time and Frequency Forum* 334–338 (Besançon, France).

[6.75] C. Audoin, J. Viennet, P. Lesage (1981). Hydrogen maser: active or passive?, *J. Phys.* **42** Suppl. 12, C8-159–C8-170.

[6.76] D.A. Howe, F.L. Walls, H.E. Bell, H. Hellwig (1979). A small, passively operated hydrogen maser, *Proc. 33rd Annual Frequency Control Symposium* 554–562 (Atlantic City, USA).

[6.77] E.M. Mattison, E.C. Blomberg, G.U. Nystrom, R.F.C Vessot (1979). Design, construction and testing of a small passive hydrogen maser, *Proc. 33rd Annual Frequency Control Symposium* 549–553 (Atlantic City, USA).

[6.78] H.E. Peters (1978). Small, very small and extremely small hydrogen masers, *Proc. 32nd Annual Frequency Control Symposium* 469–476 (Atlantic City, USA).

[6.79] N.A. Demidov, A.A. Uljanov (1990). Design and industrial production of frequency standards in the USSR, *Proc. 22nd Annual Precise Time and Time Interval Applications and Planning Meeting* 187–208 (Vienna, USA).

[6.80] R.V. Pound (1947). Frequency stabilization of microwave oscillators, *Proc. IRE* **35**, 1405–1415.

[6.81] G. Busca (1979). Passive H-maser, *Proc. 33rd Annual Frequency Control Symposium* 563–568 (Atlantic City, USA).

[6.82] L.S. Cutler, C.L. Searle (1966). Some aspects of the theory and measurement of frequency fluctuations in frequency standards, *Proc. IEEE* **54**, 136–154.

[6.83] D.A. Howe, F.L. Walls (1983). A compact hydrogen maser with exceptional long-term stability, *IEEE Trans. Instrum. Meas.* **32**, 218–223.

[6.84] E.M. Mattison (1992). Physics of systematic frequency variations in hydrogen masers, *IEEE Trans. Instrum. Meas.* **39**, 250–255.

[6.85] R.F.C. Vessot, E.M. Mattison, R.L. Walsworth, I.F. Silvera, H.P. Godfried, C.C. Agosta (1987). A hydrogen maser at temperatures below 1 K, *IEEE Trans. Instrum. Meas.* **36**, 588–593.

[6.86] M.E. Hayden, M.D. Hurlïman, W.N. Hardy (1996). Measurement of atomic-hydrogen spin-exchange parameters at 0.5 K using a cryogenic hydrogen maser, *Phys. Rev. A* **53**, 1589–1604.

[6.87] M. Arditi (1958). L'influence des gaz tampons sur le déplacement de la fréquence et la largeur des raies des transitions hyperfines de l'état fondamental des atomes alcalins, *J. Phys. Radium* **19**, 873–880.

[6.88] M. Arditi, T.R. Carver (1961). Pressure, light and temperature shifts in optical detection of 0-0 hyperfine resonance of alkali metals, *Phys. Rev.* **124**, 800–809.

[6.89] M. Arditi, T.R. Carver (1964). Hyperfine relaxation of optically pumped Rb^{87} atoms in buffer gases, *Phys. Rev.* **136A**, 643–649.

[6.90] M. Arditi (1960). A gas cell atomic clock as a high-stability frequency standard, *IRE Trans. Mil. Electron.* **4**, 25–28.

[6.91] R.J. Carpenter, E.C. Beaty, P.L. Bender, S. Saito, R.O. Stone (1960). A prototype rubidium vapor frequency standard, *IRE Trans. Instrum.* **9**, 132–135.

[6.92] M.E. Packard, B.E. Swartz (1962). The optically pumped rubidium vapour frequency standard, *IRE Trans. Instrum.* **11**, 215–223.

[6.93] M. Bloch, I. Pascaru, C. Stone, T. McClelland (1993). Subminiature rubidium frequency standard for commercial applications, *Proc. IEEE International Frequency Control Symposium* 164–177 (Salt Lake City, USA).

[6.94] P. Rochat, H. Schweda, G. Miletti, G. Busca (1994). Miniaturized rubidium clocks for space and industrial applications, *Proc. 8th European Time and*

Frequency Forum 385–404 (Munich, Germany).

[6.95] W.J. Riley (1992). The physics of the environmental sensitivity of rubidium gas cell atomic frequency standards, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **39**, 232–240.

[6.96] J.C. Camparo, R.P. Frueholz (1986). Fundamental stability limit for the diode-laser-pumped rubidium atomic frequency standard, *J. Appl. Phys.* **59**, 3313–3317.

[6.97] Y. Saburi, T. Koga, S. Kinugawa, T. Imamura, H. Suga, Y. Ohuchi (1994). Short term stability of laser-pumped rubidium gas cell frequency standard, *Electron. Lett.* **30**, 633–635.

[6.98] G. Mileti, P. Thomann (1995). Study of the S/N performance of passive atomic clocks using a laser pumped vapour, *Proc. 9th European Time and Frequency Forum* 271–276 (Besançon, France).

[6.99] M. Têtu, G. Busca, J. Vanier (1973). Short-term frequency stability of the Rb^{87} maser, *IEEE Trans. Instrum. Meas.* **22**, 250–257.

[6.100] G. Busca, R. Brousseau, J. Vanier (1975). Long-term frequency stability of the Rb^{87} maser, *IEEE Trans. Instrum. Meas.* **24**, 291–296.

[6.101] W. Paul (1990). Electromagnetic traps for charged and neutral particles, Nobel lecture, *Rev. Mod. Phys.* **62**, 531–540.

[6.102] F.G. Major, G. Werth (1973). High-resolution magnetic hyperfine resonance in harmonically bound ground-state ^{199}Hg ions, *Phys. Rev. Lett.* **30**, 1155–1158.

[6.103] M.D. McGuire, R. Petsch, G. Werth (1978). Precision determination of the ground state hyperfine separation in $^{199}Hg^{+}$ using the ion-storage technique, *Phys. Rev. A* **17**, 1999–2004.

[6.104] M. Jardino, M. Desaintfuscien, R. Barillet, J. Viennet, P. Petit, C. Audoin (1981). Frequency stability of a mercury ion frequency standard, *Appl. Phys.* **24**, 107–112.

[6.105] L.S. Cutler, R.P. Giffard, M.D. McGuire (1981). A trapped mercury 199 ion frequency standard, *Proc. 13th Annual Precise Time and Time Interval Applications and Planning Meeting* 563–578 (Washington D.C., USA).

[6.106] L.S. Cutler, R.P. Giffard, M.D. McGuire (1985). Thermalization of $^{199}Hg^{+}$ ion macromotion by a light background gas in an RF quadrupole trap, *Appl. Phys. B* **36**, 137–142.

[6.107] J.D. Prestage, R.L. Tjoelker, G.J. Dick, L. Maleki (1992). Ultra-stable Hg^{+} trapped ion frequency standard, *J. Mod. Opt.* **39**, 221–232.

[6.108] R.L. Tjoelker, J.D. Prestage, G.J. Dick, L. Maleki (1993). Long term stability of Hg^{+} trapped ion frequency standards, *Proc. IEEE International Frequency Control Symposium* 132–138 (Salt Lake City, USA).

[6.109] R.L. Tjoelker, J.D. Prestage, L. Maleki (1995). Record frequency stability with mercury in a linear ion trap, *Proc. Fifth Symposium on Frequency Standards and Metrology* 33–38 (Woods Hole, USA) ed. J.C. Bergquist (World Scientific, Singapore).

[6.110] D.J. Berkeland, J.D. Miller, J.C. Berquist, W.M. Itano (1998). Laser-cooled mercury ion frequency standard, *Phys. Rev. Lett.* **80**, 2089–2092.

[6.111] D.J. Berkeland, F.C. Cruz, J.C. Berquist (1997). Sum-frequency generation of continuous-wave light at 194 nm, *Appl. Opt.* **36**, 4159–4162.

[6.112] C. Tamm, D. Schnier, A. Bauch (1995). Radio-frequency laser double-resonance spectroscopy of trapped $^{171}Yb^{+}$ ions and determination of line shifts of the ground-state hyperfine resonance, *Appl. Phys. B* **60**, 19–29.

[6.113] M.J. Sellars, P.T.H. Fisk, M.A. Lawn, C. Coles (1995). Further investigation of a prototype microwave frequency standard based on trapped $^{171}Yb^{+}$ ions, *Proc. International Frequency Control Symposium* 66–73 (San Francisco, USA).

[6.114] D.J. Wineland, W.M. Itano (1987). Laser cooling, *Phys. Today*, June issue, 2–8.

[6.115] D.J. Wineland, J.C. Bergquist, D. Berkeland, J.J. Bollinger, F.C. Cruz, W.M. Itano, B.M. Jelenkovic, B.E. King, D.M. Meekhof, J.D. Miller, C. Monroe, M. Rauner, J.N. Tan (1995). Application of laser cooled ions to frequency standards and metrology, *Proc. Fifth Symposium on Frequency Standards and Metrology* 11–19 (Woods Hole, USA) ed. J.C. Bergquist (World Scientific, Singapore).

[6.116] J.J. Bollinger, D.J. Wineland, W.M. Itano (1995). Spin squeezing applied to frequency standards, *Proc. Fifth Symposium on Frequency Standards and Metrology* 107–114 (Woods Hole, USA) ed. J.C. Bergquist (World Scientific, Singapore).

[6.117] A. Godone, E. Bava, C. Novero (1988). Mg beam frequency standard, *Fourth Symposium on Frequency Standards and Metrology* 78–83 (Ancona, Italy) ed. A. de Marchi (Springer, Berlin).

[6.118] C. Daussy, F. Ducos, G.D. Rovera, O. Acef (2000). Performances of OsO_4 stabilized CO_2 lasers as optical frequency standards near 29 THz, *IEEE Trans. Ultrason., Ferroelec., Freq. Contr.* **47**, 518–521.

[6.119] S.N. Bagayev, A.K. Dmitriyev, P.V. Pokasov, B.N. Skvortsov (1995). He–Ne/CH_4 laser frequency standard for precise measurements, *Proc. Fifth Symposium on Frequency Standards and Metrology* 289–296 (Woods Hole, USA) ed. J.C. Bergquist (World Scientific, Singapore).

[6.120] H.G. Dehmelt (1982). Mono-ion oscillator as potential ultimate laser frequency standard, *IEEE Trans. Instrum. Meas.* **31**, 83–87

[6.121] E. Peik, G. Hollemann, H. Walther (1994). Laser cooling and quantum jumps of a single indium ion, *Phys. Rev. A* **49**, 402–408.

[6.122] W. Nagourney, E. Burt, H.G. Dehmelt (1995). Optical frequency standard using individual indium ions, *Proc. Fifth Symposium on Frequency Standards and Metrology* 341–346 (Woods Hole, USA) ed. J.C. Bergquist (World Scientific, Singapore).

[6.123] R.J. Rafac, B.C. Young, J.A. Beall, W.M. Itano, D.J. Wineland, J.C. Bergquist (2000). Sub-decahertz ultraviolet spectroscopy of $^{199}Hg^{+}$, *Phys. Rev. Lett.* **85**, 2462–2465.

[6.124] T. Udem, J. Reichert, R. Holzwarth, T.W. Hänsh (1999). Absolute optical frequency measurement of the cesium D1 line with a mode-locked laser, *Phys. Rev. Lett.* **82**, 3568–3571.

[6.125] S.A. Diddams, D.J. Jones, L.-S. Ma, S.T. Cundiff, J.L. Hall (2000). Optical frequency measurement across a 104 THz gap with a femtosecond laser frequency comb, *Opt. Lett.* **25**, 186–188.

Chapter 7

[7.1] L. Fairhead, P. Bretagnon (1990). An analytical formula for the time transformation TB–TT, *Astron. Astrophys.* **229**, 240–247.

[7.2] Bureau international de l'heure (1994). *Rapport annuel pour 1973* (Observatoire de Paris).

[7.3] B. Guinot, C. Thomas (1988). Establishment of International Atomic Time, *Rapport annuel de la section du temps du BIPM* **1**, D1–D22 (BIPM, Sèvres, France).
[7.4] J. Azoubib, M. Granveaud, B. Guinot (1977). Estimation of the scale unit duration of time scales, *Metrologia* **13**, 87–93.
[7.5] B. Guinot, M. Feissel, F. Laclare (1970). Bureau international de l'heure (1970). *Rapport annuel pour 1969* (Observatoire de Paris).
[7.6] F. Arias (2000). Private communication.
[7.7] B. Guinot (1988). Atomic time scales for pulsar studies and other demanding applications, *Astron. Astrophys.* **192**, 370–373.

Chapter 8

[8.1] B. Guinot (1979). Basic problems in the kinematics of the rotation of the Earth, *Time and the Earth's rotation, IAU Symp. 82* 7–12 (Reidel, Dordrecht, Boston, London).
[8.2] N. Capitaine, B. Guinot, J. Souchay (1987). A non-rotating origin on the instantaneous equator: definition, properties and use, *Celest. Mech.* **39**, 283–307.
[8.3] N. Capitaine, B. Guinot, D.D. McCarthy (2000). Definition of the Celestial Ephemeris Origin and of UT1 in the International Celestial Reference Frame, *Astron. Astrophys.* **355**, 398–405.
[8.4] S. Aoki, B. Guinot, G.H. Kaplan, H. Kinoshita, D.D. McCarthy, P.K. Seidelmann (1982). The new definition of Universal Time, *Astron. Astrophys.* **105**, 359–361.
[8.5] IERS Conventions (1996). *IERS Technical Note 21*, ed. D.D. McCarthy (Paris Observatory, France).
[8.6] International Earth Rotation Service. *Annual Report* (Paris Observatory, France). See Appendix 4.
[8.7] International Earth Rotation Service. *Bulletin B* (monthly edition, Paris Observatory, France). See Appendix 4.
[8.8] K. Lambeck (1980). *The Earth's variable rotation* (Cambridge University Press, Cambridge).
[8.9] A. Cazenave, K. Feigl (1994). *Formes et mouvements de la Terre* (CNRS Editions, Paris).
[8.10] G.D. Egbert, R.D. Rey (2000). Significant dissipation of tidal energy in the deep ocean inferred from satellite altimeter data, *Nature* **405**, 775–778.
[8.11] F.R. Stephenson, L.V. Morrison (1980). *Phil. Trans. R. Soc. London Ser. A* **313**, 47.
[8.12] F. Mignard (1994). Hipparcos: une mission d'astrométrie spatiale, *Bulletin Bur. Nat. Métrologie* **95**, 11–17.
[8.13] J. Kovalevsky *et al.* (1992). The FAST Hipparcos data reduction consortium: overview of the reduction software, *Astron. Astrophys.* **258**, 7–17.
[8.14] J. Kovalevsky *et al.* (1995). Construction of the intermediate Hipparcos astrometric catalogue, *Astron. Astrophys.* **304**, 34–43.
[8.15] D. Benest, C. Froeschle (eds.) (1992). *HIPPARCOS, une nouvelle donne pour l'astronomie* (Observatoire de la Côte d'Azur et Société française des spécialistes d'astronomie, Goutelas).
[8.16] A. Hewish, S.J. Bell, J.D.H. Pilkington, P.F. Scott, R.A. Collins (1968). Observation of a rapidly pulsating radio source, *Nature* **217**, 709–713.

[8.17] D.C. Backer, R.R. Kulkarni, C. Heiles, M.M. Davis, W.M. Goss (1982). A millisecond pulsar, *Nature* **300**, 615–618.
[8.18] G. Petit, P. Tavella (1996). Pulsars and time scales, *Astron. Astrophys.* **308**, 290–298.
[8.19] J.H. Taylor, A. Wolszczan, T. Damour, J.M. Weisberg (1992). Experimental constraints on strong field relativistic gravity, *Nature* **355**, 132.
[8.20] N. Mashall (1993). Des pulsars binaires aux ondes gravitationnelles, *La Recherche* **260**, 1390–1391.
[8.21] B. Guinot (1989). General principles of the measure of time, astronomical time, in *Reference frames in astronomy and geophysics*, 351–377, eds. J. Kovalevsky, I.I. Mueller, B. Kolaczec (Kluwer Academic Publishers, Dordrecht, Boston, London).

Chapter 9

[9.1] B. de Beauvoir, C. Schwob, O. Acef, L. Josefowski, L. Hilico, F. Nez, L. Julien, A. Clairon, F. Biraben (1997). Metrology of the hydrogen and deuterium atoms: determination of the Rydberg constant and Lamb shifts, *Eur. Phys. J. D* **12**, 61–93, and references therein.
[9.2] H. Dehmelt (1990). Experiments with an isolated subatomic particle at rest, Nobel lecture, *Rev. Mod. Phys.* **62**, 525–530.
[9.3] H. Gabrielse, X. Fei, L.A. Orozco, R.L. Tjoelker, J. Haas, H. Kalinowsky, T.A. Trainor, W. Kells (1990). Thousandfold improvement in the measured antiproton mass, *Phys. Rev. Lett.* **65**, 1317–1320.
[9.4] R.S. van Dick, D.L. Farnham, P.B. Schwinberg (1995). Proton–electron mass ratio and the electron's atomic mass, *IEEE Trans. Instrum. Meas.* **44**, 546–549.
[9.5] T. Kinoshita (1995). New value of the α^3 electron anomalous magnetic moment, *Phys. Rev. Lett.* **75**, 4728–4731.
[9.6] S. Weinberg (1989). Testing quantum mechanics, *Ann. Phys.* **194**, 336–386.
[9.7] J.J. Bollinger, D.J. Heinzen, W.N. Itano, S.L. Gilbert, D.J. Wineland (1989). Test of the linearity of quantum mechanics by rf spectroscopy of the $^9Be^+$ ground state, *Phys. Rev. Lett.* **63**, 1031–1034.
[9.8] R.L. Walworth, I.F. Silvera, E.M. Mattison, R.F.C. Vessot (1990). Test of the linearity of quantum mechanics in an atomic system with a hydrogen maser, *Phys. Rev. Lett.* **64**, 2599–2602.
[9.9] G.D. Fletcher, S.J. Lipson, D.J. Larson (1987). Observation of a magnetic-field-dependent g-factor ratio, *Phys. Rev. Lett.* **58**, 2535–2538.
[9.10] N. Fortson (1987). New term in atomic Zeeman energy, *Phys. Rev. Lett.* **59**, 988–990.
[9.11] A. de Marchi (1993). The high C-field concept for an accurate cesium beam resonator, *Proc. 7th European frequency and time forum* 541–548 (Neuchâtel, Switzerland).
[9.12] C.M. Will (1993). *Theory and experiment in gravitational physics* (revised edn.) (Cambridge University Press, Cambridge).
[9.13] T. Damour (1994). Gravitation and quantizations, *Proc. LVIIth Les Houches summer school, July 1992*, eds. B. Julia, J. Zinn-Justin (North Holland, Amsterdam).
[9.14] P. Wolf (1997). *Relativity and the metrology of time*, PhD. Thesis, Queen

Mary and Westfield College, University of London. (Also Monographie 97/1, BIPM, Sèvres, 1997.)

[9.15] J.C. Hafele, R.E. Keating (1972). Around-the-world atomic clocks: predicted relativistic time gains, observed relativistic time gains (two articles), *Science* **77**, 166–170.

[9.16] R.V. Pound, G.A. Rebka Jr. (1960). Apparent weight of photons, *Phys. Rev. Lett.* **4**, 337–341.

[9.17] R.V. Pound, G.A. Rebka Jr., J.L. Snider (1965). Effect of gravity on gamma radiation, *Phys. Rev. Lett. B* **140**, 788–803.

[9.18] R.F.C. Vessot, M.W. Levine (1979). A test of the equivalence principle using a space-borne clock, *J. Gen. Rel. Grav.* **10**, 181–204. (See also R.F.C. Vessot *et al.* (1980). *Phys. Rev. Lett.* **45**, 2081–2084.)

[9.19] A. Godone, C. Novero, P. Tavella, K. Rahimullah (1993). New experimental limits to the time variations of $g_p(m_e/m_p)$ and α, *Phys. Rev. Lett.* **71**, 2364–2366.

[9.20] A. Godone, C. Novero, P. Tavella (1995). Null gravitational redshift experiment with non-identical atomic clocks, *Phys. Rev. D* **51**, 319–323.

[9.21] J.D. Prestage, R.L. Tjoelker, L. Maleki (1995). Atomic clocks and variations of the fine structure constant, *Phys. Rev. Lett.* **74**, 3511–3514.

[9.22] A.A. Michelson, E.W. Morley (1887). *Am. J. Sci.* **34**, 333.

[9.23] T.P. Krisher, L. Maleki, G.F. Lutes, L.E. Primas, R.T. Logan, J.D. Anderson, C.M. Will (1990). Test of the isotropy of the one-way speed of light using hydrogen-maser frequency standards, *Phys. Rev. D (Rapid Commun.)* **42**, 731.

[9.24] P. Wolf (1995). Proposed satellite test of special relativity, *Phys. Rev. A* **51**, 5016–5018.

[9.25] P. Wolf, G. Petit (1997). Satellite test of special relativity using the GPS, *Phys. Rev. A* **56**, 4405–4409.

[9.26] I.I. Shapiro (1990). *General relativity and gravitation*, 313, eds. N. Ashby, D.F. Bartlett, W. Wyss (Cambridge University Press, Cambridge).

[9.27] K. Nordtvedt Jr. (1973). Post-Newtonian gravitational effects in lunar laser ranging, *Phys. Rev. D.* **7**, 2327–2356.

[9.28] R.N. Treuhaft, S.T. Lowe (1991). A measurement of planetary relativistic deflection, *Astron. J.* **102**, 1879–1888.

[9.29] K. Nordtvedt Jr. (1968). Equivalence principle for massive bodies: II Theory, *Phys. Rev.* **169**, 1017–1025.

[9.30] C.M. Will (1971). Theoretical frameworks for testing relativistic gravity: II Parametrized post-Newtonian hydrodynamics and the Nordtvedt effect, *Astrophys. J.* **163**, 611–628.

[9.31] *Guide to GPS Positioning* (1986). Ed. D. Wells (Canadian GPS Associates).

[9.32] B. Guinot, W. Lewandowski (1989). Improvement of the GPS time comparisons by simultaneous relative positioning of the receiver antennas, *Bull. Géodésique* **63**, 371–386.

[9.33] B. Juompan, P. Uhrich, M. Brunet (1995). Aspects temps du complément européen au Global Positioning System (CE-GPS): résultats expérimentaux des phases continentales et transatlantiques, *Bull. Bureau Nat. Métrologie* **101**, 13–24.

[9.34] M. Brunet (1992). DORIS precise orbit determination and location system, performance of ultra stable oscillators, *Proc. 6th European Frequency and Time Forum* 125–136 (Noordwijk, Holland).

[9.35] W. Schäfer, H. Wilmes (1986). Precise range and range-rate equipment (PRARE). System design and status of development, *Workshop on advances*

in satellite radio tracking (Austin, Texas).

[9.36] R.T. Schilizzi (1995). Current development in VLBI astronomy on the ground and in space, *Radio Sci. Bull.* **274**, 14–28.

[9.37] D. Benest, C. Froeschle (eds.) (1992). *HIPPARCOS, une nouvelle donne pour l'astronomie* (Observatoire de la Côte d'Azur et Société française des spécialistes d'astronomie, Goutelas).

[9.38] L. Lindegren, J. Kovalevsky (1995). Linking the Hipparcos catalogue to the extragalactic reference system, *Astron. Astrophys.* **304**, 189–201.

[9.39] CNES (1986). Announcement of opportunity: altimetry research in ocean circulation, TOPEX/POSEIDON, a CNES/NASA oceanographic satellite.

[9.40] B. Dubouis (1986). Temps fréquences dans les télécommunications françaises, *Bull. Bureau Nat. Métrologie* **63**, 58–67.

[9.41] P. Kartaschoff (1991). Synchronization in digital communications networks, *Proc. IEEE* **79**, 1019–1028.

[9.42] CCDS (1993). Rapport de la 12ème session (see p. S17 and recommendation S5, p. S29).

Index

absolute
  space, 17, 19
  time, 6, 17
absorption, 116
accuracy
  frequency, 55, 66, 87–88
  of TAI readings, 252
active clock, 126
algorithm
  for frequency stability, 243–245
  for TAI, 241
  for time scale, 60, 240–241
Algos, 247
alkali atom, 110–116
Allan variance
  definition, 74
  estimator, 76, 78
  modified, 80
  practical calculation, 76–79
  systematic effects, 79
  variation, 75
apparent solar
  day, 38
  time, 40
astronomical unit of length, 30
atomic fountain, 130, 177–180
atomic second, 56, 236
atomic time
  integrated, 56, 57
  local independent, 258

barycentric coordinate system, 29
Barycentric Coordinate Time (TCB), 30, 238
Binary Pulsar Time (TBP), 284
Boltzmann's law, 120
Breit–Rabi formula, 114, 149

caesium clock
  cold atom fountain, 177
  frequency shifts, 147–159, 168–169, 181
  inaccuracy, 161, 169, 183
  long term frequency instability, 160, 169
  magnetically deflected beam, 130–162
  medium term frequency instability, 144, 167, 181
  non-reproducibility, 160, 169
  optically pumped beam, 162–170
cavity pulling, 128
celestial
  frame, 276
  pole, 265
  sphere, 267
Celestial Ephemeris Origin, 268
Celestial Ephemeris Pole (CEP), 265
Celestial Intermediate Pole (CIP), 265
chaos, 7
civil time, 42
clock
  active, 126
  comparison, 35
  definition, 65
  in space, 183
  master, 247
  passive, 126, 129, 147
  primary, 262
  transition, 115
  transportation, 93, 291
closing error, 93
common view method, 99
comparison
  of clocks, 35
  of frequencies, 104–108
  of times, 59, 90–104
cooling atoms
  by Doppler effect, 171
  by Sisyphus effect, 174
  launching the atoms, 176
cooling ions
  collisional method, 224
  radiative method, 229

coordinate
  quantity, 22, 28
  simultaneity, 35
  synchronisation, 35, 37
  time (definition), 11, 28, 37
coordinate system, 17
  barycentric, 29
  geocentric, 29
Coordinated Universal Time, 39, 255–259
cycling transition, 122, 166

date, 17, 38
day, length of, 266
definitive time, 45
Dicke condition, 124, 223
Doppler cooling, 171
Doppler effect, 23
  Dicke condition, 124, 223
  elimination, 123–125
  for positioning, 301
  for satellite tracking, 272
  frequency shift, 123
  second order, 125, 151
DORIS, 272, 301, 306
double resonance method, 211
duration, 7

Earth Reconnaissance Satellites (ERS), 302, 303, 305
Earth rotation, 264
  irregularities, 47, 273–276
  measurement, 269–273
  parameters, 265, 278
Echelle atomique libre (EAL), 242
Einstein's equivalence principle, 21, 109
ellipsoid, reference, 303
energy–time uncertainty relation, 118
ephemeris, 10, 13, 276
Ephemeris Time, 15, 50–51
  definition, 278–279
  determination, 279–280
equation of time, 40, 41
equinox, 269
event, 17
excited state, 110

fine structure, 110
fluctuations
  model, 72–73
  power spectral density, 71
fluorescence, 117
frequency
  instantaneous, 68
  nominal, 57
  normalised, 57
frequency accuracy, 55, 66, 87–88
frequency comparison, 104–108
frequency instability
  definition, 67
  measurement, 70
  of atomic clocks, 147
  of TAI, 254
frequency noise, 73
frequency shift
  Doppler, 123
  gravitational, 32, 292
  hydrogen maser, 207
  in caesium clock, 147–159, 168–169, 181
  in rubidium clock, 215
  mercury ion clock, 228
frequency standard
  characteristics, 234
  inaccuracy, 234
  magnesium beam, 231
  mercury ion, 224–229
  optical, 231
  primary, 129
  primary and secondary, 65
  single ion at rest, 231
  ytterbium ion, 229
frequency steering, 243, 246

Galilean relativity principle, 19
geocentric coordinate system, 29
Geocentric Coordinate Time (TCG), 30, 237
geoid, 31, 303
Global Positioning System, 60, 98, 297–301
GLONASS, 98, 100, 301
GLONASS Time, 262
GMT, 42
GPS, 60, 98, 297–301
GPS Time, 98, 261
gravitation, 15, 26, 291
gravitational
  frequency shift, 32, 292
  potential (terrestrial), 32
Greenwich Mean Time, 42
ground state, 110

HIPPARCOS, 277, 302
hydrogen maser, 184–211
  active, 191–201
  cavity tuning, 192–195, 199–201
  cryogenic, 210
  frequency shifts, 207
  inaccuracy, 209
  induced magnetic moment, 193
  long term frequency instability, 209
  medium term frequency instability, 206
  oscillation condition, 191
  passive, 201–206
hyperfine
  filter, 212
  structure, 111–116

ICRF, 31, 264, 302
inaccuracy, 88
  of caesium clock, 161, 169, 183
  of frequency standards, 234
  of hydrogen maser, 209
  of mercury ion clock, 229
  of TAI frequency, 254
induced emission, 117
inertial system, 19
instability
  frequency, 67, 70
  time, 67, 69, 81
instantaneous
  axis of rotation, 265
  frequency, 68
  longitude origin, 267
instrumental delay, 90
integrated
  atomic time, 56, 57
  time scale, 12
International Atomic Time, 60, 237–239
  accuracy of readings, 252
  algorithm, 241
  frequency inaccuracy, 254
  frequency instability, 254
  origin, 237
  scale unit, 106, 237, 260
International Celestial Reference Frame, 31, 264, 302
International GPS Service (IGS), 272, 300
International Laser Ranging Service (ILRS), 272
International Terrestrial Reference Frame, 31, 264
interval, 22, 25
ion confinement
  micromotion, 221
  Paul trap, 218
  restoring force, 221
  secular motion, 222
  three-d quadrupole field, 219
  two-d quadrupole field, 218
ITRF, 31, 264

Josephson effect, 288
Julian Date (JD), 39

laser ranging, 272, 306
LASSO, 102
leap second, 256
length of day, 266
Loran-C, 59, 97
Lorentz transformation, 21
lunar
  month, 39
  reflecting panels, 272
lunation, 39
lunisolar precession–nutation, 265

magnesium beam frequency standard, 231
magnetic
  deflection, 120, 130
  shielding, 125
master clock, 43, 247
mean solar time, 40, 46
mercury ion clock
  frequency shift, 228
  inaccuracy, 229
  long term frequency instability, 229
  medium term frequency instability, 228
mercury ion frequency standard, 224–229
metric, 22
  tensor, 25
Minkowski metric, 22
Modified Julian Date (MJD), 39

N-sample variance, 74
Navsat, 272, 301
Newton's first law, 18
noise
  frequency, 73
  shot, 144
  thermal, 206
nominal frequency, 57
non-rotating origin, 268
normalised frequency, 57

optical
  detection, 122
  frequency standard, 231
  molasses, 173
  pumping, 121
origin
  instantaneous longitude, 267
  non-rotating, 268

passive clock, 126, 129, 147
Paul trap, 218
pendulum clock, 83
physical constants, 289
piezoelectric effect, 84
pole
  celestial, 265
  terrestrial, 265
power spectral density, 71
PPN formalism, 294
PRARE, 302
precession–nutation, lunisolar, 265
primary clock, 262
proper
  frequency, 87, 104
  quantity, 28
  second, 36
  time, 10, 21, 65, 69, 236

pulling
  cavity, 128
  Rabi, 157
  Ramsey, 157
pulsar spin-down, 283
Pulsar Time, 64, 281–285
pumping
  optical, 121
  transition, 122
  two-laser, 166

quality factor
  of atomic resonance, 118
  resonant cavity, 155
quantum Hall effect, 288
quartz oscillator, 84, 140–144, 196–199, 204, 228
quasar, 24, 29, 31, 100, 270

Rabi
  frequency, 189
  pedestal, 133
  pulling, 157
Ramsey
  fringe, 133
  pulling, 157
Ramsey method
  by separated oscillating fields, 131
  Rabi pedestal, 133
  response curve, 135
  time domain, 227, 230
reference ellipsoid, 303
relative frequency offset, 57, 68
relativity principle, Galilean, 19
reproducibility postulate, 7
resonance curve, 133, 135, 141, 151, 165, 179, 202, 227
resonant cavity
  quality factor, 155
rise time, 91
rotation, 24
  of the Earth, 47, 264, 269–276
rubidium clock, 211–217
  double resonance, 211
  frequency shifts, 215
  hyperfine filter, 212
  long term frequency instability, 216
  medium term frequency instability, 215

Sagnac effect, 96
scale unit, 28, 237
  TAI, 106, 237, 260
second
  atomic, 56, 236
  leap, 256
  of Ephemeris Time, 50
  of mean time, 46
  proper, 36
  unit of duration, 12
selection rules, 116
Selective Availability (SA), 99, 299
servo control
  of amplitude, 155
  of frequency, 143
  of magnetic field, 150
  of phase, 198
shot noise, 144
sidereal time, 43, 269
signal transmission, 92, 95, 100
simultaneity, 17
  coordinate, 35
Sisyphus effect, 174
speed of light, 19
spontaneous emission, 117
stability algorithm, 243–245
standard frequency emissions, 260
Stark effect, 116
  second order, 150
state preparation, 119
steering, of frequency, 243, 246
stellar angle, 267
stored ion clock, 217–230
synchronisation
  coordinate, 35, 37
  Einstein convention, 20

TAI, 60, 237–239
  accuracy of readings, 252
  algorithm, 241
  frequency inaccuracy, 254
  frequency instability, 254
  origin, 237
  scale unit, 106, 237, 260
telecommunications, 306
terrestrial
  frame, 276
  pole, 265
Terrestrial Ephemeris Origin, 268
Terrestrial Time (TT), 33, 238, 259
thermal noise, 206
three-corner hat method, 83
tidal potential, 23, 29, 32, 303
time
  absolute, 6, 17
  coordinate, 11, 28, 37
  proper, 10, 21, 65, 69, 236
time comparison
  methods, 90–104
  need for, 59
time instability
  characterisation, 81
  definition, 67
  measurement, 69

time interval counter, 91
time scale
  algorithm, 60, 240–241
  integrated, 12, 57
  notion, 10
time signal, 256
time standard, primary, 129
time zone, 41
timekeeper, 43
TOPEX–POSEIDON, 272, 301, 303–306
transition
  clock, 115
  cycling, 122, 166
  pumping, 122
tropical year, 38
two-laser pumping, 166
two-sample variance, 74
two-way time transfer, 92, 100

uncertainty relation, energy–time, 118
unit of proper time, 22
Universal Time, 43, 264–278
  definition, 264–266
  measurement techniques, 269–273
  other forms, 269
  use, 276
UT1, 266–269
UTC, 39, 255–259

VLBI, 270, 302, 306

World Geodetic System, 99
worldline, 26

ytterbium ion frequency standard, 229

Zeeman effect, 114
  second order, 148

The Measurement of Time

This book provides a comprehensive introduction to the physics of time and time measurement, from an historical perspective to the modern day. It discusses the stability and accuracy of atomic frequency standards, including different types of oscillators and atomic clocks, and covers the most recent developments and uses of these devices. The precision of atomic clocks and the atomic time scale are considered in the context of fundamental physical research, with relation to general relativity and searches for possible time variation in the fundamental constants. The authors also discuss International Atomic Time and its relationship to Coordinated Universal Time and the time scales used in astronomy, as well as applications such as the Global Positioning System (GPS).

The book will be ideal as an introduction for graduate students or researchers entering the fields of time and frequency metrology and precise astronomical observation. It will also be useful as a reference for scientists working in these and other applied areas, such as geophysics, atomic physics, astronomy and telecommunications. General readers with a background in science should also find this a fascinating book.

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 0-521-00397-0

9 780521 003971