THE INTERNET BOOK

Everything You Need to Know about Computer Networking and How the Internet Works

FIFTH EDITION

Douglas E. Comer

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

The Internet Book

Everything You Need to Know about Computer Networking and How the Internet Works

Fifth Edition

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

The Internet Book

Everything You Need to Know about Computer Networking and How the Internet Works

Fifth Edition

Douglas E. Comer
Departments of Computer Science and ECE
Purdue University
West Lafayette, IN

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20180730

International Standard Book Number-13: 978-1-138-33133-4 (Hardback)
International Standard Book Number-13: 978-1-138-33029-0 (Paperback)

Library of Congress Cataloging-in-Publication Data

Names: Comer, Douglas, author.
Title: The Internet book : everything you need to know about computer networking and how the Internet works / Douglas E. Comer.
Description: Fifth edition. | Boca Raton : Taylor & Francis, CRC Press, 2018. | Includes bibliographical references and index.
Identifiers: LCCN 2018021320 | ISBN 9781138330290
Subjects: LCSH: Internet. | Computer networks. | World Wide Web.
Classification: LCC TK5105.875.I57 C65 2018 | DDC 004.67/8--dc23
LC record available at https://lccn.loc.gov/2018021320

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Everyone Who
Is Curious

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

About The Author

Douglas Comer is a Distinguished Professor at Purdue University in the departments of Computer Science and Electrical and Computer Engineering. He has created and enjoys teaching undergraduate and graduate courses on computer networks and internets, operating systems, computer architecture, and computer software. One of the researchers who contributed to the Internet as it was being formed in the late 1970s and 1980s, he has served as a member of the Internet Architecture Board, the group responsible for guiding the Internet's development. Comer is an internationally recognized expert on computer networking, the TCP/IP protocols, and the Internet, who presents lectures to a wide range of audiences. In addition to research articles, he has written a series of textbooks that describe the technical details of the Internet. Comer's books have been translated into many languages, and are used in industry as well as computer science, engineering, and business departments around the world. He is a Fellow of The Association for Computing Machinery (the major professional society in computer science) and editor of the scientific journal, *Software — Practice and Experience*.

Professor Comer had dial-up Internet access from his home in the late 1970s, has enjoyed a direct connection with 24-hour-per-day service since 1981, and uses the Internet daily. He wrote this book as a response to everyone who has asked him for an explanation of the Internet that is both technically correct and easily understood by anyone. An Internet enthusiast, Comer displays *INTRNET* on the license plate of his car.

Additional information can be found at:

www.cs.purdue.edu/people/comer

and information about Comer's books can be found at:

www.comerbooks.com

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Preface

The Internet Book explains how computers communicate, what the Internet is, how the Internet works, and what services the Internet offers. It is designed for readers who do not have a strong technical background — early chapters clearly explain the terminology and concepts needed to understand all the services. When you finish reading, you will understand the technology behind the Internet, will appreciate how the Internet can be used, and discover why it is so exciting. In addition, you will understand the origins of the Internet and see how rapidly it has grown.

Instead of using mathematics, algorithms, or computer programs, the book uses analogies from everyday life to explain technology. For example, to explain why digital communication is superior to analog, the text uses an analogy of sending signals through fog with a flashlight. To explain how audio can be played back for the user at a steady rate when packets arrive in clumps, the text uses the analogy of smart phones arriving at a distribution center in one shipment, but being sold to customers one at a time.

In addition to explaining the services users encounter such as email, video streaming, instant messaging, and web browsing, the text covers key networking concepts such as packet switching, Local Area Networks, protocol software, and domain names. More important, the text builds on fundamentals — it describes basic Internet communication facilities first, and then shows how the basic facilities are used to provide a variety of services.

The fifth edition retains the same general structure as the previous edition, but adds three new chapters (19, 26, and 32), and updates material throughout. Chapter 19 explains NAT, a technology many Internet subscribers now have in their home. Chapter 26 explains blogs and wikis, two recent Internet applications. The third new chapter, 32, explains Virtual Private Networking, a technology that allows a trusted user, such as an employee, to access an organization's network safely from an arbitrary remote location.

As with the previous edition, the book is divided into four main parts. The first part begins with fundamental concepts such as digital and analog communication. It also introduces packet switching and explains the Local Area Network technologies that are used in most businesses and in many homes.

The second part of the book gives a short history of the Internet research project and the development of the Internet. Although most of the history can be skipped, readers should pay attention to the phenomenal growth rate, which demonstrates that the

technology was designed incredibly well — no other communication technology has remained as unchanged through such rapid growth.

The third part of the book explains how the Internet works, including a description of the two fundamental protocols used by all services: the Internet Protocol (IP) and the Transmission Control Protocol (TCP). Although they omit technical details, the chapters in this part allow students to understand the essential role of each protocol and gain perspective on the overall design.

The fourth part of the book examines services available on the Internet. In addition to covering browsers, web documents, and search engines used with the World Wide Web, chapters discuss email, bulletin boards, file transfer, remote desktops, wikis, blogs, and audio and video communication. In each case, the text explains how the service operates and how it uses facilities in the underlying system. The fourth part concludes with a discussion of network security, Virtual Private Networks, and electronic commerce.

The Internet Book serves as an excellent reference text for a college-level course on the Internet. Although presented in a nontechnical manner, the material is scientifically accurate. More important, in the twenty-first century, an educated person needs to know more than how to use a browser or set up a web page — they should have some understanding of what goes on behind the scenes. They can acquire such knowledge from this text.

Instructors are encouraged to combine classroom lectures with laboratory sessions in which students see and use the technology first-hand. In all courses, early labs should focus on exploring a variety of services, including sending email, using a browser, using a search engine, downloading files, listening to audio, and using an IP telephone, if one is available. I encourage all students, even those who have no interest in computers, to build a trivial web page by hand. In addition to helping them see the relationship between tags in an HTML document and the resulting display, it shows students how a server transfers files on a computer disk to a browser. Seeing the relationship in labs helps one better understand as they read about the underlying process.

Lab projects later in the semester depend on the type of course. Business-oriented courses often focus students on using the Internet or constructing a case study — labs require students to search the Internet for information and then write a paper that analyzes the information. Other courses use labs to focus on tools such as programs used to create a web page. Some courses combine both by having students search for information and then create a web page that contains links to the information. In any case, we have found that students enter Internet courses with genuine enthusiasm and motivation; a professor's task is merely to provide perspective and remind students throughout the semester why the Internet is so exciting.

The author thanks many people who have contributed to editions of this book. John Lin, Keith Rovell, Rob Slade, and Christoph Schuba read early versions and made suggestions. Dwight Barnette, George Polyzo, Donald Knudson, Dale Musser, and

Dennis Ray sent the publisher reviews of a previous edition. Scott Comer offered perspective. Sharon Comer and Mark Kunschke checked details and provided many useful suggestions for this edition. As always, my wife, Christine, carefully edited the manuscript, solved many problems, and improved the wording.

Douglas E. Comer

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

This book was typeset by the author and sent across the Internet in digital form to a publishing company where it was edited and sent to be printed.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

The Internet Is Everywhere

A revolution has occurred. It started quietly, and has grown to involve the entire world. It is the Internet. On a given day, people around the world use Internet services:

- A college student uses a smart phone to record a stunning touchdown, and then uploads the video to allow others to view it.
- While walking down the street in a city, a teenager runs an app to find others nearby who are interested in playing an online game.
- A person suffering from a chronic disease wears a battery-powered monitor that sends an update to their doctor every fifteen minutes.
- Parents use a laptop computer to view the weather where their child lives, and are relieved to see the storm has passed.
- A family on vacation in Switzerland uses a smart phone to contact their home security system and see views of the interior of their home.
- A teenager uses a smart phone to listen to a sample of music. Later, he uses the smart phone to purchase and download a copy of the song.
- A grandparent uses a laptop computer to view photos their grandchild has uploaded. Later, the grandparent uses the laptop to find airline flights, make a reservation, and purchase a ticket to visit the grandchild.
- Two friends keep in touch by posting a log of their activities for each other to view on a social media site.
- A group of company executives holds a meeting. Each sits in front of a computer that has a camera and microphone, and they see video of one another on their screens and hear each other's voices.

- A company runs a computer program at the close of business each day that sends encrypted copies of the daily activity to multiple storage locations.
- A group of friends are on a road trip when bad weather closes the road. They use a smart phone to find a hotel room, and then use a map program on their smart phone to navigate to the hotel.
- An author finishes writing a short story, and publishes the story on a web site; readers around the world download a copy.

We begin our study of the Internet with a basic question:

"What is the Internet?"

Most people think the Internet is a set of services. They might start with Facebook, Amazon, Netflix, and Google, and then go on to list other services, such as Instagram, Snapchat, and YouTube. They may give more general categories, such as "the Web" or "email." It doesn't matter. They are all incorrect because the Internet is none of those.

So, what is the Internet? In this book, we will learn that it is a global computer communication system that has made all the services possible. In short, the Internet has enabled the revolution that has changed the way we live, work, and play.

Let's look at some facts:

- Scientists and researchers have been using the Internet since the 1980s, long before most services were invented.
- The Internet reaches every populated area of the planet and connects billions of people.
- Celebrities use the Internet to reach their fans.
- The majority of couples getting married in the US met online; the trend is the same in other developed countries.
- Most companies use the Internet to conduct business.
- Schools, ranging from elementary through college, have access to the Internet, and students use the Internet routinely to obtain assignments from teachers, look up facts, and submit their work.
- Military organizations use the Internet; it played a role in military actions as early as Operation Desert Storm in the early 1990s.
- Government agencies in many countries use the Internet.

1.1 Basic Facts Do Not Tell The Story

The most common assessments of the Internet's significance focus on the number of devices that connect to it or the number of people that use it every day. However, such numbers only tell part of the story. The Internet reaches ships at sea, planes in the air, and vehicles on land. Internet connected devices surround us, and include security systems, vending machines, surveillance cameras, and televisions and other common household appliances. In short, the Internet is everywhere.

To assess the impact of the Internet, one might ask, "What has it affected?" The answer is, "Almost everything." So, the question becomes:

How does the Internet affect you daily?

1.2 Imagine Life Without The Internet

To appreciate how the Internet impacts you every day, imagine that you are transported back in time before the Internet. There would be no Facebook, Instagram, Twitter, Google, Snapchat, Netflix, DuckDuckGo, or Amazon. There would be no smart phones, no Wi-Fi hot spots, and online games. You could not access iTunes, Reddit, YouTube, or online dating sites. In fact, there would be no online shopping, no photo sharing, and no email.

Without the Internet, you would feel cut off from the easy, instantaneous access to information that we take for granted. If you saw an item in a store or in a mail-order catalog, you could not search online for evaluations and reviews. You could not compare prices without visiting other stores or waiting for other catalogs to arrive. When information does become available, it would seem stale. For example, instead of immediate online access to weather radar whenever you want to view it, you would read the weather forecast in the morning newspaper, knowing that the information had been compiled and printed the night before. The media — radio, television, and newspapers — would present news stories and summaries of the previous day's sports events. If you wanted to know more than the story reported, you could not search online. You could not view photos or videos of an event that spectators had taken with their devices, and could not read the opinions of others. Instead, you would need to wait for later follow-up stories, and hope that instead of reusing the same photos, the media would print new pictures.

The point is:

> *If we imagine life without the Internet, we can see that Internet services have become deeply embedded in our daily lives, and that instantaneous access to information has changed just about everything.*

1.3 Why You Should Understand Internet Technology

Why should an average person care how the Internet works? It may seem that the technology is irrelevant, and that users can enjoy the Internet as if it is magic. However, basic knowledge can help in two ways. On the one hand, the Internet has become such an exciting and inescapable part of life that every educated person should understand what it provides and what it can do for them. Understanding the technology allows one to dream of new ways it can be used. On the other hand, learning about the Internet will help you avoid fraud, scams, and exaggerated marketing claims that arise whenever a new technology arises. Understanding how the Internet works will help you be less vulnerable and make smarter decisions. A personal example will explain how individuals can be vulnerable to false claims.

One day, a salesman came to my door selling "a new, higher speed Internet service." I knew all about the Internet because I had worked on the Internet project since the 1970s, had written books explaining the technology to engineers, and had consulted for many of the major companies that build and use Internet equipment. Of course, the salesman had no idea who I was, and launched into his typical spiel. Within a few sentences, I realized the salesman's claims were completely invalid. So, I stopped him and asked a series of questions. "If I sign up for your service, will I be able to download an HD movie in less than 15 seconds?" "Will I always receive tweets before all my neighbors?" "Will email I send be delivered much faster than it is now?" "When I click on a link, will the new page always appear without delay?"

The salesman gave an incorrect answer to every question. Yet, he seemed quite confident and earnest, never wavering in his conviction that the new service solved all problems. Why are salespeople so convincing? There's a joke among network companies that explains the situation:

> What's the difference between a network salesperson and a used car salesperson? The used car salesperson knows when they are lying.

After the salesman left, it occurred to me that without any knowledge of how the Internet works, a consumer is not able to distinguish fact from fiction, and is susceptible to hype, especially from someone who seems sincere, confident, and knowledgeable. It happens with all new technologies. To avoid being duped, one must understand the basics.

This book considers the Internet in the broadest sense. It removes some of the mystery and helps the reader understand how the technology works.

1.4 Learning About The Internet

Learning about the Internet is not something one can complete in an afternoon — learning never stops because the Internet keeps changing. When new information ap-

pears, it replaces older information, and when new services appear, they replace older services. Each time you visit the Internet, you can find something new.

Of course, information on the Internet changes much more rapidly than humans can imagine. In fact, because information on the Internet comes from computers and automated systems, the changes can seem instantaneous to humans. For example, if one accesses weather information twice in a single minute, the information obtained from the two accesses can differ because computers can measure weather and change the report constantly.

Like a traditional library, the Internet offers tools and services that catalog information and aid users who are searching for specific topics. Unlike a traditional library, however, Internet search services use computers that can update the search information fast enough to keep up with constant change.

1.5 Understanding The Big Picture

Grasping all details of the Internet is impossible because the Internet continues to change. Thus, no one can know the locations of all the interesting data or the way to obtain the lowest price for an item. More important, because new applications are being invented, no one can obtain a complete description of all the services available. Finally, because individual computers and software programs differ, one cannot expect the same details to apply to all computers.

To avoid becoming overwhelmed with details, we will examine the fundamentals of the Internet. Instead of focusing on how to use a particular computer, a particular brand of software, or a particular Internet service, we will consider the basics of how the Internet works and how information services use the basic mechanisms. In essence, we will examine the capabilities and structure of the Internet.

Understanding Internet capabilities makes it much easier to use the Internet. In particular, because most "help resources" specify the details of how to accomplish a task without describing why one needs to perform the task, beginners often find the instructions difficult to follow. Knowing how the Internet works and the purpose of each service helps put the instructions in perspective.

1.6 Terminology And Technology

A complex technology, the Internet has spawned a terminology that can be daunting. This book clearly explains the Internet technology using analogies and examples. It shows how the pieces fit together, emphasizing basics instead of details. It discusses the difference between the Internet and the services that are offered, explains how the Internet has been designed to permit new services to be created, and describes what happens when you use a service.

In addition, this book defines technical terms used with computer networks and the Internet. Instead of merely providing a long list of terms, Chapters 2 through 6 present definitions in a historical perspective that shows how communication systems evolved. For example, Chapters 3 and 4 explain the difference between digital and analog information. Instead of using computer networks as an example, the chapters relate the terminology to everyday experiences.

1.7 Growth And Adaptability

Part of the mystique surrounding the Internet arises from its rapid success. While the Internet was developing, dozens of other attempts to provide the same services failed to deliver on their promise. Meanwhile, the Internet has continued to expand by adapting to change, both technical and political. We will examine why Internet technology has worked so well, and you will understand how it has adapted to accommodate change.

Another amazing part of the Internet story is its incredible growth. We will look at how the Internet continues to grow and the consequences of such growth.

1.8 The Impact Of The Internet

Perhaps the most significant aspect of the Internet is its impact on society. Once restricted to a few scientists, it quickly became universal. It reaches governments, businesses, schools, and homes worldwide. As we examine services, you will see both how the Internet has changed our lives and what we can expect in the future.

1.9 Organization Of The Book

The first section (Chapters 2 through 6) introduces communication system concepts and terminology. If you already understand the terms digital and analog, universal service, and the concept of binary data, you may choose to skim this section. The second section (Chapters 7 and 8) reviews the history of the Internet and its incredible growth. The section documents the rate at which the digital revolution occurred, and provides background that will help you appreciate the significance of the underlying design. The third section (Chapters 9 through 18) describes basic Internet technology and capabilities. It examines how Internet hardware is organized and how software provides communication. Be sure to understand this section; it provides the foundation for later chapters, and will help you ask good questions and make better decisions when salespeople offer Internet products and services. The final section describes application services currently available on the Internet. For each service, the book explains both what the service offers and how the service works.

1.10 A Personal Note

I still remember an occasion many years ago when a colleague bluntly asked me the question, "What is the Internet?" I had been involved with Internet research for many years, and understood the technology. I knew many details about the hardware and software systems that constituted the Internet, how the computers were connected, and the details of communication. I also knew most of the researchers who were working on technical improvements. What puzzled me most was that the person asking already knew basic technical details and had a copy of my textbook. What could I say?

As I contemplated the question, my colleague guessed that I misunderstood and said, "I do not want to know about computers and wires. I mean, in a larger sense, "What *is* the Internet, and what is it becoming?" Have you noticed that it is changing? Who will be using it in twenty years, and what will they do with it?"

The questions were important because they pointed out a significant shift. Early in its history, most users of the Internet were the experts who helped build it. The Internet had outgrown its research beginnings and had become a powerful tool for the general public. Now, the Internet is being used in ways — both good and bad — that the original designers had not imagined.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Before The Internet

A Gentle Introduction To Communication Systems Concepts And Terminology

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

2 Telephones Everywhere

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

Telephones Everywhere

2.1 Introduction

This chapter introduces the concept of *universal service*. It uses a familiar example to show how the assumption of universal service can affect our view of a communication service, and explains why the Internet is becoming a necessity as it becomes universal.

2.2 A Communication Service

The Internet is a communication technology. Like the landline telephone before it, the Internet makes it possible for people to communicate in new ways. To the average person living now, digital communication offers expanded opportunities, just as telephone communication did after the telephone was invented. We can learn many lessons from the story of telephone service that apply directly to the Internet.

2.3 Selling Communication

To understand how a new communication technology infiltrates society, think back to the early 1900s. Imagine yourself as someone in an average town in the U.S. who has the job of selling telephone service.

All things considered, the economic times you face are full of promise. Excitement and optimism pervade industry. After all, society is experiencing an industrial revolution. Everywhere you find that mechanization has replaced manual labor. The

steam engine has replaced water wheels and animals as a source of power; some industries are starting to use engines that run on petroleum products. Factories are producing more goods than ever before.

Of course, a telephone salesperson of a century ago would have had little or no firsthand experience using a telephone. Indeed, he or she may have had only a few demonstrations before going out to sell telephone service.

Imagine that you walk into a small company and talk to the owner about telephone service. What can you say? You could tell the owner that the company needs a telephone because it will allow customers to place orders easily. Or you could say that a telephone will allow employees to check with suppliers, order raw materials, or trace shipments that do not arrive on schedule. Maybe you would ask the owner if he or she goes out to lunch with other business owners, and point out that a luncheon could be arranged in a few seconds over a telephone. You could say that a telephone is easy to use. Or, you might take a more serious approach and point out that if fire struck the business, a telephone could be used to reach the firehouse instantly: the speed might save property or lives.

How do the owners react to your telephone sales pitch? Some are interested; many are skeptical. A few are delighted, but others are wary. Although some will think the idea has merit, many will laugh. Some want to redesign business practices, but most resist. A few want a telephone just because it is new and lends status to their establishment. Despite what they say, most owners believe that they will continue to conduct business without using a telephone.

2.4 Limited Access

Selling telephone service without having used it can be difficult. But let's make the task of selling easier. Suppose that you had grown up in a world with telephone service, and that you had used telephones all your life. Then suppose that you were transported back in time to the early 1900s and tried to share your enthusiasm about telephones. You might think that it would be easy to convince people to adopt telephone service knowing how it can be used, but you would be surprised by what you face.

The first shock you encounter when trying to sell telephone service is learning that the service at that time did not work the same way as modern telephone service. Back then, telephone service meant low-quality, local, landline service. A landline phone was attached to the wall, and could not be moved. A given phone line was shared among multiple subscribers, which meant that a subscriber had to wait if someone else was using the line. Instead of clear, high-quality sounds, telephones were noisy and the volume was low. The service was local because each town or village decided independently when to run wires, hire a switchboard operator, and establish phone service. More important, each town chose a telephone technology that met its needs and budget. As a result, although many phone systems existed, they were incompatible — running wires from one town to the next did not guarantee that the telephone systems in the two

towns could work together. From a business perspective, even if a company installed a telephone, it could not be used to order supplies from other parts of the country. You quickly discover:

> *Having an independent local telephone service in each town limits the usefulness of a telephone.*

2.5 High Cost

Another shock you encounter when trying to sell telephone service approximately a century ago is learning that even when it was available, telephone service was expensive. An average family could not afford a telephone in their home. In addition to buying the telephone itself, many telephone companies charged each subscriber the true cost of installation. The first customer on a given street had to pay for running the wires from the telephone office to the street; subsequent customers only paid for running wires down the street to their houses. As a consequence, it was often more difficult to enlist the first subscriber in a given neighborhood than to enlist additional subscribers. More important, for a large part of the population who lived in rural areas, the installation cost meant telephone service was out of the question.

After many attempts to sell telephone service to individuals fail, you would reach a key conclusion:

> *Telephone service will not be a viable business until the cost of service becomes low enough for an average family to have a phone installed.*

2.6 The Difficult Transition

In a world with only a few telephones, convincing a business to install one may seem impossible. If the business cannot use it to call suppliers in remote parts of the country and local customers do not have their own telephones, a business will have little economic justification for acquiring a phone. In fact, after thinking about the world of telephone service that we enjoy and the world of telephone service a century earlier, you realize:

> *The single most important idea behind a communication service arises from its adoption — if no one else uses the service, it is useless; if everyone else uses the service, it is a necessity.*

The transition between the two extremes is difficult. It requires businesses and individuals to invest in a new communication technology before the economic benefit is obvious. If they choose a technology that does not catch on, they lose their investment. Even if others adopt the technology, it may have insufficient subscribers to sustain the expenditure. Many people remain reticent when a new technology arrives. They wait to see what everyone else will do, hoping to minimize their financial risk. The financial decision is more difficult for a business, which must decide how many phone lines to install. If the business has too few lines, callers will receive a "busy signal";† if the business has too many, the phone lines sit idle, meaning that the business has wasted resources.

2.7 Ubiquitous Access

Why did everyone in the U.S. eventually choose to subscribe to telephone service? If you are a student of history, you know the answer: because the U.S. government decided that ubiquitous telephone service was important for the country. The governments of most other countries reached the same decision. The U.S. government helped turn American Telephone and Telegraph (AT&T) into a regulated monopoly. It mandated that telephone service be available to every home and business, and regulated rates to ensure that telephone service was affordable to the average family. It required the telephone system to reach rural areas as well as cities. More important, the government encouraged AT&T to interconnect all the local telephone services, providing a single, large system.

Because one company owned and operated much of the U.S. telephone network, many tasks were easy. For example, AT&T could specify the technical details of how the phone system in one city interconnected with the phone system in another. Having one company own the system made it easy to deploy new technology. A single company also made it easy to define a global numbering system so that a subscriber in one city could directly dial the telephone number of a subscriber in another city.

In short, the result of the government action was universal telephone service available at a price an average family could afford. Within a few decades, most businesses and a large portion of the population could be reached by telephone. Of course, universal telephone service could have occurred without government intervention; we can only speculate about what might have happened. The important point is not that the government intervened, but that popularity of telephones surged as universal service became a reality. Businesses understood that universal phone service would mean a change in business procedures. As businesses and individuals started acquiring telephones, it became apparent to everyone that telephones were important. Acquiring one became a necessity. Telephone service changed from a luxury reserved for the rich to something expected by the average family.

†Early phone systems did not have voicemail; if the phone was in use, a caller heard a tone that became known as a *busy signal*.

In the U.S., the telephone system became the communication system of choice in the twentieth century because the government mandate of universal telephone service and many adopters guaranteed that subscribing would be beneficial.

2.8 Relevance To The Internet

Like the telephone system, the Internet provides communication. Also like the telephone system, the Internet had to make the awkward transition between limited access and universal service. Initially, only a few dozen people — the scientists and engineers who created the Internet — had access. Although it contributed to Internet development, the U.S. government did not mandate universal service. Neither did the governments of most other countries. Thus, unlike the phone system, Internet growth has relied on economics. As a result, growth has proceeded in a haphazard manner. In the mid-1990s, major businesses decided that they would benefit from an Internet connection, and began to mention their web pages in general advertising. By 2000, millions of homes had slow-speed Internet access, and businesses were upgrading both their computers and Internet connections to handle the increased traffic. By 2010, higher speed Internet connections to homes had become widely available in developed countries, and smart phones provided a new way to access Internet services. Now, both wired and wireless high-speed Internet access is commonplace.

During the early days of the Internet, convincing someone that it offered exciting possibilities was like trying to sell telephone service before a universal phone system was in place. Often, people who saw Internet technology smiled politely and nodded, while thinking to themselves, "That's all very nice, but why would *I* want it?"

The answer, of course, is that as more and more people connected to the Internet, access became more valuable. As we will see, adoption of the Internet proceeded much faster than adoption of the telephone. As businesses advertised products and services available over the Internet, consumer interest rose. ISPs, including America Online (AOL), ran effective advertising campaigns that enticed users to subscribe to their services. Now, in most developed countries, children think having a smart phone and access to Internet services is an absolute necessity. They have grown up in a world where the Internet provides a universal communication service.

2.9 Summary

A principle says that the value of a communication service depends on the number of adopters. Just like the telephone system, the Internet had to grow from an early facility that connected a few dozen sites to a global communication system with billions of adopters around the world. The availability of affordable access and exciting services helped spur adoption.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

3 The World Was Once Analog

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3

The World Was Once Analog

3.1 Introduction

The Internet uses digital technology to carry many forms of information, including documents, photos, songs, audio clips, and videos. Such information has not always been stored and communicated in digital form. This chapter describes the predecessor of digital, analog, and explains how analog signals are converted to digital form. The discussion uses audio as an example.

3.2 Sound, Vibrations, And Analog Recording

Highway engineers use a simple mechanism to warn motorists to slow down. They install a series of small bumps in the roadway informally called *rumble strips*. When a car drives over the bumps, the tires vibrate. Humans feel the tire vibration and hear it as sound, which alerts drivers that they must reduce speed.

The first mechanical phonographs used the same basic idea to reproduce sound. Sound was recorded by cutting a grove in a cylinder or disc. As the groove was cut, sound vibrations caused the cutting device to vibrate, leaving small bumps in the groove. To play back the recording, a stylus traveled across the surface of the cylinder or disc. As the stylus ran across the bumps, it vibrated a flat diaphragm, producing vibrations that humans perceive as sound. The diaphragm attached to a tube that focused and directed the sound, similar to the tubes used in musical instruments, such as a trumpet. Figure 3.1 illustrates the idea.

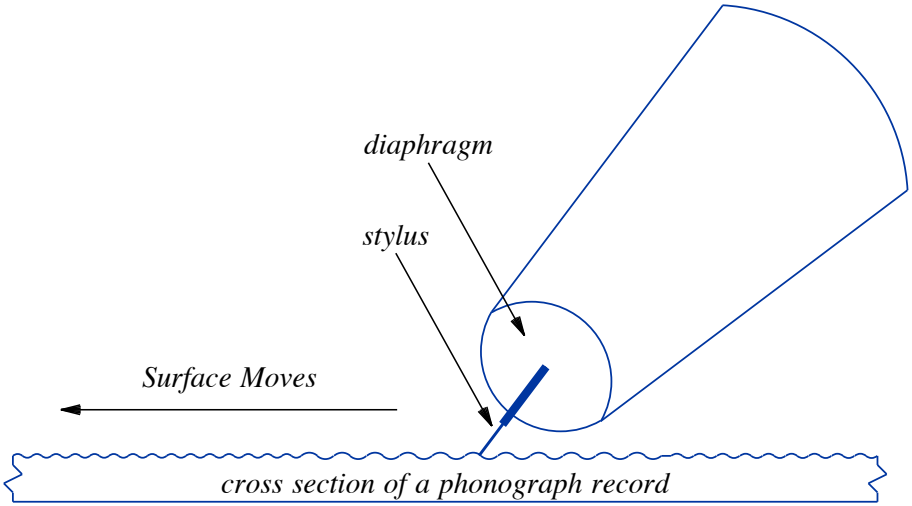


Figure 3.1 An illustration of how bumps on an early phonograph record cause a diaphragm to vibrate as the record passes under the stylus.

Devices like a phonograph are called *analog devices* because they record and play an analog of sound. That is, bumps on a phonograph recording are exactly analogous to the vibrations that make sounds. For example, the height of the bumps controls the volume. When the sound is soft, the bumps are nearly flat; when the sound is loud, the bumps are higher. If there are no bumps at all, a phonograph produces no sound.†

To summarize:

> *An analog device maintains an exact physical analog of information. For example, bumps on an early phonograph recording correspond to vibrations that we perceive as sound.*

3.3 Analog Electronic Devices

Although early phonographs were entirely mechanical, modern equipment that reproduces sound uses electronics. For example, an AM radio that broadcasts signals through the air uses analog technology. An AM radio station transmits an electromagnetic signal that varies in an exact analog of sound. When the sound is loud, a stronger signal is transmitted than when the sound is soft. In fact, analog can best be understood by thinking about an amount of one substance being proportional to another: the amount of signal is proportional to the volume of sound.

†In practice, a phonograph always produces some noise because the recording surface is not perfectly flat; it contains minor scratches that become worse each time the record is played.

When a radio receiver is tuned to the same channel as a transmitter, an electronic circuit in the receiver captures the incoming radio waves (i.e., the signal), and produces an electric current that is an exact analog of the signal. When the signal corresponds to a soft sound, the current is weak; when the signal corresponds to a loud sound, the current is stronger. Thus,

An electronic device is analog if the amount of electrical current it generates is proportional to its input.

3.4 Many Electronic Devices Are Analog

Many electronic devices still use analog technologies. For example, in addition to AM and FM radios, stereo systems, wireless microphones, televisions, and even smart phones use analog electronic circuits to provide audio for earphones and speaker systems. Early electronic devices used analog circuits for everything:

At one time, most electronic devices used analog techniques to transmit, amplify, and emit pictures or sounds.

3.5 The First Analog Communication

Analog communication was an important part of early telephone systems. The first telephones had two basic parts: a microphone to convert sound waves into an analog electrical signal and an earpiece to convert an analog electrical signal into sound waves. Whenever a person spoke into the microphone, the electrical signals carried an analog of the sound along the wire to another telephone where it was converted back into sound. Because the system used analog signals, a loud sound caused more electric current to flow than a soft sound. Figure 3.2 illustrates the basic idea.

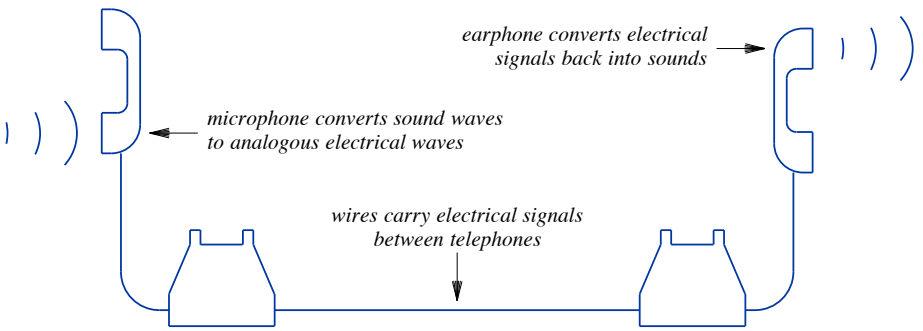


Figure 3.2 An illustration of an analog phone system in which electrical signals vary according to the sound received at a microphone.

Early telephones used an analog scheme to send voice from one place to another; the amount of electrical current sent between two telephones was proportional to the sound received by the microphone.

3.6 Sending An Analog Signal Across A Wire

Whenever an electric current passes along a wire, the signal becomes weaker. Although engineers use the term *signal loss*, physicists tell us that energy is not really lost. Instead, it is simply converted to heat. The consequence for analog electrical signals is important: as an electric signal passes along a wire and some of the energy is converted to heat, the signal becomes weaker and weaker. For example, if an electrical signal is an analog of sound, the volume of the sound will be lower after the signal passes across a long wire than it was at the start.

For an analog telephone system, the signal loss causes a problem. It means that the signal becomes weaker as it travels from one telephone to another. If the telephones are far apart, the signal will be so weak that the sound cannot be heard. In early telephone systems, the signal loss problem was so severe that telephones only worked in a small local area.

As telephone service expanded, telephone companies solved the problem of signal loss by adding *amplifiers* (i.e., devices to boost the signal) to the system. Amplifiers are currently used to boost audio at rock concerts and guitar amplifiers boost the audio from a guitar. In the analog phone system, amplifiers were placed periodically along wires, giving the signal enough additional energy to travel along wires to the next amplifier. Eventually, the signal reached its destination. Figure 3.3 illustrates the idea.

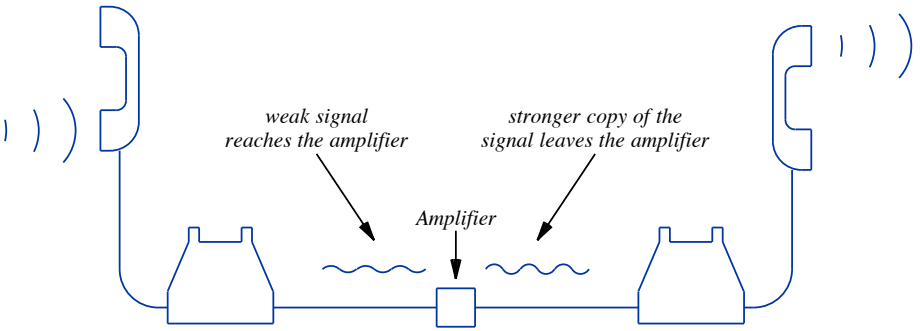


Figure 3.3 Telephone systems that use analog communication need to amplify the signal if it travels a long distance.

Unfortunately, analog electronic devices are never perfect. Each amplifier along the path between two telephones distorts the signal and adds a little noise that is amplified, along with the signal, by the next amplifier. The analog telephone system included special filters to help block distortion and noise, but doing so meant the system also blocked some legitimate sounds. The filters themselves distorted the signal because they eliminated unwanted sounds.

3.7 Analog Is Simple But Inaccurate

Analog devices are the easiest to understand because most of what we do is analog. When a human uses a muscle to open a door, the door moves in an analog of the force exerted on it. The volume of a human voice changes in exact analog to the force exerted by the person's diaphragm. Similarly, the pitch of a human voice is an exact analog of the force a person applies to stretch their vocal cords.

Although analog may be natural and easy for a human to understand, analog electrical devices have drawbacks. In general, it is impossible to produce an exact analog of all possible inputs. A microphone, for example has parts that detect vibrations and then convert the vibrations into electrical signals. Like any mechanical device, a microphone cannot capture all sounds. For example, when a performer expels a breath directly into a microphone, the microphone can become overwhelmed and miss softer sounds that occur simultaneously.

Inaccuracies also arise because amplifiers are not perfect. In general, every analog electronic device changes its input signal in some unintended ways. It may reduce the signal strength or add background noise. We call the changes *distortion*. One can hear the distortion produced by an audio amplifier by turning the volume to maximum when no input is connected. In summary:

An analog electronic device always distorts the input and adds noise.

3.8 A Definition Of Digital

We use the term "digital" to mean that numbers are used. In particular, a digital technology does not use a physical analog:

A technology is digital *if it uses numbers to store and transfer information instead of a physical analog like bumps on a record, magnetism on a tape, or electric current in a receiver.*

Unlike the analog devices discussed earlier, a computer is indeed a *digital device*. We classify computers as "digital" because they use numbers to store information, including images, sounds, and video.

A computer is digital because inside a computer, all information is represented by numbers.

For example, when the user presses a key on a computer keyboard, the keyboard sends a number to the computer. When the computer paints text or graphics on the screen, it does so using numbers.

3.9 Digital Music

Using numbers to represent sounds may seem impossible. After all, we know that sound is a sequence of vibrations of varying pitch and volume. We call them *sound waves*, and they seem to have nothing in common with numbers.

People alive in the 1980s were surprised when the music industry began selling recordings on a medium known as a *compact disc* (*CD*). Advertisements boasted that because they are digital, CDs produce better music than older analog media, such as vinyl records and magnetic tapes. When CDs arrived, most people had no idea what to expect from "digital music," or why it was supposed to be better.

Digital recording only works because computer circuits operate at much higher speeds than the vibrations a human ear can hear. At such high speed, it is possible for a computer circuit to translate analog waves into a sequence of numbers.

3.10 Recording Sound As Numbers

To understand digital recording, think of the temperature on a summer day. In the early morning it can be cool, but the temperature rises rapidly following sunrise. Around noon it peaks, and begins to fall in early evening. Suppose you wanted to re-create the exact outdoor temperatures of a summer day in a greenhouse in the fall. Let's assume the greenhouse has a heating system with a control that let's you choose the temperature.

To re-create the temperatures that occur on a given day, you must record the temperatures that day. You take a thermometer outside and record the temperature periodically (for example, every half hour). All you need to write down is a list of times and the temperature at each time (i.e., a list of numbers). Later. you can take the list of temperatures to the greenhouse. By setting the thermostat in the greenhouse every half hour to exactly match the temperature you recorded, you can make the greenhouse mimic the temperature on the summer day.

Computer circuits use the same technique when they record sound digitally. A conventional microphone generates an analog electrical signal (i.e., an electrical wave that varies exactly like the sound wave). The signal travels to a digital device that measures the incoming signal and generates a number that tells the level of the signal at that instant. Because a computer circuit operates quickly, it can take tens of thousands

of measurements per second. The set of numbers is saved and transferred to a file inside the computer.

When someone plays a digital recording, a computer reads the numbers and uses them to re-create an analog electrical signal that the numbers specify. That is, the computer reproduces the original analog signal from the numbers. Once a signal has been created, it is sent through an amplifier and then to earphones or a loudspeaker.

A simplified example will help explain. Suppose we start with a sound wave that vibrates smoothly. Figure 3.4 illustrates the wave vibrating up and down as time passes.

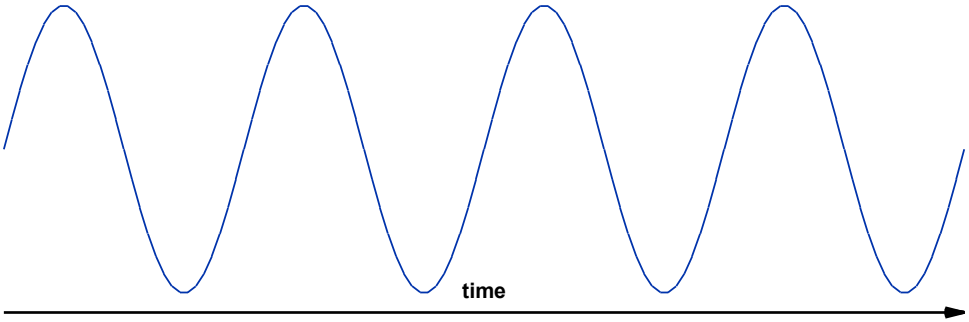


Figure 3.4 An analog signal that vibrates up and down as time passes.

To convert the wave to digital, a computer circuit measures the height of the wave periodically, and records the measurements. The measurements are known as *samples*. Figure 3.5 illustrates one possible set of samples for our wave.

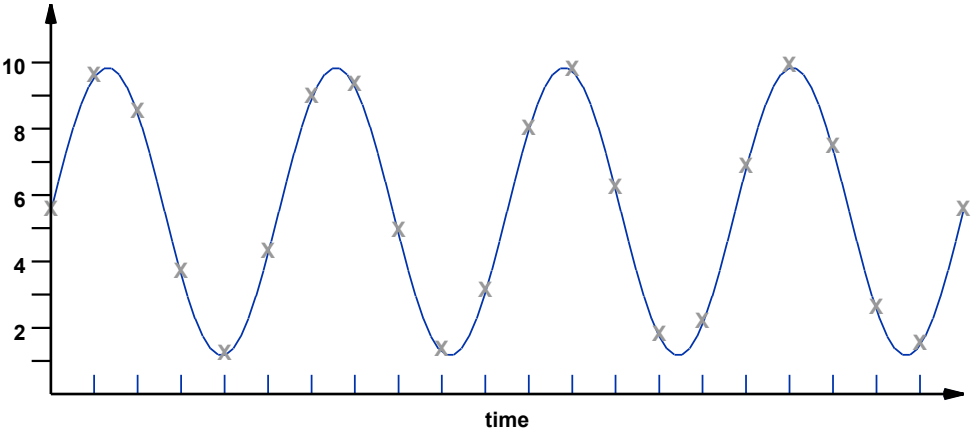


Figure 3.5 An illustration of samples taken periodically to measure the wave.

How well do our samples work? Figure 3.6 shows what happens if we use the samples to re-create a wave — the result has the same general shape as the original wave, but is not an exact match.

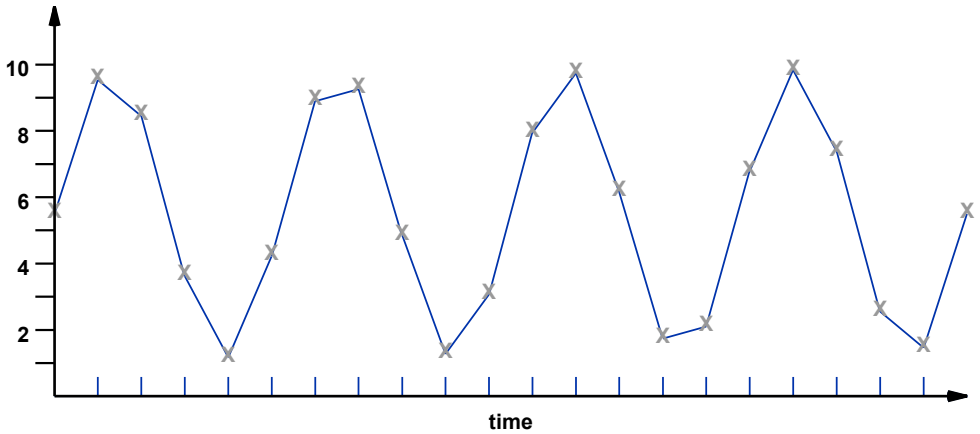


Figure 3.6 An illustration of samples taken periodically to measure the wave.

Can we do better? Yes! If we take samples more frequently, we will be able to generate a wave that is a closer match to the original. In the case of making the temperature in a greenhouse match a summer day, the idea should be clear. If we only record three samples — morning, noon, and evening — it would be impossible to know how fast the temperature rose in the morning or when it started to fall in the afternoon. If we record measurements every hour, the greenhouse temperature will be a better match to the summer day. Recording a measurement every half-hour would improve the match further. Figure 3.7 applies the idea to our example wave, and shows how taking more samples results in a closer match.

In terms of digital audio, a mathematician from Bell Labs named Nyquist worked out the details — to reproduce a sound exactly, a digital recording must sample the sound twice as fast as the fastest vibration. Engineers use Nyquist's mathematical theory when they build systems that convert sound to digital form. For example, because most humans can hear sounds up to 20,000 vibrations per second, a digital audio system must sample at least 40,000 times per second. A technology used by the music industry samples 44,100 times per second, and one used by the Hollywood movie industry samples 48,000 times per second, which means that unless your hearing is exceptional, the recordings may contain high-pitched sounds that only your dog can hear.

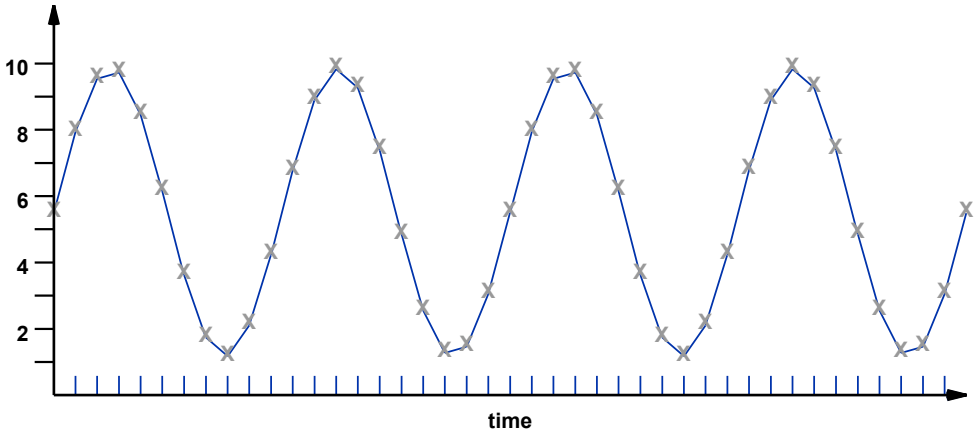


Figure 3.7 Taking samples more frequently results in a closer match to the original wave.

3.11 Converting Between Analog And Digital Forms

An electronic circuit used to convert an analog signal into a sequence of numbers is known as an *Analog-to-Digital converter*, and is often abbreviated *A-to-D converter*. An A-to-D converter samples (i.e., measures) an electrical signal periodically, and produces a sequence of numbers that specify the level of the signal when each sample was taken. To convert sounds to digital form, the sequence of values produced by an A-to-D converter are stored in a file. Figure 3.8 illustrates the conversion.

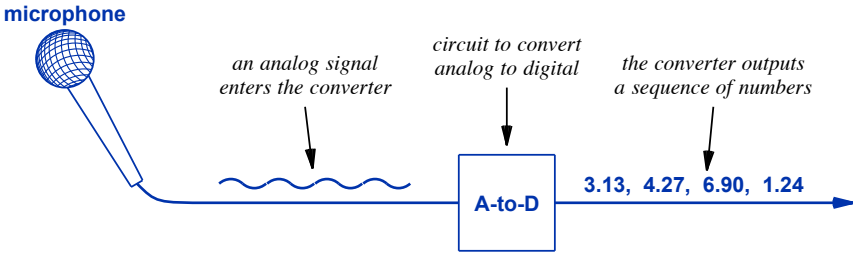


Figure 3.8 An electronic circuit performs analog-to-digital conversion by sampling an analog signal and generating a sequence of numbers that give measurements of the signal.

A computer needs another electronic circuit to reproduce sound from a stored set of numbers. The circuit is known as a *Digital-to-Analog converter* (*D-to-A converter*). Figure 3.9 illustrates how the conversion works.

To reproduce sound from a digital music file, a computer reads the sequence of numbers from a file (or a music CD), and passes the sequence to a D-to-A converter. The converter generates an analog electrical signal, which can be played through earphones or amplified and played through a loudspeaker. A computer can send numbers into the converter so quickly that our ears hear the result as continuous sound.

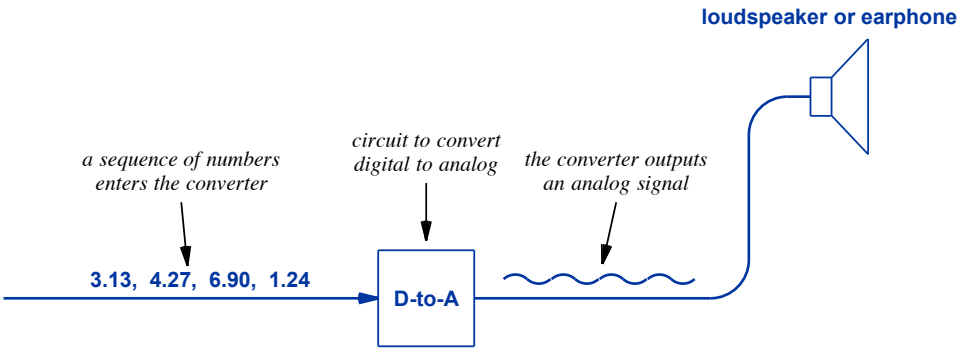


Figure 3.9 An electronic circuit performs digital-to-analog conversion by taking a sequence of numbers as input and generating an analog signal that matches the numbers.

3.12 Why Did Digital Music Take Over?

What makes digital music more valuable than its analog predecessors? First, unlike a vinyl record or magnetic tape, a digital recording does not "wear down" as it is played. Only numbers are stored, and the numbers remain unchanged as the song is played. So, exactly the same sounds will be generated the one-millionth time the song is played as the first time the song is played. Second, unlike analog media that always has background noise, a digital recording does not inject extra noise. For example, to leave a silent gap between songs, a digital recording can contain zeros, meaning that no electrical signal will be produced. Third, digital media stores much more music in a given space than an analog storage system. For example, when magnetic tape was used, a reel of tape seven inches in diameter and approximately one-half inch thick could only store ninety minutes of music. A digital device of the same size can store hundreds of hours of music. The small storage requirements allowed the music industry to develop small, portable music players.

3.13 Summary

Sounds are vibrations. The most natural representation of audio information is an analog form in which the amount of a physical quantity varies in exact proportion to the sound. For example, the height of bumps on a phonograph record and the amount of electrical current generated by a radio receiver each correspond to the loudness of the recorded sound.

Audio information can be represented using a digital form, which means using a sequence of numbers to represent the sounds. A digital representation is the most convenient form for computers and smart phones because they represent everything as numbers.

It will be important to keep the following ideas in mind throughout the remainder of the book:

- Information, including text, photos, audio and video, can be encoded in digital form (i.e., as a set of numbers).
- Every device that can record digital audio (e.g., a laptop and a smart phone) contains an A-to-D circuit that converts the analog signal from a microphone into a sequence of numbers.
- Every device that can play digital audio (e.g., a digital TV, and a portable music player) contains a D-to-A circuit that converts a previously-recorded sequence of numbers back into an analog signal.
- A key advantage of using a digital representation arises because the information does not become distorted while being stored, read, copied, or communicated over a computer network.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

4

The Past And Present Digital Network

4.1 Introduction

This chapter discusses the concept of digital communication, and shows how digital information can be encoded for transmission using only two basic symbols.

4.2 The World Was Previously Digital

Chapter 3 asserts that, at one time, most of the electronic devices in the world were analog. Indeed, they were. Before digital music appeared, AM and FM radios, stereos, telephones, and televisions all used analog electronic circuits. Surprisingly, the earliest electronic communication systems were not analog. Decades before the first telephones made analog communication popular, the world used digital communication!

In the 1850s, it was possible to send a message from one town to another in a matter of minutes. The technology was known as a *telegraph*, and became so popular that telegraph lines spread quickly across the United States and other countries.

A telegraph operates on the same principle as a wall switch that controls an electric lamp. The switch is located at a convenient height on the wall, remote from both the source of power and the lamp itself. A pair of wires that reach the switch carry power to the switch and current back to the lamp. When the switch is in the "off" position, the circuit is open and no current flows to the lamp. When the switch is in the "on" position, the circuit is complete and current flows to the lamp.

The basic telegraph also used a switch with wires running to it. In the telegraph, however, distances were much longer: the switch was located in one town, and the device it operated was located in another. In addition, a telegraph did not use light. Instead, a telegraph used a small electrically operated device that made an audible click when it received an electric signal. To send a message across the telegraph, a person in one town moved a switch back and forth, while a person in another town listened to the clicks generated by the telegraph device.

To an untrained person, a telegraph sounded like an unending series of clicks with no perceptible pattern. Some of the clicks had a short duration (i.e., the switch was held down a very short time), while other clicks were longer. Sometimes, a sequence of extremely short, rapid clicks occurred before the mixture of short and long clicks began again. A trained telegraph operator could distinguish individual letters among the clicks, and could transcribe the message onto paper as fast as it arrived.

4.3 A Telegraph Was Digital

A telegraph was a digital device because instead of sending a continuous signal that is an exact analog of the input, the telegraph used clicks to send the individual characters of a message. Although telegraph clicks may seem unrelated to numbers, to mathematicians, the set of clicks represent digits of a number system. Specifically, a mathematician thinks of the two types of clicks used in a telegraph transmission as the digits 0 and 1. We can now define *digital* more precisely: any device that uses a fixed set of discrete values to represent information is digital. To summarize:

The telegraph was a digital technology because it used discrete clicks to transfer information instead of a continuously varying signal.

4.4 Morse Code

Samuel Morse invented a code that became popular among telegraph operators; we use the term *Morse code* to refer to Morse's system. Morse code is simply a way to represent letters and words using a series of clicks and pauses. For example, Morse code uses one short click followed by one long click to represent the letter *A*.

When assigning code values, Morse tried to use short sequences for letters that occurred frequently. The result is that one can send a message faster using Morse's code than codes that are not planned as carefully. For example, in common English text, the letter *E* occurs more frequently than any other letter. Morse code uses a single, short click to encode *E*.

Short clicks are called *dots*, and long clicks are called *dashes*. During transmission, a short pause occurs after the dots and dashes that constitute a single letter, and a longer pause occurs after each word. A trained operator uses the pauses to detect when each letter and word ends.

4.5 Letters And Digits In Morse Code

In addition to codes for all the letters, Morse specified codes for the digits zero through nine and a few punctuation symbols. Figure 4.1 shows a version of Morse's code that was adopted for international use.

A	• –	X	– • • –
B	– • • •	Y	– • – –
C	– • – •	Z	– – • •
D	– • •	0	– – – – –
E	•	1	• – – – –
F	• • – •	2	• • – – –
G	– – •	3	• • • – –
H	• • • •	4	• • • • –
I	• •	5	• • • • •
J	• – – –	6	– • • • •
K	– • –	7	– – • • •
L	• – • •	8	– – – • •
M	– –	9	– – – – •
N	– •	,	– – • • – –
O	– – –	.	• – • – • –
P	• – – •	?	• • – – • •
Q	– – • –	;	– • – • – •
R	• – •	:	– – – • • •
S	• • •	'	• – – – – •
T	–	–	– • • • • –
U	• • –	/	– • • – •
V	• • • –	(	– • – – •
W	• – –	)	– • – – • –

Figure 4.1 Examples of Morse code, which uses a unique sequence of dots and dashes to represent each letter, digit, and punctuation symbol.

Morse did not assign a code to all possible symbols. For example, although extensions were added later, Morse did not define codes for the dollar sign or for the percent sign, even though such characters do occur in written text. Furthermore, Morse did not attempt to include letters and symbols used in languages other than English.

4.6 Telegraph Users Did Not Encounter Morse Code

Although all messages passed across a telegraph in Morse code, only telegraph operators needed to know it. A person who wanted to send a telegram wrote the message on a piece of paper and handed it to an operator. The message itself could contain any sequence of letters and numbers. In fact, because the cost of sending a telegram depended on the length of the message, people often invented abbreviations, similar to the abbreviations used in text messages.

A skilled telegraph operator could translate between Morse code and text quickly; two operators were required for transmission across a telegraph system. At the sending end, the operator read a message from paper and tapped out Morse code. At the receiving end, the operator listened to the Morse code and wrote the text. After the message was received, it was delivered to the intended recipient.

Three ideas from the telegraph are relevant to the Internet:

- It is possible to encode all letters and digits using only two basic code values: dot and dash.
- A code used for message transmission defines a basic alphabet of characters that can be sent; the code can be useful even if it does not include all possible characters.
- A customer of a telegraph service never encountered or understood the underlying encoding scheme.

4.7 Virtually Instant Communication

When the telegraph was invented, it seemed like magic. Until then, sending a message to a remote location meant using a human courier, usually on horseback. Suddenly, the world changed, and it became possible to learn about events as they occurred. With a telegraph, for example, people located far away from a financial market could learn about current stock prices and could send orders to buy or sell stock. People far from the location where ballots were counted could learn the results of an election immediately. Travelers could stay in touch with friends or family at home.

4.8 Speed Is Relative

Although the telegraph changed the world because it was much faster than a courier on horseback, we would think of communication by telegraph as relatively slow. Imagine communicating with a friend via telegraph instead of texting. After writing a message, you must hand it to a telegraph operator and wait while the operator translates it into Morse code. Only the best operators can send more than a dozen words per

minute. Furthermore, both the sending and receiving operators must be equally adept for a transfer to succeed. If the receiver misses a character or word, he or she must ask the sender to transmit it again. As a result, holding a dialogue via telegraph was more inconvenient and much slower than texting.

It should be obvious why the telephone caused so much excitement when it appeared. Instead of writing a message and passing it to an operator, a person on one end of a telephone call can speak directly to a person on the other end. The telephone system carries the speaker's voice to the other end immediately, and conveys something that cannot easily be expressed in written form: emotions. Hearing a voice makes it possible to identify an individual and to distinguish anger from humor or reticence from excitement.

Users who could afford it, switched from telegraph to telephone communication quickly. Many engineers who worked on communication systems abandoned efforts to improve the slow, digital telegraph, and spent their time working on analog technology for telephones.

4.9 The Telephone Eventually Became Digital

Although voice communication may seem inherently analog, many modern telephone systems use digital encoding for voice transfer. At one end, the system converts an analog voice signal into a series of numbers exactly as described in Chapter 3. Computers transfer the numbers across the phone network, where they are converted back into an analog signal and played for the user. In Chapter 26, we will learn about VoIP technology used to transfer digital phone calls over the Internet.

Using digital technology to carry voice has a significant advantage for a telephone company. To understand why, consider a phone call. Recall that an analog system needs amplifiers to boost the signal and that each amplifier injects a little noise that is amplified along with the phone call audio. As a result, audio quality deteriorates as the signal passes across an analog phone network. By contrast, a digital call does not deteriorate — once the audio has been converted to a sequence of numbers, the numbers are sent to the other end, and the original audio can be re-created. Because digital calls provide much higher quality, modern telephone systems are almost entirely digital.

4.10 Relevance To The Internet

Like the early telegraph, the Internet provides digital communication. Because computers store information in digital form, digital communication works well in a computer network. When the information moves from one computer to another, a digital mechanism saves time and effort.

4.11 Binary Encoding Of Data On The Internet

The Internet is like a telegraph in another way: it uses exactly two values to encode all data items. While the values used in Morse code are commonly called dot and dash, we usually think of the values used in the Internet as zero and one, the two "digits" of the binary number system. In the Internet, as in most computer systems, the values are known as *bits*; the term *bit* is an abbreviation for *binary digit*. The next chapter explains the modern equivalent of Morse code used on the Internet by describing the sequences of bits used to represent individual letters and digits.

4.12 Why Use Two Symbols?

Using two symbols for digital encoding is not limited to the Internet — all digital electronic devices use bits to encode data. To understand why, think of communicating through a light fog. Suppose you have a lamp and want to signal a friend. It would be easiest for another person to tell whether the lamp is "on" or "off"; it would be much more difficult for the person to distinguish among "off," "dim," "medium," and "bright." Similarly, electronic circuits can sense "off" and "on" much easier than they can distinguish multiple levels of an electrical signal. In addition to eliminating ambiguity, limiting digital systems to two values allows engineers to build circuits that operate faster.

The Internet is like a telegraph in another way: even though the Internet encodes data in bits, the details are completely hidden from users. Like a person who sent or received a telegraph message, someone who uses the Internet never sees and never needs to understand the binary encoding that is used. To summarize:

> *Although the Internet uses a binary encoding for all data transferred, users usually remain completely unaware of the encoding.*

4.13 Summary

The Internet is similar to its early predecessor, the telegraph, in three ways. First, the Internet provides a digital communication service. It allows one to transfer a set of numbers from one computer to another. Numbers stored in a computer can be used to encode almost any information including the letters in a document, sounds, or pictures. Second, like the telegraph, at the lowest level, the Internet encodes all data using two values. The Internet uses zero and one, the two binary digits. Third, the Internet hides the details of data encoding from the user, allowing the user to send text, photos, and digital audio without knowing any details.

Chapter Contents

5 Basic Communication

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Basic Communication

5.1 Introduction

Computer networks interconnect computers and electronic devices, such as smart phones, that contain computers so they can exchange data.† Although modern computer networks are complex combinations of hardware and software, early computer networks were much less sophisticated.

This chapter outlines the development of basic communication technologies, and shows how networks evolved. It introduces terminology, and explains how modems work. We will see that modems provide Internet service over a cable system, phone line, and even wirelessly to connect cell phones. The concepts defined here provide the foundation for network technologies described in successive chapters.

5.2 Communication Using Electricity

Since the discovery of electricity, inventors, scientists, and engineers have worked on ways to use electrical signals for communication. The principles discovered have resulted in fast, reliable communication systems. Our knowledge of digital communication can be divided into roughly three historical stages. The first stage focused on the properties of signals. The second stage focused on how to use signals to send bits and how to organize the bits into characters. The third stage focused on how to detect and correct errors that occur during transmission.

†Throughout this text, we will use the term *computer* to refer to devices, large or small, fixed or portable that contain computer circuits.

5.3 Sending Signals

Researchers first studied how electromagnetic signals propagate. They learned, for example, that electrical signals lose energy as they travel from the source. That is the reason modern networks limit the length of interconnecting wires or require the use of electronic devices to amplify signals after a certain distance. They learned that whenever an electrical signal passes across a wire, electromagnetic energy is emitted, much like a miniature radio transmitter. Finally, they learned that electromagnetic energy in the environment can interfere with other signals. In particular, a wire acts like a radio receiver, and incoming electromagnetic energy can interfere with the signals that are passing across the wire. To prevent such interferences, cable TV connections use a special coaxial cable that encloses wires in a metal shield. The radiation generated during a thunderstorm may be so strong that it interferes with signals despite the shielding around the wires. Later, we will discuss how the Internet handles such interference.

5.4 Using Signals To Send Information

Once they understood the physics of sending electromagnetic signals across wires, scientists and engineers studied ways to use electrical signals to carry information. Much of the pioneering work focused on finding ways to transmit a human voice across telephone lines with minimum distortion, but the techniques that were discovered apply to communication in general.

A key idea emerged from early research that is used throughout the Internet: *modulation*. The idea of modulation derives from basic physics: an electrical signal that oscillates back and forth regularly travels much farther than an electrical signal that is merely on or off. For example, an oscillating signal is used to send data from your device to a service provider. We call the oscillating signal a *carrier*. To send information, the sender changes the carrier signal slightly. We say that the carrier wave has been *modulated*. The receiver detects the changes, and reconstructs the information that was sent.

Conceptually, two electronic circuits are needed to convey information over a carrier wave. The sender must have a circuit called a *modulator*. The modulator starts with a carrier wave and uses the information to be communicated to change the carrier slightly before sending it. At the receiving side, an electronic circuit known as a *demodulator* performs the reverse function, which is known as *demodulation*. By measuring how much the incoming signal deviates from a perfect carrier, a demodulator can recover the information that was sent.

Modulation is not new. AM and FM broadcast radio stations use modulation to send audio. In fact, the *M* in AM and FM stands for *modulation*. The *A* and *F* specify the type of modulation, *amplitude* or *frequency*. When a user tunes a broadcast receiver to a station, the user selects a particular carrier wave. The receiver detects changes in the incoming wave, and plays the changes, which a user hears as audio. Broadcast television stations use modulation as well, and encode a combination of audio and

video. The modulation used with computer networks differs in two ways. First, instead of analog audio or video, the modulation encodes digital data. Second, the modulation is substantially more sophisticated than the modulation used with broadcast radio and television. Nonetheless, it relies on the same underlying principle.

5.5 Modem: A Modulator And A Demodulator Combined

Modulation technology is used throughout the Internet — when computerized devices communicate across a long cable, they need a modulator at one end and a demodulator at the other. Modulation technology is also used for wireless Internet communication (Wi-Fi as well as 3G, 4G, and 5G cellular systems).

Internet communication differs from broadcast radio in a significant way: Internet communication is two-way. For broadcast radio, only the radio station needs a modulator because only the radio station transmits. To listen to a broadcast, a user's radio receiver needs a demodulator that detects changes in the carrier. However, a receiver does not need a modulator because it does not transmit information.

To handle two-way communication, engineers invented an electronic device known as a *modem* (an abbreviation for *mo*dulator/*dem*odulator). In fact, a modem has two independent electronic circuits inside: a modulator used to transmit outgoing data and a demodulator used to decode incoming data. Figure 5.1 illustrates the concept.†

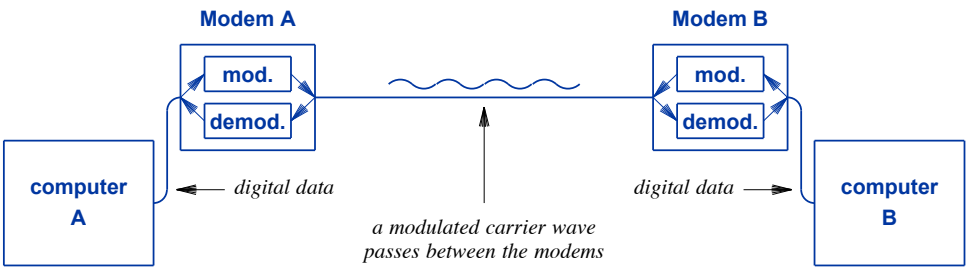


Figure 5.1 Illustration of modems used to send data across a long distance. Each side contains a modulator to send data and a demodulator to receive data.

In the figure, when computer A sends to computer B, the modulator in modem A is used, and demodulator in modem B is used. When computer B sends, the modulator in modem B is used, and the demodulator in modem A is used.

In the early days of the Internet before high-speed Internet service was available using cable modems and DSL (Digital Subscriber Line), most Internet users relied on dial-up modems. Dial-up modems allow digital data to be communicated over an ordinary dial-up telephone connection. Instead of an electromagnetic carrier wave, a dial-up modem uses an audible tone as the carrier, and modulates (changes) the tone to send

†Although the figure shows a modem as physically separate from a device, some modems are inside a device (e.g., inside a smart phone).

data. For various technical reasons, a dial-up modem cannot transfer data quickly. A DSL connection can transfer data over 30 times faster than dialup, and some cable modems can transfer data over 700 times faster than dial-up. Thus, once such technologies became available, most users abandoned dial-up modems.

5.6 How Modems Allow Two-Way Traffic

Each modem contains both a modulator and a demodulator, which allows data to be sent in either direction. Some modems are arranged to take turns transmitting — first the modem on one ends sends data, then the modem on the other end sends data. Other modems use an interesting technique that permits data to be sent in both directions simultaneously: each side uses a different carrier. You can think of the technique as having separate "channels" for each direction. Using separate channels means data moving in one direction does not interfere with data moving in the other direction.

To summarize:

> *A modem is a device that transfers digital data over a carrier wave, either on a wire or wireless network. A pair of modems can transfer data in both directions because each modem contains a modulator used to encode data in an outgoing signal and a demodulator used to extract data from an incoming signal.*

5.7 A Character Code For Digital Information

As they studied voice transmission, researchers also considered transmission of digital information. They found ways to encode digital values in an electrical signal (e.g., using a positive voltage to encode *1* and a negative voltage to encode *0*). In addition, they devised a sequence of bits (zeros and ones) to represent each letter and digit.

Although the character codes used on modern computer networks use two basic values, they differ from Morse code because each character is assigned a code with the same number of bits. For example, Morse code uses a single dot for the letter *E* and three dots for the letter *S*. By contrast, many modern character codes assign a sequence of seven bits (0's and 1's) to each letter. Having a uniform number of bits for all characters makes character processing faster and the hardware less expensive. It also simplifies character storage because each character occupies a constant number of bits regardless of the specific character.

The *American Standard Code for Information Interchange* (*ASCII*) is among the most popular and widespread character codes used throughout the computer and network industry. ASCII defines a bit sequence for most characters used in English: upper- and lowercase letters, digits, punctuation, and a few miscellaneous symbols such as the mathematical symbols for equal, plus, and minus.

The details of the ASCII encoding are unimportant because most people who use computers or networks never see the encodings. However, the examples shown in Figure 5.2 will help clarify the idea. ASCII uses the 7-bit sequence 1000101 to represent the letter "E," the sequence 1010011 to represent the letter "S," and the sequence 0101100 to represent a comma.

To summarize:

When it sends textual data across a network, a device encodes the data in zeros and ones. For example, the ASCII code assigns a 7-bit code to each letter and digit. Most users never see the encoding because it is an internal detail that remains hidden.

Why is a standard encoding important? Anyone who creates a communication system quickly realizes that both sides must agree on the exact form and meaning of messages. Having a standard means that one doesn't have to choose an encoding for each correspondent. For example, if everyone agrees on the meaning of bits, it is possible to send an email message to many recipients and be sure they will all see the same characters displayed.

A 1000001	S 1010011	a 1100001	s 1110011
B 1000010	T 1010100	b 1100010	t 1110100
C 1000011	U 1010101	c 1100011	u 1110101
D 1000100	V 1010110	d 1100100	v 1110110
E 1000101	W 1010111	e 1100101	w 1110111
F 1000110	X 1011000	f 1100110	x 1111000
G 1000111	Y 1011001	g 1100111	y 1111001
H 1001000	Z 1011010	h 1101000	z 1111010
I 1001001	0 0110000	i 1101001	. 0101110
J 1001010	1 0110001	j 1101010	, 0101100
K 1001011	2 0110010	k 1101011	? 0111111
L 1001100	3 0110011	l 1101100	(0101000
M 1001101	4 0110100	m 1101101) 0101001	
N 1001110	5 0110101	n 1101110	/ 0101111
O 1001111	6 0110110	o 1101111	& 0100110
P 1010000	7 0110111	p 1110000	+ 0101011
Q 1010001	8 0111000	q 1110001	- 0101101
R 1010010	9 0111001	r 1110010	= 0111101

Figure 5.2 Examples of the ASCII encoding. The upper- and lowercase letters each have a binary code. Users do not see the encoding.

5.8 Bits And Bytes

Inside a digital device, all values are stored in bits. We will learn that network performance is measured in the number of bits a network can deliver per second. However, data is usually measured in *bytes*. What is a byte, and how does it relate to bits? A byte is a group of eight bits; each byte can store one English character.† You can remember that a byte is larger than a bit because the word *byte* contains more letters than the word *bit*. You can remember that a byte is eight bits because the last letter of byte is "e."

Data sizes are measured in bytes; each byte contains eight bits.

ASCII uses seven bits. Why use eight bits per byte instead of seven? Or why not use ten? Humans think in decimal because they have ten fingers. Computers use binary because it provides a more convenient way to build digital circuits. Humans say that 10, 100, 1000, and so on are "round numbers" because they have trailing zeros. In binary (base 2), the round numbers are 2, 4, 8, 16, 32 because when written in binary they have trailing zeros. For example, in binary, 8 is 1000 and 32 is 100000. Consequently, eight bits was chosen instead of seven or ten because eight makes sense in binary.

5.9 Detecting Errors

Much of the early work on digital communication focused on error detection and correction. Researchers studied the errors that occur when sending electrical signals across copper wires or broadcasting wireless signals, and found ways to detect the errors. For example, they knew that natural phenomena like lightning can cause random electrical signals to appear on wires and interfere with wireless transmissions. They also found that electrical signals can become distorted when they pass through a strong magnetic field (e.g., when a network passes near the electric motor in a household appliance).

When electric or magnetic interference disrupts signals, data can be damaged or lost. For example, if voltage is used to represent a bit, a bolt of lightning that strikes near a wire can cause the voltage to change even if lightning does not hit the wire directly. The point to remember is:

When using signals to communicate digital information, electrical or magnetic interference can cause the value of one or more bits to be changed.

---

†The character codes used with other languages require multiple bytes. For example, *Unicode* requires two bytes per character.

To guard against corruption of information caused by random electrical noise, researchers devised mechanisms to detect and correct the problem. The basic idea is straightforward: when sending a message, include additional information that can be used to verify that the message arrived intact.

The idea of adding extra information sounds appealing, but two questions arise:

- What extra data should be sent?
- How large is the extra data?

It may seem, for example, that the best solution involves sending an extra copy of all the data. However, doing so would reduce the network performance substantially. Fortunately, researchers devised a clever way to solve the problem without a significant change in network performance: use a mathematical formula.

To use the scheme, a sender treats an entire message as a sequence of numbers. That is, the sender uses the underlying binary values, and instead of treating them as letters and punctuation marks, processes the integer value. The numbers are fed into a mathematical formula, which produces a single integer value as a result. The value is sent along with the message. On the receiving side, the receiver uses exactly the same formula to compute an integer from the message that arrives. If the value the receiver computes differs from the value that was sent along with the message, one or more of the bits in the message must have been damaged (i.e., changed) during transmission.

The interesting aspect of the error detection scheme arises from its ability to detect an error with little extra data. If the mathematical formula is chosen carefully, the scheme has an extremely high probability of detecting errors, even though only one extra integer value is sent.

The point is:

> *To handle errors that occur when lightning or other electromagnetic interference damages bits during transmission, a small amount of extra data is sent along with each message. The data is an integer computed from a mathematical formula. A careful choice of formula makes the probability of detecting errors high even though the extra data only consists of a small integer.*

5.10 Summary

Researchers have studied the properties of electrical signals, and have learned how to use signals to encode information, including both analog audio and digital information (i.e., bits). They devised codes that assign each character a unique string of bits. In particular, they devised the ASCII code that is used to send text over the Internet.

Transmission involves an oscillating signal known as a carrier wave. To send information, an electronic circuit known as a modulator changes the carrier wave slightly. A receiver extracts the information by measuring how the incoming carrier wave deviates from a perfect carrier.

A device known as a modem is used to send data a long distance across a wired or wireless network. A modem can provide two-way communication because it includes both a modulation circuit for outgoing data and a demodulation circuit for incoming data. Modems are used throughout the Internet, including wired connections to ISPs over cable and DSL, as well as wireless communication, including communication with cell phones.

Researchers also studied transmission errors and found mechanisms that hardware can use to detect when interference has damaged bits during transfer. Although adding a small amount of extra information can help detect errors, it does not solve the problem completely. The Internet uses more powerful error detection techniques, which we will discuss later.

EXERCISES

5.1 Find out if you have a modem at home: if you have cable or DSL Internet service, look on the bottom of the device that connects your computer to the cable or phone, and see if it is labeled "modem."

5.2 When you use a wireless network (e.g., when you connect to Wi-Fi or the cellular phone network), where is the modem?

5.3 What is the name of the technique a radio station uses to send audio and a modem uses to send data?

5.4 During a severe lightning storm, the picture on Bob's TV became scrambled and then returned to normal. Explain why the picture was temporarily scrambled.

Chapter Contents

6 Local Area Networks

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

6

Local Area Networks

6.1 Introduction

Motivated by the need for better telephone communication, much of the early work on communication focused on ways to span large geographic distances. In the late 1960s and early 1970s, new networking technologies emerged that had a more immediate impact on the average person. This chapter examines the new technologies, and describes how they changed the economics of computer communication.

6.2 The Digital Revolution

The digital world became possible when scientists at Bell Laboratories invented a solid-state switch called a *transistor*. The digital revolution began a short time later when scientists and engineers devised ways to combine transistors in an *integrated circuit*† built out of silicon crystals.

An integrated circuit consists of many electronic components interconnected together, all built on a square a few tenths of an inch per side. Through intensive research, manufacturers have found ways to reduce the size of transistors and to make integrated circuits more sophisticated. Currently, a manufacturer can create an integrated circuit that contains billions of transistors.

The importance of integrated circuits lies in their economy. Because integrated circuits can be manufactured in mass quantities at low cost, it became possible to mass produce complex circuits that were too expensive to build using individual components.

†An integrated circuit is informally called a *chip*.

Many integrated circuits are designed primarily for use in computers and related devices, such as smart phones. For example, a microprocessor is an integrated circuit that forms the heart of a modern digital device — it contains all the electronic circuitry needed to add, subtract, multiply, divide, or compare numbers. In addition, a microprocessor can fetch numbers from a computer's memory or store results into memory. Low-cost integrated circuits are particularly relevant to the Internet because they form the basis for communication hardware that is used in the Internet and in user's devices that connect to the Internet.

Although low-cost integrated circuits made Internet hardware possible, another trend provided motivation for computer communication. The next sections describe the trend.

6.3 The Move To Multiple Computers

In the 1960s, a computer was a large, bulky behemoth called a *mainframe* that cost several million dollars. By 1970, advances in electronics resulted in smaller, less expensive computers known as *minicomputers*. The low cost of minicomputers changed computing. When each computer cost more than a million dollars, most organizations could only aspire to have a single computer. As computers became inexpensive, however, it became obvious that each organization could benefit from having several computers. Instead of one mainframe computer serving an entire company, each department could afford its own minicomputer. People also began to understand that computers could help with many of the tasks in an ordinary office.

6.4 Removable Media And Manual Transfer

Having multiple computers in a given organization introduced a new problem: data was no longer centralized. If the accounting department needed data that was on a computer in the payroll department, a copy of the data had to be transferred from one computer to another. Data transfers involved *removable media* storage devices, usually magnetic tapes or disks. Although the early media were physically large and bulky, they worked like modern flash drives. An operator placed a blank disk into a disk drive attached to the payroll computer, and used a program to write the needed data on the disk. Once the copy was complete, the operator removed the disk, carried it to the accounting computer, and inserted it into the disk drive. Finally, the operator used a program to load the data onto the accounting computer.†

It quickly became obvious that an organization with multiple computers could benefit from a mechanism to move data among them without requiring a human to carry a device from one to the other. Engineers and computer scientists began to investigate ways to build networks that connected computers and allowed them to share information.

---

†Decades later, after computer networks became common, networking professionals jokingly referred to the manual transfer of information as "sneaker net," implying that operators wore tennis shoes to help hasten transfers.

6.5 Early Computers Used Circuit Boards

To understand how computer networks were formed, one must understand the basics of how computers are built. Inside a computer, electronic components reside on thin, flat rectangular boards called *printed circuit boards* or simple *circuit boards*. A circuit board contains both electronic components and the wires that connect them. Computers had a main circuit board known as a *motherboard*. To make customization possible, a computer also had a set of sockets where the owner could plug in additional circuit boards, which were known as *daughterboards*.

The first type of hardware that engineers built for local communication consisted of a pair of daughterboards connected by a cable. One of the circuit boards was plugged into a computer, and the other was plugged into a second computer. Figure 6.1 illustrates the approach.

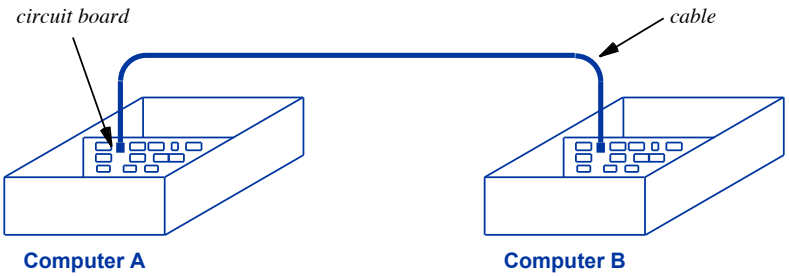


Figure 6.1 Illustration of an early computer communication system formed using two daughterboards plugged into sockets in two computers.

Once circuit boards were plugged into the two computers and connected by a cable, the computers could use them to transfer data. All transfers were controlled by software. On the sending side, software told the circuit board what data to send and when to send it. On the receiving side, software told the circuit board where to store a copy of the data that arrives over the cable.

The chief advantage of a dedicated connection from one computer to another was speed — only two computers used a given cable, so data could be transferred at any time. The chief disadvantages of dedicated connections arise from inconvenience and cost. The technology was inconvenient because installing circuit boards was tedious. It was expensive because a new pair of circuit boards had to be installed for each pair of computers. For example, Figure 6.2 illustrates the additional two boards needed to add a connection between Computers B and C.

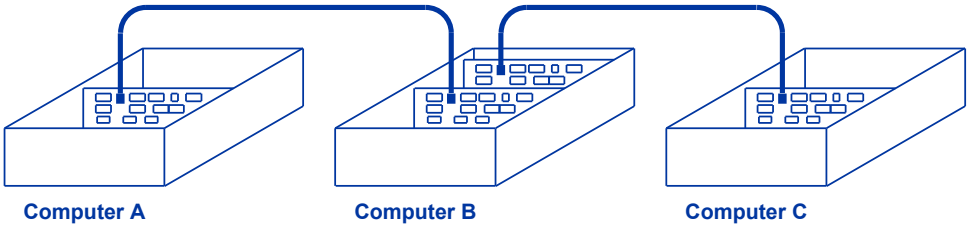


Figure 6.2 Illustration of Computer B connected to both A and C. A pair of circuit boards must be installed for each new connection; another pair will be needed to Computer A to communicate with Computer C.

6.6 LANs

A set of new technologies emerged called *Local Area Networks* (*LANs*). As the name implied, a LAN is intended for use in a small area. Unlike the telephone system that spanned a large geographic distance (e.g., a continent), a LAN is designed to connect computers in a room or in a small building.

Why limit the distance a network can span to a few hundred feet? The answer lies in economics. Building equipment for long-distance communication is expensive. In addition to the cost of running wires from one city to another, specialized hardware is needed. For example, in addition to modems (described in Chapter 5), amplifiers must be used to boost a signal as it travels over a long distance.

A LAN design does not face the same challenges as long-distance communication systems. Even a small computer can generate a sufficiently strong signal to reach across a room or down an office hallway. More important, inexpensive electronic circuits can be used.

6.7 The LAN Approach

LAN technologies solve the problem of computer communication in a way that is convenient, inexpensive, and reliable. Instead of connecting one computer directly to another, LAN technologies use hardware that can interconnect multiple computers. The network hardware exists independent of the computers themselves. Instead of one circuit board for each computer connection, a computer only needs one circuit board that allows it to communicate with the LAN. Furthermore, a computer can be added to a LAN at any time, without requiring new circuit boards to be installed in other computers. Figure 6.3 illustrates the LAN approach.

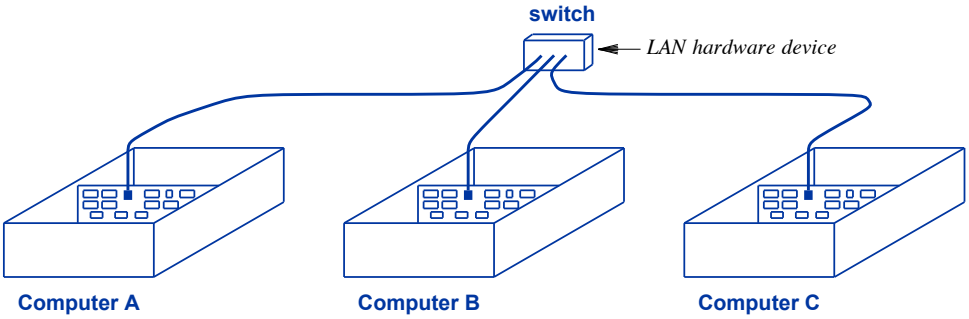


Figure 6.3 Illustration of a LAN. Each computer attaches to a switch with a cable; all computers can then communicate.

6.8 LAN Hardware

As Figure 6.3 shows, the electronic device that forms the "center" of a LAN is known as a *switch*. We use the term *port* to refer to the socket on a switch to which a cable can attach. When a vendor sells a switch that can attach up to four computers, the vendor calls it a *4-port switch*, and a switch that can attach up to sixteen computers is called a *16-port switch*. A port on a switch is really a specialized socket, and the cable that attaches a computer to a switch has a plug that fits into the socket. So, attaching a computer to a switch merely means plugging in a cable.

The figure uses a small box to illustrate a switch. In reality, a 4-port switch is only a few inches wide and sells for less than thirty dollars. Larger switches are available that can connect more computers. The largest switches, which are used to interconnect computers in an enterprise, stand several feet tall and offer hundreds of ports.

Interestingly, switch vendors take a modular approach to building large switches. First, they create a switch with a moderate number of ports (e.g., a 24-port switch or a 48-port switch). They then devise a way to connect multiple copies of the switch together such that it will function like one giant switch. For example, a company that has 130 computers might need to purchase three 48-port switches and configure them to operate as a single, large 144-port switch.

The use of modular hardware makes it possible for a LAN to span a large building without running a wire from every office to a single, centralized switch. The company constructs a switch for each floor of a building. A 96-port switch might be used on a floor with ninety computers, and a 24-port switch might be used on a floor that has twenty. Once each floor has a switch, the switches are interconnected with special high-capacity cables, and all switches are configured to operate like one giant switch.

Placing a switch on each floor of a building means that the cables between computers and the switch only need to run along one floor of the building. Consequently, no amplifiers are needed to boost signal strength; a cable can run from each office directly to the switch. In office buildings, the cables can run down hallways in the ceiling.

To summarize:

> *A LAN uses a device called a* switch *to connect multiple computers. A cable must run from each computer to the switch, but modular switch hardware means a switch can be placed on each floor of an office building and then all the switches can be interconnected to act like one giant switch.*

6.9 Wireless LAN (WLAN) Connections

A *wireless LAN* (*WLAN*) uses the same general approach as a wired LAN. The network hardware consists of a central electronic device that computers use for communication. Formally, the device is known as a *wireless access point*. Less formally, it is sometimes referred to as a *base station*; advertisements for consumers use the term *hotspot*. Figure 6.4 illustrates the idea.

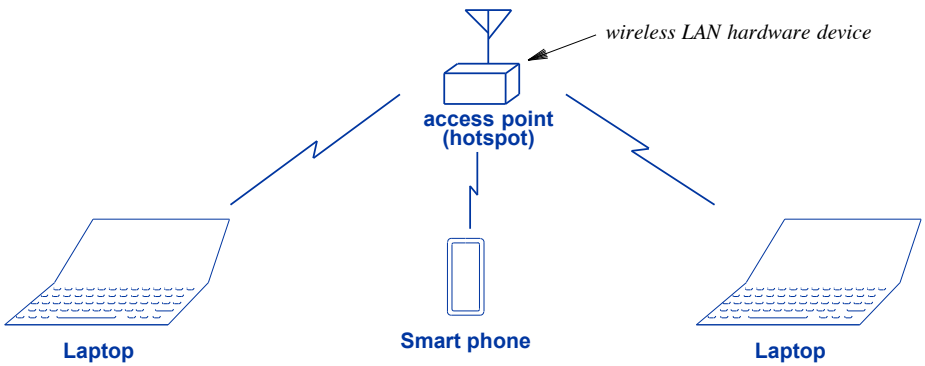


Figure 6.4 Illustration of a wireless LAN that uses radio waves.

As the figure shows, a wireless LAN has the same general structure as a wired LAN that uses cables. The only difference is that instead of connecting with cables, computers use radio waves to communicate with a hotspot.

6.10 Wired And Wireless LAN Technologies

At one time, many wired and wireless LAN technologies existed. Vendors had created a variety of LAN systems with various types of cables and connectors, data speeds, and prices. Although a few specialized LAN technologies still exist, two technologies now dominate the LAN marketplace: *Ethernet* and *Wi-Fi*.

Ethernet. Ethernet now dominates the wired LAN marketplace. Most desktops and many laptops come with Ethernet built in, as do some printers and televisions. More significantly, Internet Service Providers rely on Ethernet as the de facto standard for interconnection — the modems supplied for cable or DSL Internet service each have an Ethernet port to which a computer connects.

Wi-Fi. The wireless LAN market has also converged on a single technology: Wi-Fi. As with wired LANs, only a few special cases exist. Most portable devices, including laptops and smart phones, come with a Wi-Fi adapter built in. Many coffee shops, stores, airports, hotels, and other establishments offer free Wi-Fi to customers.

Interestingly, both Ethernet and Wi-Fi technologies have evolved. The available data rates have increased dramatically. Initially, Ethernet only supported transfers of 10 Megabits per second. A later version supported 100 Megabits per second, and then 1000 Megabits per second (1 Gigabit per second). High-end Ethernet switches can support 40 Gigabits per second. Similarly, Wi-Fi has made a dramatic increase in speed by defining new ways to modulate carrier waves. How did Ethernet and Wi-Fi survive the changes and come to dominate their markets? In each case, the designers used the same technique to make the transition to higher speeds painless: *backward compatibility*. When a computer is plugged into an Ethernet switch or when a computer communicates with a Wi-Fi access point, the hardware negotiates. Each side declares which speeds it can use. The two sides then pick the highest speed that they have in common. Thus, when an old computer is plugged into a modern Ethernet switch or when an old computer contacts a new Wi-Fi access point, communication is still possible because the new hardware reverts to the older speed for that connection.

6.11 Wireless PAN Technology

It may seem that our description of wireless LANs omits *Bluetooth*, a popular wireless technology. Bluetooth wireless has been used to connect a smart phone to a car, and to connect a smart phone to a door lock or other control mechanism. It has also been used to connect a computer to a wireless mouse, trackpad, or headphones.

Despite its many uses and popularity, Bluetooth does not compete with Wi-Fi because Bluetooth is not a LAN. Instead, Bluetooth is categorized as a *Personal Area Network* (*PAN*) technology. In general, a PAN spans a much shorter distance than a wireless LAN (a few feet instead of dozens of feet), and transfers data at a much slower rate. Consequently, PANs are typically used for communication with a headphone or mouse, but are not a useful way to transfer large amounts of data.

6.12 Connecting A Device To An Ethernet

Connecting a computer to an Ethernet is foolproof. The user plugs one end of an Ethernet cable into the computer, and then plugs the other end of the cable into an Ethernet switch. It doesn't matter which end of the cable is plugged into the computer and

which is plugged into the switch. The plug is designed so it cannot be inserted incorrectly. A small plastic tab on the plug clicks to lock the plug in the socket and prevent disconnection in situations where someone moves the cable or the computer.

Ethernet cables are carefully designed to make installation easy.

Of course, a computer must have the appropriate network hardware before it can attach to an Ethernet. The hardware can be internal (i.e., built into the computer when the computer is manufactured) or external (hooked onto a computer after the computer has been purchased). Networking professionals use the term *Network Interface Card* (*NIC*) to refer to the hardware. The term is archaic — it derives from early computers where a physical circuit board (i.e., a "card") was plugged into a computer. The terminology survives even though modern computers do not have a separate circuit board for each network interface.

Consumers and retail stores that sell computers tend to use the term *network adapter* or *Ethernet adapter* instead of NIC. For example, someone might ask, "Does your computer have an Ethernet adapter?"

An external Ethernet interface is only needed if a computer does not have an adapter built in. Suppose, for example, that a user wants to connect a laptop to an Ethernet, but the user did not include an Ethernet adapter when purchasing the laptop. The solution consists of using an *Ethernet dongle*, a small device that plugs into the computer and provides a socket for an Ethernet cable.

How does a dongle connect to a computer? The computer must have at least one external connection. For example, most computers have a USB port. To use the USB port for an Ethernet connection, a user must purchase a dongle that has a USB connector on one end and an Ethernet socket on the other.

Don't let the appearance of a dongle fool you. Most dongles are so small that they appear to consist of a few inches of cable with a connector (i.e., a plug) on each end. However, most dongles contain electronics — a vendor hides a tiny integrated circuit in the plastic housing that surrounds a connector. The circuit receives power from the computer, and performs the functions necessary to handle data transfer on each side of the dongle. To summarize:

Although it appears to be a tiny cable with a connector on each end, a dongle used to provide an external network interface is actually an electronic device with an integrated circuit hidden in the plastic housing surrounding one of the connectors.

6.13 Connecting A Device To A Wi-Fi Network

Each Wi-Fi hotspot is given a name, known as an SSID.† An SSID is up to thirty-two characters long, and is case-sensitive. Thus, the SSID *Main_street_cafe*

†The SSID acronym expands to *Service Set IDentifier*, a technical term taken from a networking standard.

differs from *Main_Street_cafe*. To connect to a hotspot, a computer must transmit a message that specifies the hotspot's SSID and requests access. The hotspot responds, and the two are connected. How does a device know the correct SSID? There are two options: a user can select from a list or configure an SSID manually.

Selecting from a list. A hotspot can "advertise" its SSID. To do so, it periodically broadcasts a message that any nearby computer can receive. The Wi-Fi software on most computers collects the advertisement messages, and forms a list of all the currently available hotspots. Software on the computer allows the user to choose one of the entries on the list and connect. The selection approach makes it easier to see all possible choices.

Manual configuration. As an alternative to selection, a user can manually enter the SSID of a hotspot. Manual typing is tedious, but can avoid cyber scams where someone impersonates a well-known hotspot and then intercepts or copies all the messages you send. Manual selection also allows the owner of a hotspot to make it difficult for others to discover the hotspot — the owner configures the hotspot so the hotspot does not broadcast its SSID. For example, a family might keep their hotspot "hidden" to prevent neighbors from accessing it.

6.14 Wi-Fi Security

Keeping a Wi-Fi hotspot hidden (i.e., not advertising the SSID) does not guarantee that communication will remain secure. In particular, because Wi-Fi uses radio waves to transmit messages, a specialized radio receiver can be created that snoops on a conversation (i.e., makes a copy of every message your computer sends to the hotspot and a copy of every message the hotspot sends to your computer).

To provide secure access, Wi-Fi technology includes a set of optional *encryption* mechanisms. Chapter 30 describes encryption in detail; for now, it is sufficient to understand a basic idea: encryption uses a password and a mathematical algorithm to change a message into code before the message is transmitted. Even if an outsider captures a copy of an encrypted message, the outsider cannot decode it without knowing the password.

6.15 The Importance Of LAN Technology

LAN technologies changed the way people used computer networks. Before LAN technologies were available, computer communication was extremely expensive. Once less expensive LAN technologies emerged, people began to use networks to connect machines within a room or within a building.

One of the most significant changes that LAN technologies produced was resource sharing. Before LAN technologies, most computers existed in a self-contained island. Each computer had a specific set of input and output devices, such as printers and disks, and each computer had one copy of the software that users could access. Once LAN

technologies became available, a set of computers could share resources. For example, a printer could be connected to a network and accessed by any of the computers on the network.

The ability to share resources changed the economics of computing dramatically. Because a network connection was much less expensive than a set of I/O devices, it became sensible to hook many computers to a network and to use the network to provide shared access to the I/O devices. To summarize:

Local Area Networks changed the economics of computing because they made it possible to use inexpensive computers that shared access to resources like printers and disks.

6.16 Relationship To The Internet

When the Internet project began, Local Area Network technologies were just emerging. A research lab at Xerox Corporation had invented *Ethernet*, and Xerox gave several universities a prototype version of the new LAN. Internet researchers who had used a LAN imagined a future in which LAN technology would become extremely inexpensive and widely available. They assumed, for example, that each organization would use one or more LANs to interconnect all its computers. They designed the Internet with the assumed future in mind, and it turns out that their assumption was correct.

EXERCISES

6.1 Suppose a company rents space in a building that consists of an extremely long passageway with forty offices spread down the passageway. The company plans to place an Ethernet switch in the center and connect all the offices to it, but discovers that the distance from the center to the outer offices is thirty feet longer than the maximum cable length allowed by Ethernet. How can the company still use Ethernet?

6.2 A user finds that a computer's battery drains slightly faster if the user leaves an Ethernet dongle plugged in all the time. Explain why.

6.3 Someone who lives in a city complains that when they try to find the hotspot for their favorite coffee shop, their smart phone shows a long list of SSIDs. A friend who lives in a small town says that their phone only lists a few entries. Explain the difference.

A Brief History Of The Internet

...how and why the Internet grew from its humble beginnings to become the largest network in the world

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7

Internet: Motivation And Beginnings

7.1 A Proliferation Of LANs

By the late 1970s, computer networking began to blossom. Several computer manufacturers had introduced small minicomputers with sufficient computational power to handle a few users. Because such computers were less expensive than older mainframe computers, individual departments could afford their own computer. As the previous chapter explains, Local Area Networks had appeared, and many organizations began installing them. Each department had sufficient budget to fund LAN installation and operation; the department could decide who had access, and could devise policies regarding use of the network.

Autonomy had the advantage of allowing each department to choose computers and LAN technologies that fit the department's needs. However, autonomy had some severe drawbacks. Allowing each group to act independently encouraged a proliferation of LAN technologies. Autonomy also meant differing policies — uses that were permitted in one department were forbidden in another. More important, heterogeneity had a negative economic impact on the overall organization. Because each department ordered equipment separately, the company could not negotiate a large quantity discount. In addition, networking staff were not interchangeable — differences in equipment meant that a technician who installed and managed a network in one department might not understand the equipment in another department.

The proliferation of multiple LAN technologies had another downside: data available in a given department could not be shared easily (e.g., across a LAN). To understand why, one must know that each LAN technology chose its own message format

and electrical signals. Thus, a company with multiple LAN technologies could not form a large, company-wide network merely by hooking cables between all the LAN systems.

7.2 No Technology Solves All Problems

In the 1960s, as work began on computer networks, many of the scientists and engineers expected data networking to follow the same approach as the telephone system. Individuals might start experimenting, but eventually, a single, large data communication system would exist that spanned the world and allowed computers to communicate, analogous to the way the phone system allows humans to communicate. They asked the question,

What is the best network technology for a global network system?

To answer the question, they started to explore possibilities. Engineers designed wired and wireless network technologies, and built experimental systems. They worked on ways computers could use networks, and they measured performance.

Looking for one technology to solve all problems seemed like a good approach, but it was flawed. As time went on, researchers found that a network that is ideal for one situation is not ideal for another. A technology that spans a long distance is substantially more expensive than a technology that spans a short distance. Because short wires pick up less interference, data can be transferred over a short wire at a much higher rate than it can be transferred over a long wire.

7.3 Wide Area Network Technologies

The search for a network technology that allowed all computers to communicate led scientists and engineers to experiment with networks that could connect computers across large geographic distances. Initially, they were called *long-haul networks*, but the name was changed to *Wide Area Networks* (*WANs*) to contrast with LANs. Chapter 5 describes the basics of modems that use a modulated carrier wave to send data over a long distance. WANs use modems to send signals across long-distance transmission lines, but a WAN does much more than merely connect a pair of computers. A given WAN consists of transmission lines interconnecting sites plus specialized hardware systems at each site that unifies the transmission lines into a coordinated system.

The hardware system at each site is known as a *WAN switch*. All the transmission lines coming from other sites connect to the WAN switch, as do local computers at the site. Figure 7.1 shows an example WAN that connects four sites.

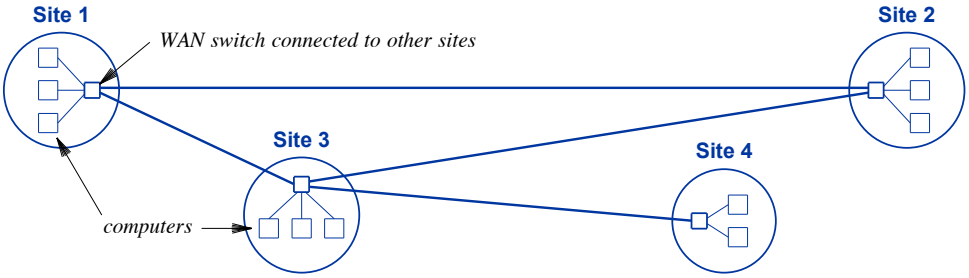


Figure 7.1 Illustration of a WAN that spans four sites.

Because it operates independent of computers, a WAN switch can keep running, even when all computers at the site are powered down. The WAN switch accepts messages sent from other sites. If a message has been sent to one of the local computers, the WAN switch delivers the message; otherwise, the WAN switch forwards the message on toward its destination. In the figure, for example, when a computer at Site 2 sends a message to a computer at Site 4, the WAN switch at Site 2 forwards the message to the WAN switch at Site 3, which forwards the message to the WAN switch at Site 4. Finally, the WAN switch at Site 4 delivers the message to the computer.

To summarize:

> *A Wide Area Network (WAN) technology uses a special-purpose hardware device called a* WAN switch *at each site. The WAN switch connects to transmission lines that lead to other sites, and forwards messages to their final destination.*

To understand the importance of WANs, imagine a company with offices in four cities: New York, Chicago, Los Angeles, and Austin. The company can install a WAN that links computers in each office. Physically, the WAN might consist of four leased transmission lines, similar to those in Figure 7.1. Conceptually, the WAN functions much like a giant LAN that allows the computers at all sites to communicate.

7.4 Can We Build A Global WAN?

It may seem that a WAN could satisfy the need for a single, global computer network. A WAN allows computers at a given site to communicate with one another. A WAN also allows a computer at any site to communicate with an arbitrary computer at any other site.

Although WAN technology looks promising, it does not solve the problem. There are two big drawbacks. The first drawback arises from cost because the equipment and transmission lines are far too expensive for an average user. The second drawback

arises from the general structure. A WAN is a great way to connect multiple sites, but not a great way to connect billions of individual subscribers. Additional technology would be needed to extend the WAN to homes and small offices without requiring them to install a WAN switch.

7.5 U.S. Department Of Defense Networking Research

In the 1960s, the U.S. Department Of Defense became interested in using computer networks. Because the idea of computer networking was new, little was known about how to build networks or how they could be used. Through the *Advanced Research Projects Agency* (*ARPA*),† the military funded research on networking using a variety of technologies. In the 1970s, ARPA had several operational computer networks and had begun to pass technology on to the military. ARPA projects included a WAN called the *ARPANET*, as well as networks that used satellites and radio transmission for communication.

DARPA realized the military would face the same problem that many organizations with multiple network systems faced: each network connected a set of computers, but no path existed between computers on separate networks. In essence, each network formed an isolated island that connected a set of computers, with no path between the islands.

7.6 Experimental Research

DARPA research examined how to interconnect all machines from a large organization. DARPA started with a few basic ideas, awarded grants to researchers in both industry and academia, and arranged for the researchers to cooperate in solving the problem. Researchers discussed their findings, and generated new ideas at regular meetings.

Instead of allowing researchers to engage in theoretical discussions, DARPA encouraged them to apply their ideas to real computers. DARPA chose researchers interested in experimental work, and insisted that they build prototype software to test their ideas.

7.7 The Internet Emerges

In 1973, two people working with DARPA, Vinton Cerf and Robert Kahn, wrote a paper that proposed a completely new approach. Instead of trying to find a single technology that would work well in all situations, interconnect the networks and add new software that passes data across the interconnection. They used the term *internetwork* to emphasize the interconnection of networks, and it quickly became shortened to *internet*. The name was applied to both the research project and to the prototype communi-

†The agency switched to the name *Defense Advanced Research Projects Agency* (*DARPA*), and after the research community switched preferences a few times, the name DARPA survived.

cation system that was built. To distinguish the idea from the specific prototype, researchers working on the DARPA project adopted the convention of writing *internet* in lowercase when referring to internetworks in general, and writing *Internet* with an uppercase I when referring to their experimental prototype. The key point is:

> *DARPA researchers investigated ways to solve the problem of incompatible networks. The research project produced a prototype system known by the name* Internet.

Chapter 13 discusses the structure of an internet in detail. It explains how the Internet solves the problem of connecting together LANs and WANs, and how seemingly incompatible networks work together.

7.8 The ARPANET Backbone

The ARPANET was especially important to the Internet project, and was often called the *backbone* network because it was the central WAN that tied researchers together. Each researcher working on the Internet project had a computer connected to the ARPANET.

Although having a Wide Area Network in place helped researchers communicate, the ARPANET became a key part of the Internet project because it allowed researchers to attach more than one computer at each site. Researchers took advantage of the feature and used ARPANET for two purposes. First, they used the ARPANET like a conventional WAN to connect a computer at each site. Second, they added an additional connection at each site, and arranged to use the additional connection to experiment with new ideas. Thus, the ARPANET served as both a standard network that permitted researchers to move data among sites involved in the project, and as an experimental network that allowed researchers to evaluate new network techniques and new applications.

7.9 Internet Software

Computer software forms an important part of Internet technology because software, not hardware, is key in making it possible to interconnect networks. Later chapters discuss the software that makes internetworking possible and efficient. For now, it is sufficient to know that DARPA's Internet project uncovered scientific principles and engineering optimizations that resulted in the software that made Internet communication possible and efficient. Although the software has many subparts, researchers worked to ensure that the software formed an integrated system. The end result is a smooth, apparently seamless software design. The parts work together so well that most users do not sense the underlying complexity.

7.10 The Name Is TCP/IP

Two pieces of the Internet software stand out as particularly important and innovative. *Internet Protocol* (*IP*) software provides basic communication, and the *Transmission Control Protocol* (*TCP*) software provides key functions that applications need. Consequently, the entire set of Internet communication software is known by the initials of these two important parts; usually the term is written with a slash between the names: *TCP/IP*.†

When a more formal name is needed for the set of software specifications, researchers use *The TCP/IP Internet Protocol Suite*. The formal name is more accurate because it points out that the entire set contains more than just the two protocols. In the end, however, the simpler name has persisted — both vendors who sell the technology as well as users who acquire and install it use the term *TCP/IP*.

7.11 The Surprising Choice Of Open Standards

To encourage vendors to adopt Internet technology, DARPA decided to make the research results public. Whenever a researcher discovered a new technique, measured network performance, or extended the TCP/IP software, DARPA asked that the researcher document the results in a report. All the specifications needed to build TCP/IP software, and all the experience installing and using it were documented. DARPA made the reports available to everyone.

DARPA's practice of publishing network specifications was surprising because it ran counter to accepted practice. Commercial companies that developed network technologies had kept their discoveries and technical specifications private. In fact, most companies had filed patents to guarantee that no other company could use the same techniques. The idea was derived from standard business practice:

> *Prevailing business opinion suggested that a company selling computer networks could maximumize profits by protecting their networking technology with patents that ensured no other vendor's computers could attach to the network without paying a fee.*

In the mid-1970s, major computer companies that sold network systems made sure that only their computers could connect to their network technology. Various LAN and WAN technologies had been created, but each vendor guarded their proprietary technology. To emphasize that a technology is not available to outsiders, computer professionals use the term *closed*. A closed technology is proprietary to one company, and not available for use by others without a fee.

From its inception, the Internet project aspired to produce an *open* set of standards that would permit computers from all vendors to communicate and permit any type of network to be included. The open philosophy meant that instead of filing patents,

†One pronounces the name by spelling out the letters "T-C-P-I-P."

researchers made their discoveries about the Internet public, and published the specifications needed to build TCP/IP software for anyone to use. Although it now seems noncontroversial, the approach was shocking at the time.

> *A network technology is* closed *if a company owns the technology and uses patents and trade secrets to prevent other companies from building products that use it. By contrast, the Internet is* open *because all specifications are publicly available and any company can build a compatible technology.*

7.12 Open Communication Systems Win

Computer companies found that, despite their efforts to sell closed systems, customers began to acquire several brands of computers. Advances in processor and memory hardware made new computer designs possible. Plummeting costs made personal computers affordable. Organizations like the U.S. military realized that as computer technology evolved, vendors would continually offer new models. Furthermore, not all software worked on all computers. A large organization usually has many brands and models of computers because it needs software systems and computers for many purposes. Only an open network system can be used to interconnect computers from multiple vendors. In summary:

> *Because large organizations have multiple types and brands of computers, a closed network that is restricted to one vendor's products is inadequate. Only an open communication system allows computers from multiple vendors to communicate.*

7.13 Placing Internet Technical Documentation Online

The Internet project pioneered another idea: using a computer network to make technical documents available. When DARPA began the Internet project, computer networks were so new that only a few people had ever heard of them; very few individuals had actually used a network. Fortunately, most of the researchers that DARPA chose to work on the Internet were among the group that had experience with networks. They had helped design and build the ARPANET, and had devised a few networking applications. They decided to use the ARPANET to exchange technical information, and placed technical documents in computer files accessible over the ARPANET. It was a pioneering idea at the time.

Initially, the Internet researchers planned to issue technical documents in two steps. When a report was first written, it would be made available to other researchers for comments. After a short time, the author would incorporate all comments and issue a final version of the report. To implement the two steps, researchers established two

series of reports: *Requests For Comments* (*RFCs*) and *Internet Engineering Notes* (*IENs*).

Unfortunately, the best laid plans often go astray. Researchers found that some of the initial reports were so well done that they did not need revision or improvement. Other reports were rewritten completely, but reissued as an RFC for another round of comments. Most researchers found it more productive to continue investigating new ideas than to edit old reports. In the end, RFC reports became the official record of the project and the IEN series was dropped. The irony is that each of the documents that specifies the technology of the largest, most successful computer network in history has a label that implies the work is somehow unfinished and the author is still waiting tenuously for comments. To summarize:

> *For historical reasons, the technical documents that define TCP/IP and related Internet technology are called* Requests For Comments.

Researchers working on the Internet project had access to all RFCs because they were stored on a computer attached to the ARPANET. Each RFC was assigned an integer number, and an index was kept that listed the title of each number. At any time, a researcher who wanted to know the details of a particular piece of Internet software could use the ARPANET to retrieve the RFC that contained the information. If the researcher did not remember which RFC was needed, they could retrieve the index.

Keeping the project documentation accessible across the network enabled everyone working on the project to coordinate their activities and keep software up-to-date with the specifications. More important, rapid communication among the researchers increased the speed at which the project progressed.

> *Because documents that specified technical details of TCP/IP and the Internet project were accessible over the ARPANET, work on the project proceeded more quickly.*

As the Internet project progressed, the technology reached a stage where prototype software could be deployed and tested. A fledgling Internet was born. One of the first applications that researchers devised for the new Internet was a mechanism that could be used to access RFCs. In fact, almost all the initial applications for the Internet were motivated by the needs of the researchers building it.

7.14 The U.S. Military Adopted TCP/IP

By 1982, a prototype Internet was in place and the TCP/IP technology had been tested. A few dozen academic and industrial research sites had been using TCP/IP regularly. Then the U.S. military started to use TCP/IP on its networks.

In the beginning of 1983, DARPA expanded the use of TCP/IP to include all military sites that connected to the ARPANET. The date marked a transition for the Internet as it began to change from an experiment to a production communication facility.

7.15 Summary

The Internet began as a research project funded by DARPA. Researchers studied ways to interconnect computers that used various kinds of networks. The name *Internet* refers to both the project and the prototype network system that researchers built.

Known by the name *TCP/IP*, the software used to make the Internet operate contains many complex pieces of software that work together to provide communication. The software works so well that it hides the details of the underlying hardware and provides the illusion of a seamless system.

The Internet standards are *open* because the specifications needed to build TCP/IP software or use the Internet are freely available to everyone. Researchers who devised the Internet published technical information in a series of reports that describe the Internet and the TCP/IP software it uses. For historical reasons, each document in the series is labeled *Requests For Comments*.

EXERCISES

7.1 Many individuals have tried to claim credit for the Internet. An easy way to spot a fake claim is to look for dates before 1973 when the Internet idea was initially published as a research paper. Search online and see if you can spot a fake claim.

7.2 After the Internet demonstrated the value of open standards, others adopted the idea. Look up "open software" on Wikipedia and write a one-paragraph summary of the idea.

7.3 In the 1970s, one of the main arguments in favor of using closed technologies focused on economics: companies asked how they could make money if other companies could build the same products. Look online and compare the histories of Cisco Systems, one of the first companies to build products using the open Internet standards, and Digital Equipment Corporation, a company that sold proprietary DECNET protocols. Which was most successful?

7.4 Perform an Internet search of humorous or funny Internet history, and find something that makes you smile.

7.5 Although those of us working on the Internet project were using an early version in the 1970s, most people only heard about the Internet in the 1990s. Ask older family members about their first encounter with the Internet, and have them guess when the Internet was created.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

8 The Incredible Growth

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

8

The Incredible Growth

8.1 Introduction

During the years between 1980 and the 2000s, the Internet changed from a small, experimental research project into the world's largest computer network. In 1981, the Internet connected approximately one hundred computers at research sites and universities. By 2000, over seventy-two million computers were attached, and the Internet continues to grow. The introduction of the smart phone changed the Internet considerably; in 2016, more users accessed the Internet through a smart phone than through a laptop or desktop computer.

This chapter chronicles the phenomenal growth of the Internet and the changes that accompanied it. It discusses steps that were taken to stimulate growth, and concludes by explaining some of the consequences and opportunities that arose from rapid adoption.

8.2 Stimulating Adoption

In 1980, the Internet was merely a research project. A handful of universities and research labs had copies of the TCP/IP software. By 1985, it was becoming a production network system. Experimental TCP/IP software was available for several brands of computers, and it was used every day. The Internet reached researchers at few dozen academic and industrial research labs.

Before the U.S. military could use TCP/IP for production work, however, the technology needed to become more robust. The software needed to be polished and tested,

and the whole system needed more tuning. DARPA considered the next step in its research program carefully.

8.3 Meanwhile, Back In Computer Science

While DARPA worked on the Internet research project, another technology came from a research lab and swept the computer science community: an *operating system*. Although vendors use the term *operating system* to refer to all the software that comes with a computer or smart phone, computer scientists use the term to describe the main piece of software that manages the computer, runs apps, controls input and output devices, and provides file storage. Operating systems are so complex that scientists and engineers spent years in the 1960s trying to understand them. By 1970, computer vendors had adopted the *closed* paradigm for operating system software, and vendors had created a proprietary operating system for each of their computers.

In the early 1970s, a small team of computer scientists at Bell Laboratories built a new operating system called the *Unix Time-sharing System*. Because Bell Laboratories used a variety of computers, the researchers wanted an operating system that could run on any hardware. So, they designed the system to be general — they created the software carefully, and made it easy to move a copy onto new computers.

Bell Labs decided to allow universities to obtain copies of the Unix system for use in teaching and research. Because they were interested in measuring its portability, Bell Labs gave away copies of the code, and encouraged universities to try running the system on new computers. As a result, the Unix system became one of the first operating systems that students could study.

A group of faculty and graduate students from the University of California at Berkeley became interested in the Unix system. They wrote application programs and modified the system itself. They added new features and experimented with applications that communicated over a Local Area Network. To make the work available to other universities, researchers at Berkeley established a software distribution facility. When a university wanted a copy of the software, the distribution facility mailed a magnetic tape that contained the software. The Berkeley version of the Unix system, known as *BSD Unix*,† became popular at other universities.

8.4 The Internet Meets Unix

DARPA realized that the Berkeley work on operating systems reached many universities, and decided to use it to disseminate Internet software. They negotiated a research contract with Berkeley. Under the terms of the contract, DARPA gave researchers at Berkeley a copy of the TCP/IP software that had been developed as part of the Internet project. Berkeley incorporated the software into their version of the Unix system, and modified application programs to use TCP/IP.

---

†The acronym BSD stands for *Berkeley Software Distribution*.

When Berkeley issued its next major software distribution, most computer science departments received TCP/IP software at virtually no cost. Although only a few computer science departments had computers connected to the Internet, most of them had a Local Area Network or were about to install one. They knew that their students needed to study networking. They also knew that using a network would make computing easier because it would allow users to share resources like printers.

For many departments, TCP/IP was the first viable networking software they had encountered. It offered a low-cost, efficient way to provide a departmental network and a technology that could be studied in classes. Thus, in a short time, most computer science departments had TCP/IP software running on their Local Area Networks, even though most had not yet connected to the Internet. The point is:

> *Computer science departments in universities received TCP/IP software along with a release of Unix system software from U.C. Berkeley. Although only a few departments had computers connected to the Internet, most of them used TCP/IP on their Local Area Networks for teaching, research, and production computing.*

8.5 The U.S. Military Makes A Commitment

By the early 1980s, the Internet operated reliably. It interconnected academic and research sites. More important, the Internet demonstrated that the basic principles of internetworking were sound. Convinced of the Internet's viability, the U.S. military started to connect computers to the Internet and to use TCP/IP software.

In 1982, the U.S. military chose the Internet as its primary computer communication system. Consequently, a cutoff date was planned. At the beginning of 1983, the ARPANET and associated military networks stopped running old communication software and switched to TCP/IP. On the cutover date, any computer that did not understand TCP/IP could not communicate. The point is:

> *Although the U.S. military funded Internet research and eventually chose to use the TCP/IP software, internetworking was developed and tested at civilian sites. Only after Internet technology had been demonstrated did the military switch its computers to the new technology.*

8.6 The Internet Doubled In Size In One Year

Before the U.S. military started using TCP/IP on all its computers, the Internet interconnected approximately two hundred computers. One year later, it had doubled in size. In retrospect, the increase seems trivial. It involved hundreds, not thousands or millions of computers. At the time, however, the increase was significant.

Anyone who has written a computer program knows that the program has specific size limits built into it. For example, some parts of the TCP/IP software use lists of computers and the addresses used to access them. When the Internet only contained dozens of computers, programmers chose maximum sizes that seemed huge (e.g., 300). As new computers joined the Internet, the list of computers exceeded the limits, and the software had to be revised to accommodate longer lists. At first, researchers made small increments to the software. They increased the capacity by ten or twenty percent. Soon, they found that it was insufficient, and further increases were needed. As the Internet continued to grow, the process of changing the software kept pace.

In addition to uncovering limitations in the software, the Internet growth revealed limits in manual and clerical procedures. For example, each time a new computer was added to the Internet, several people had to take action. Someone had to review the reasons for the connection and its relationship to the project before approving the connection. Someone else had to assign a name to the computer, and then enter it in a database. Finally, someone had to make a physical connection between the computer and the network.

During the period of rapid growth, researchers were busy updating the software and had little spare time to help with manual procedures like registration; the duties began to pass to a professional staff. We can summarize what happened:

> *As new computers were added to the Internet, it doubled in size in a single year. The rapid growth forced researchers to tune administrative procedures as well as the communication software.*

8.7 Internet For Every Computer Science Department

In the late 1970s, many computer scientists recognized the importance of networking. A small group of researchers proposed a networking project to the *National Science Foundation* (*NSF*)† with a goal of devising a computer network to connect all computer science researchers in the U.S.

After reviewing the proposal, the National Science Foundation funded a project to build the *Computer Science Network*. The project, which also had support from DARPA, became known by the acronym *CSNET*. To reach all computer scientists in the country, CSNET had to contend with the problem of providing network service to small universities in rural areas as well as major universities in metropolitan areas. DARPA encouraged CSNET to run TCP/IP software and connect researchers to the Internet. For smaller institutions that could not afford direct connections, CSNET devised ways to provide limited network services at much lower cost.

By the time the U.S. military selected the Internet as a primary computer communication system, many of the top computer science groups in industry and academia were already using it. Over the next few years, CSNET worked to provide Internet connections to the remaining computer science departments. As a result, by the mid-1980s, most computer scientists had Internet access.

---

†NSF, a U.S. federal agency, is responsible for funding research and education in science and engineering.

8.8 Graduate Student Volunteers Contribute

Connecting computer science researchers to the Internet had an interesting effect. Although some computer scientists work in industrial research labs, many are professors who work in universities, where they also teach classes and advise students. The professors talked to students about the Internet project, the technology and software that it used, its success, and the remaining research problems. The professors' enthusiasm was contagious.

Students became interested in learning more about TCP/IP and the Internet. Graduate students who were searching for research topics began to investigate the technical details of TCP/IP software. They studied ways to extend the Internet technology, and devised experiments to measure its capabilities. They considered new applications, and found ways to extend the functionality. The result was synergistic: students gained valuable knowledge and experience with computer networks, while their creative energies helped advance Internet technologies.

8.9 Internet Governance: The IAB And IETF

Scientists and engineers working on the Internet held regular meetings to discuss new ideas, review the technology, share discoveries, and exchange technical information. DARPA originally named the group the *Internet Advisory Board*. With the Internet growing rapidly, DARPA decided that the group of scientists should have a more formal structure and more responsibility for coordinating TCP/IP research and Internet development. It renamed the group the *Internet Activities Board*. Following military tradition, the board became known by its acronym, IAB.

DARPA appointed a chairman of the IAB, who was given the informal title *Internet Architect*, even though the Internet was already growing too rapidly for a single person to be responsible for an architectural plan. Another member of the IAB was designated to be the *RFC Editor*, and given responsibility for reviewing and editing all RFCs before they were published. Other scientists on the IAB were each assigned a specific problem to investigate.

To study an assigned problem, each member of the IAB gathered volunteers from the research community to serve on a *task force*. Each task force held meetings to discuss ideas, resolve issues, generate new approaches, and report on experiments. If a task force reached a consensus on a new approach, members would build prototype software to demonstrate how their ideas worked in practice, and then would generate and submit a specification as an RFC.

The IAB guided the development of the Internet for several years. In 1989, it was reorganized to add more representatives from commercial companies. The IAB's duties and interactions with other groups were reorganized again in 1992, when it became part of the *Internet Society*. At the time of its second reorganization, the IAB kept the acronym, but changed its name to the *Internet Architecture Board*. The IAB divested most

of its technical responsibilities, passing more control to subordinate groups, and leaving the board as the ultimate arbiter of policies and standards.

Among the subordinate groups, one stood out: the *Internet Engineering Task Force* (*IETF*). The IETF has survived reorganizations, and has taken over responsibility for generating new Internet technologies, revising older technologies, and creating, revising, and publishing standards documents. Most RFCs now originate within the IETF from committees, which are called *working groups*.

The IETF is partitioned into areas of interest, with an area director assigned to coordinate groups within each area. The IETF holds open meetings approximately three times per year. When it holds a meeting, thousands of people attend, most from commercial companies. Attendees are volunteers who attend to hear about the latest developments and participate in efforts to refine and improve the software.

We can summarize:

> *The group responsible for guiding the research and development of the Internet is known as the* Internet Architecture Board *(IAB). The primary subgroup responsible for technical matters is known as the* Internet Engineering Task Force *(IETF).*

8.10 NSF Led Internet Expansion

During the years following the military adoption of TCP/IP, rapid growth continued. Government agencies, such as *NASA* (the *National Aeronautics And Space Administration*) connected to the Internet. By the mid-1980s, the National Science Foundation (NSF) realized that eminence in science would soon demand computer communication. Before computer networks, scientists exchanged ideas by publishing them in scientific journals, which took many months, sometimes years. Computer communication was about to change the way scientists did research. A computer network makes it possible to share data as an experiment proceeds, making it possible for many other scientists to analyze the results without traveling to the site of the experiment.

Recognizing how important the Internet was becoming to science, NSF decided to fund further Internet growth. In 1985, NSF announced that it intended to connect researchers at 100 U.S. universities to the Internet. NSF advised the U.S. Congress of the plan, and received additional funds to support networking. NSF consulted experts in the field, devised a plan, and began a program that resulted in major changes to the Internet.

Scientists often use sophisticated, high-speed computers called *supercomputers* to analyze data from their experiments. Because supercomputers were expensive, NSF had established five supercomputer centers around the country. A scientist working on an NSF project had to travel to the nearest supercomputer center to process their data.

As the first step of Internet expansion, NSF built a Wide Area Network that interconnected its five supercomputer centers. The network used TCP/IP, and provided a

connection to the Internet. Named *NSFNET*, the network was initially much smaller, and not any faster, than the ARPANET. Scientists found the network useful, but not exciting.

8.11 NSF Target: All Of Science And Engineering

Spurred by the success of CSNET and the initial NSFNET, NSF launched a new program to keep the U.S. competitive. The program had an ambitious goal:

> *NSF decided that the U.S. would not remain competitive unless every science and engineering researcher had Internet access.*

To achieve the goal, NSF decided to use its funds to create a major new Internet that had significantly more capacity than the existing Internet. After examining available technologies and reviewing its budget, NSF decided that it could not afford to pay for the entire project. Instead, it decided to offer partial support, in the form of federal grants. Companies and other organizations submitted written proposals to NSF to request funding to work on the project.

NSF divided the grants into two types. First, NSF funded a group that wanted to build and operate a new high-speed Wide Area Network to connect parts of the Internet. The new WAN had to replace parts of the ARPANET as well as the original NSFNET. Second, NSF funded groups wanted to interconnect computers in a small region of the country and attach them to the new WAN. For example, NSF thought that each state might choose to apply as a group. Originally, the groups were referred to as *NSF Regional Networks*. Later, when it became clear that some of the groups spanned large geographic areas, NSF began referring to them as *NSF Mid-Level Networks*, but most professionals still called them *regionals*.

Because most universities or companies already had LANs connecting their computers, NSF decided to use its funds to help pay for long-distance connections; individual companies and schools paid for their own internal networks.

8.12 The NSFNET Backbone

NSF used a competitive bidding process when it awarded a grant for the new Internet WAN, which became known as the *NSFNET backbone*.† In 1987, NSF asked for proposals and used a panel of scientists to help assess them. After considering the alternatives, NSF selected a joint proposal from three organizations: IBM, a computer manufacturer; MCI, a long-distance telephone company; and MERIT, an organization that built and operated a network connecting schools in Michigan.

The three groups cooperated to establish a new Wide Area Network that became the backbone of the Internet in the summer of 1988. MCI provided long-distance

†The term *backbone* is used as an analogy to a human spine that forms a central structure to which many other bones attach.

transmission lines, IBM provided the dedicated computers and software used in the WAN, and MERIT operated the network. Most people referred to the new backbone using the same name applied to its predecessor, *NSFNET*.

8.13 On To The ANS Backbone

Eventually, as traffic on the new WAN reached capacity, NSF approved reorganizing the network slightly and tripled the capacity of each transmission line. By the end of 1991, it became clear that the Internet was growing so fast that the NSFNET backbone would soon be completely saturated. NSF realized that the federal government could not afford to pay for the Internet indefinitely. They wanted private industry to assume some responsibility. To solve the problem, IBM, MERIT, and MCI formed a nonprofit company named *Advanced Networks and Services* (*ANS*).

During 1992, ANS built a new Wide Area Network to serve as the Internet backbone. Known as *ANSNET*, the WAN used transmission lines with 30 times the capacity of the NSFNET backbone it replaced. Figure 8.1 illustrates the ANSNET connections.

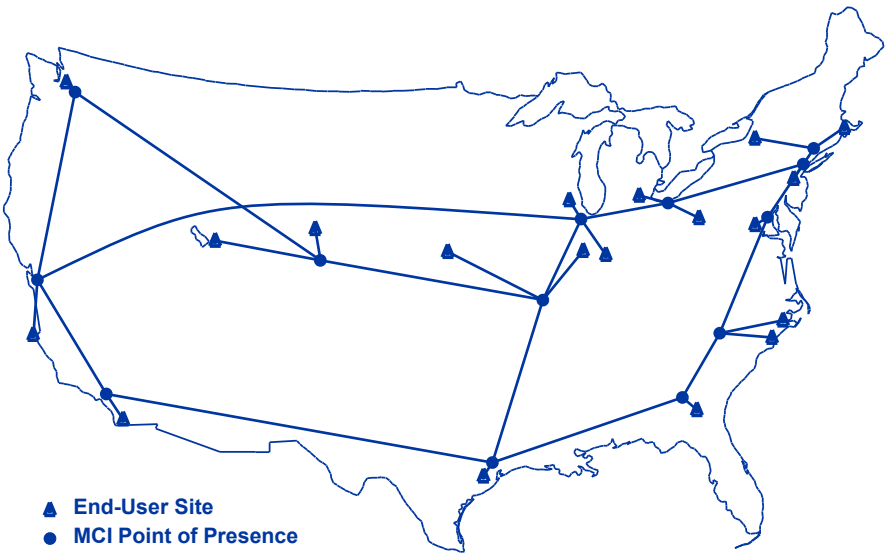


Figure 8.1 The backbone of the Internet in 1995, known as ANSNET. Funding came from NSF, IBM, MCI, and MERIT.

8.14 Commercialization

The move to ANSNET and associated regional networks represented a major shift in the Internet. For the first time, the Internet had become commercial. When DARPA and NSF provided Internet funding, they had to adhere to government rules. In particular, the Internet had an *Acceptable Use Policy* (*AUP*), that allowed scientists and engineers to use it for research and teaching, but not for money-making activities. The legal rules changed when ANS, not the federal government, owned the transmission lines and computers that constituted the network. As ownership began to transfer to private companies, the Internet took its first steps toward commercialization and privatization.

NSF played a role in creating another aspect of the Internet that we now take for granted: *Internet Service Providers* (*ISPs*). In the early days, the Internet grew chaotically. When a new site wanted to join, the site paid for a transmission line to the nearest Internet site. When NSF decided to fund regional networks that each provided service to a group of subscribers in an area, the connectivity paradigm changed. When a site wanted to join the Internet, the site contacted the appropriate regional network for service. When NSF transferred ownership of regional networks to the private sector, they each became an ISP.

8.15 Exponential Growth

As NSF connected scientists and engineers, the Internet grew incredibly fast. In 1983, the Internet connected 562 hosts. Ten years later, it connected over 1,200,000 hosts and was still growing quickly. Such staggering growth can best be understood by considering individual hosts:

> *By 1999, the Internet was growing so fast that, on the average, a host was added to the Internet every second. By 2006, the average exceeded ten hosts per second, and by 2016, the use of smart devices made growth impossible to measure.*

Although the Internet did not grow at exactly the same rate in all years and most of the hosts were added in recent years, a trend of doubling can be identified. In round numbers, the Internet has experienced sustained growth of approximately *10* percent per month, doubling in size approximately every *10* months. Mathematicians call such growth *exponential*. The table in Figure 8.2 illustrates growth from 1983 through 2018.

Exponential growth has some interesting properties. For example, although the Internet has been around for many years, exponential growth means that approximately half the people connected to the Internet have gained access in the past year. Interestingly, that same statement could have been made in any previous year. In fact, the following summarizes the incredible growth:

At any time from 1983 through 2007, approximately half the Internet growth occurred in the previous 10 to 12 months.

What happened to growth starting in 2007? The answer is the smart phone happened. The numbers listed in Figure 8.2 are a count of host computers with permanent IP addresses. As Chapter 26 explains, the cell phone system issues temporary IP addresses to smart phones, making it impossible to obtain an accurate count. Thus, numbers in Figure 8.2 for years after 2007 are low, and in January of 2018, fewer hosts had permanent addresses than the year before.

Year	Computers	Year	Computers
1981	213	2000	72,398,092
1982	235	2001	109,574,429
1983	562	2002	147,344,723
1984	1,024	2003	171,638,297
1985	1,961	2004	233,101,481
1986	2,308	2005	317,646,084
1987	5,089	2006	394,991,609
1988	28,174	2007	433,193,199
1989	80,000	2008	541,677,360
1990	313,000	2009	625,226,456
1991	535,000	2010	732,740,444
1992	727,000	2011	818,374,269
1993	1,313,000	2012	888,239,420
1994	2,217,000	2013	963,518,598
1995	4,852,000	2014	1,010,251,829
1996	9,472,000	2015	1,012,706,608
1997	16,146,000	2016	1,048,766,623
1998	29,670,000	2017	1,062,660,523
1999	43,230,000	2018	1,003,604,363

Figure 8.2 Internet hosts with permanent IP addresses each year from 1983 through 2018. The counts for years 2007 on are low because smart phones cannot be counted accurately.

8.16 When Will Growth End?

At various times in the past, people have predicted the imminent collapse of the Internet by observing that some small piece of the technology was reaching its limit. By 1990, for example, someone had predicted that the Internet could not survive past March of 1993. In 1995, a group predicted that the Internet would collapse in the summer of 1997. Then in 1999, another group predicted collapse in 2004. The predictions of doom have been incorrect, and the Internet keeps growing. Each time the traffic has approached the capacity of a backbone network, a new backbone technology has been developed and deployed with significantly more capacity. When the traffic approached the capacity of the systems that forward data across the Internet, faster systems have been created. At one time, a group observed that Internet growth must be curtailed because it was about to overtake the worldwide production of computers. However, the group focused on PCs, and was surprised when tablets and smart phones came along.

Another group calculated the end of growth by carefully estimating the world population growth and the rate at which users were being added to the Internet. They confidently predicted a date when Internet growth would stop because every person on earth would have a computer hooked to the Internet. Since that prediction two things occurred. First, smart devices mean many people have multiple devices for use in their business and personal lives. Second, as Chapter 24 explains, the latest Internet expansion is occurring because users are connecting many small devices to the Internet.

The point is that both technology and the way we use the Internet keeps changing, making accurate prediction difficult.

Although researchers agree that growth cannot continue unchecked forever, The Internet has persisted in growing beyond predictions of its end.

EXERCISES

8.1 Various groups estimate the number of Internet users. Search online to obtain an estimate of how many users worldwide access the Internet on an average day.

8.2 Extend the previous exercise and find out what percentage of the world's population has Internet access.

8.3 Search the Internet to find programs and projects that are being undertaken to deliver Internet services to the few groups of people who do not yet have access.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Inside The Internet

An Explanation Of The Underlying Technology And Basic Capabilities Of The Infrastructure

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

9

Packet Switching

9.1 Introduction

This chapter begins an exploration of the basic communication technology that the Internet uses. It describes the fundamental mechanism all computer networks use to transfer data, and explains why the scheme works well. Succeeding chapters show how the Internet uses the mechanism. Understanding how networks function is important because knowing about technology allows one to appreciate mechanisms, understand possibilities, and distinguish between apparent magic and advanced technology.

9.2 Sharing To Reduce Cost

Chapter 7 claims that using a dedicated connection between each pair of computers is too expensive. To understand the expense, look at some examples. If four devices need a wired connection between each possible pair, a total of only six connections are needed. For seven devices, the total is twenty-one, and for twenty devices, one hundred ninety wires are required. If an organization has fifty devices, over twelve hundred connections are needed!

To avoid the expense and inconvenience of running a dedicated connection between each pair of communicating devices, a computer network arranges for multiple devices to share the underlying transmission facilities. We can summarize:

Because running connections between every pair of communicating devices is prohibitively expensive, networks arrange for multiple communications to share a given transmission path.

9.3 Sharing By Taking Turns

How can multiple computers share a transmission path? They take turns. A good analogy comes from early telephone systems. Subscribers who wanted to lower costs could choose a *party-line* service. Instead of running a separate set of wires to each subscriber's house, a party-line service meant one set of wires ran down the street and multiple telephones connected to the set of wires. If one of the subscribers on the party line was using the phone, another user on the party line who picked up their phone would hear the conversation. Etiquette dictated that if a subscriber heard someone else talking, they were to hang up immediately and try again later.

Many computer networks use the party-line approach — multiple computers share a transmission system. When a computer has data to send, the computer sends immediately if the network is idle, and must wait for the transfer to complete if the network is in use. The scheme is used for wireless networks, such as Wi-Fi, as well as wired networks. The point is:

Only one data transfer can occur on a transmission path at a given time. When computers share a network, they take turns sending data.

9.4 Avoiding Long Delays

We have all experienced the downside of taking turns to share a resource: delay. For example, consider an office where employees share a photocopier. Imagine the frustration of needing to copy one page and finding that you must wait for someone who is copying a 900-page document.

If computer networks followed the same approach, long transfers would leave users frustrated. For example, suppose all the subscribers on your street shared a connection that led to your ISP, and suppose all the subscribers took turns using the connection. Imagine how frustrating it would be if you had to wait twenty minutes while a neighbor streamed an HD movie before you could use the connection.

To avoid situations where the data transfer by one device leaves others waiting, networking researchers invented a system that prevents long delays. The idea is straightforward: instead of allowing a given device to use the network for an arbitrarily long time, limit the amount of data that a device can transfer on each turn. The idea, which was invented in the 1960s, is called *packet switching*, and the unit of data that can be transferred is called a *packet*. Figure 9.1 illustrates how devices use packet switching.

The figure shows four devices attached to a network. Device *A* is sending data to Device *C*, while Device *B* sends data to Device *D*. Each sender divides its outgoing message into packets, and they take turns sending packets. First *A* sends a packet, then *B* sends a packet, then *A* sends a packet, and so on.

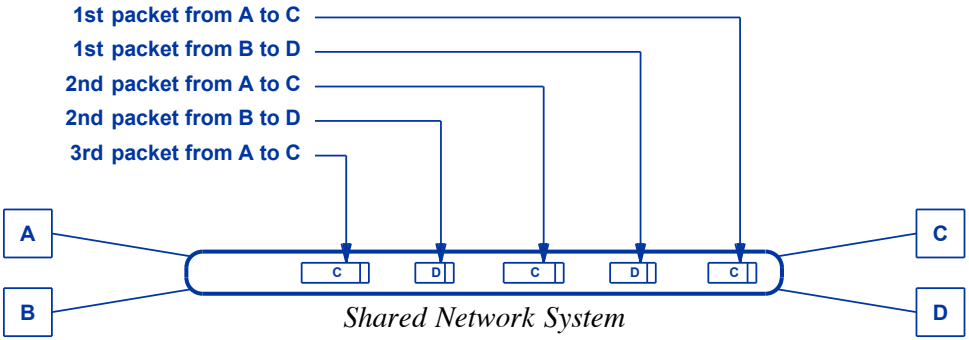


Figure 9.1 An illustration of devices taking turns sending packets across a network. Device *A* communicates with device *C*, while device *B* communicates with device *D*.

9.5 Long Messages And Short Packets

Both LANs and WANs use packet switching, as do both wired and wireless networks. Furthermore, packet switching is used with all applications. Whether you send a text message, photo, document, video clip, or download a song, the data is always sent in packets. When a user requests a data transfer, software on the sending device divides the item to be sent into packets before sending; on the receiving device, software collects the series of incoming packets, and reconstructs the original item. A user remains unaware that packets are being used.

The maximum size of a packet is set by a network designer. For Ethernet, the maximum size is 1500 bytes of data; for Wi-Fi, 2304 bytes of data can be sent in a single packet. The details are not important, but remember that when you transfer a large item (e.g., a movie), the item is divided into many packets. If an item is short enough, the entire item can fit into a single packet. Because devices take turns, a single-packet message can be sent without waiting for a long transfer to complete. From a user's point of view, a short message appears to "sneak in" and use the network while a long transfer continues.

9.6 Each Packet Contains Extra Information

Each packet sent across a network originates at one device and is destined for another. If devices take turns sharing, how can the network hardware tell which device should receive a given packet? The answer is that computer networks operate the same

way as a postal system: in addition to the data being sent, each packet contains extra information that specifies the device to which the packet is being sent and the device that sent the packet.

We use the term *packet header* (or just *header*) to refer to the part of the packet that holds the extra information. The term was chosen because the header precedes the data. That is, each packet starts with a header and ends with the data being sent. Think of the header as a label that specifies two things: the device that sent the packet and the device that should receive the packet.

9.7 Devices Have Addresses

How is a device identified? Each device on a network is assigned a unique number known as the device's *address*. One popular addressing scheme uses the term *MAC address*. To identify the pair of communicating devices, the header at the beginning of a packet contains two important MAC addresses: the MAC address of the device that sent the packet and the MAC address of the device to which the packet is sent. The sender's address is called the *source address*, and the receiver's address is called the *destination address*. When they travel in a packet, the addresses are actually binary numbers. However, humans who manage networks use more convenient forms (e.g., decimal numbers) to represent addresses. The important idea is:

> *Each device attached to a network is assigned a unique number called its* address. *In addition to the data being sent, a packet contains the address of the device that sent the packet and the address of the device to which the packet is sent.*

9.8 Packet Size

How large is a packet? The packets used with modern networks usually hold fifteen hundred characters (bytes) of data. If you download a movie, packet after packet will each contain fifteen hundred bytes of data.

Although packet switching technologies limit the amount of data in a packet, they allow the sender to transmit any size packet up to the maximum. For example, suppose you are engaged in a chat session and send a short reply, 'OK, it's a date." The entire reply takes sixteen characters. So, the chat application will send a single packet that carries sixteen bytes of data. Similarly, when you use a network to transfer a large file or a video, the final packet of the transfer will not be full (unless you happen to transfer a file that is an exact multiple of the packet size).

9.9 To Humans, Packet Transmission Seems Instantaneous

When thinking about packets traveling across computers networks, we must remember that network hardware operates incredibly fast. For example, sending a fifteen-hundred-byte packet across an inexpensive Ethernet LAN takes approximately

0.000012 seconds = 12 millionths of a second (12 microseconds).

Events measured in millionths of a second are hard to imagine. To put it another way, it takes less than two tenths of a second to transfer ten thousand completely full packets across an Ethernet. To summarize:

To avoid long delays, a packet switching system divides each transfer into small packets and arranges for the devices that share a network to take turns sending packets. The time required to send a packet is measured in millionths of a second.

9.10 Sharing Occurs On Demand

Imagine a network with one hundred devices attached. What happens if only one device has packets ready to send? After sending a packet, does the network hardware check each of the ninety-nine computers, find they have nothing to send, and then allow the first computer to send another packet? No. In most networks, only devices that have something ready to send take turns sending. If only one device has packets to send, the device can use the network continuously. If two devices have packets ready to send, they alternate.

The sharing scheme allows a device to participate or stop participating at any time. Once a device has sent its last packet, the device stops using the network. Thus, at any time a device receives an equal share of the network with all other devices that are using the network. If only two devices are actively sharing a network, each will send one-half of the total packets. If three devices are actively sending packets, each device will send one-third of the packets, and so on.

Network sharing is completely automatic because network hardware handles all the details. The hardware does not need to know how many devices are using the network simultaneously, but instead uses a method that allows all active devices to "contend" for access. The key point is:

A packet switching system allows devices to start or stop sending packets at any time. Each device that has packets to send receives a "fair share" of network resources because the hardware is arranged so that the devices with packets to send take turns sending packets.

9.11 Relevance To The Internet

Like most computer networks, the Internet is a packet switching system. Packet switching allows many communications to proceed across the Internet simultaneously without requiring one user to wait for another user to finish their communication. As a consequence of the technology, whenever a user transfers a data item across the Internet, network software on the sending machine must divide the data into packets, and network software on the receiving machine must reconstruct packets to produce the original item. For example, a photo must be divided into packets for transfer across the Internet, and then reassembled into a copy of the photo at the receiving side. To summarize:

> *All data is transferred across the Internet in packets. A sender divides a message or document into packets and transfers the packets across the Internet. A receiver reassembles the original message from the packets that arrive. Packets from many machines traverse the Internet at the same time.*

9.12 Summary

The fundamental technique that computer networks use to ensure fair access to shared network resources is known as packet switching. Before data can be transferred across a packet switching network, the data must be divided into individual packets. A typical packet can hold up to fifteen hundred characters (bytes) of data.

Every device is assigned a unique number known as the device's *address*. Each packet contains a header that specifies the address of the device that sent the packet and the address of the device to which the packet is sent. Devices that share access to a network take turns sending packets. On each turn, a given device can send one packet.

EXERCISES

9.1 If you have DSL or cable Internet service, look at the label on the bottom of your DSL or cable modem and find its MAC address. Hint: a MAC address consists of twelve characters, including digits 0 through 9 and letters A through F.

9.2 Sue and Paul are sitting in a coffee shop using the Internet when suddenly everything stops working. After ten seconds, Paul says that he heard that devices on the Internet use packets, and they are probably blocked waiting because someone else is sending a packet. Can Paul be correct? Explain.

9.3 Think about a coffee shop. Name five types of devices (either that customers have or the coffee shop has) that use the Internet.

9.4 Suppose that every individual on earth has both a laptop and a smart phone, and suppose each device needs a unique address. How many total addresses would be needed? (Hint: search online to find the world's population.)

9.5 On a police drama when a criminal is accessing a victim's computer, the script has the police capture data and use the data to identify the criminal's computer. What information in a packet can be used to identify the sending computer?

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

10

Internet: A Network Of Networks

10.1 Introduction

The previous chapter describes packet switching and shows why dividing long messages into packets allows computers to share a transmission path without introducing arbitrarily long delays. This chapter describes how multiple packet switching networks can be interconnected to form an Internet that functions like a single, large network.

10.2 Building A Global Network

How can we build a global communication system? As Chapter 7 points out, no single technology suffices for all purposes. Many packet switching technologies exist because each has been designed to meet constraints of speed, distance, and cost. So, the question becomes: can we allow groups to choose a network technology that meets the group's needs, but still be able to connect multiple networks together? Unfortunately the simplest approach — connect the wires from one network to the wires of another — doesn't work. To understand why, one must know that designers do not make all technologies compatible. Details, such as the electrical voltages used and the maximum size of a packet often differ. In fact, directly connecting wires can permanently damage the hardware. Consequently, when they designed the Internet, researchers did not envision merely connecting wires of incompatible networks.

How can we cope with incompatible hardware? The Internet uses an approach that allows each group to select the network technology that best meets the group's needs, and manages to interconnect completely incompatible networks safely. The next sections explain how the Internet makes such an interconnection possible.

10.3 Two Fundamental Concepts

Two basic ideas will help explain the technology the Internet uses to connect networks:

- The structure of network interface hardware
- Multiple interfaces

The structure of network interface hardware. Recall from Chapter 6 that to connect to a network, a device uses a piece of hardware known as a *network adapter* or *network interface*. The terms *interface* and *adapter* were both chosen to indicate that the hardware is designed with two electronic circuits that operate with differing signals. One circuit is *internal facing*, and the other is *external facing*. The internal circuit communicates with the processor. That is, an internal facing circuit connects to the device, and uses the same signals as other components in the device. The external facing circuit uses the electrical signals that the network hardware uses. An adapter allows the two parts to work together despite using different electrical signals (e.g., a device that uses 3 volts can attach to a network that uses 5 volts). Figure 10.1 illustrates the idea.

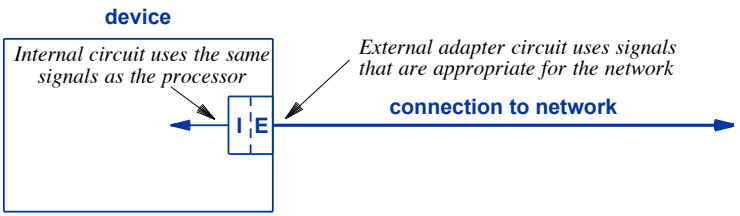


Figure 10.1 Illustration of a network adapter in a computer that connects the computer to a network.

You do not need to understand hardware details; you just need to appreciate that an adapter keeps the electrical signals used on the network electrically isolated from the electrical signals used in the computer. To summarize:

> *A network adapter keeps signals used externally separate from signals used internally.*

Multiple interfaces. The second idea that will help explain Internet technology is straightforward: a given device can connect to two or more networks. To do so, the computer must have a network adapter circuit for each network. Many users are already aware that multiple connections are possible because they have a device that has a Wi-Fi adapter as well as an adapter for a cellular network. Users may also have seen a laptop that includes both an Ethernet adapter and a Wi-Fi adapter.

From the Internet perspective, a more interesting case occurs when a device has adapters for two wired networks. The structure of an adapter described above explains why such a configuration is possible: each adapter isolates the electrical signals used on a network from the signals used inside the computer. Thus, a device can connect to two networks with incompatible electrical signals because the adapters keep the signals for the networks isolated from each other and from the signals inside the computer. Figure 10.2 illustrates the idea.

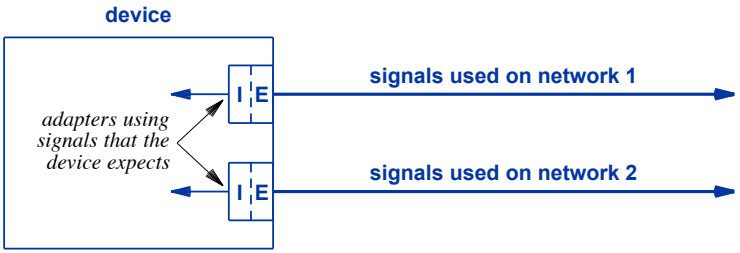


Figure 10.2 A device with two network adapters that keep the signals used on the two networks isolated from each other.

10.4 Using A Specialized Computer To Interconnect Networks

The ideas above explain one of the most fundamental pieces of the Internet: we can use a specialized computer to connect dissimilar networks together. Figure 10.3 illustrates the idea.

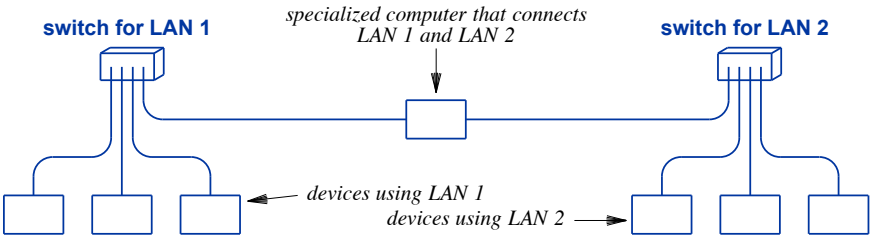


Figure 10.3 Illustration of a specialized computer used to interconnect networks.

As Figure 10.4 shows, a computer used to connect the two networks attaches to each network (i.e., each switch). What makes the interconnection so useful? In terms of hardware, a computer used to interconnect networks resembles a conventional computer. It has a processor, memory, and network adapters. We will learn, however, that there is one major difference: the software used. Instead of running conventional desktop applications, computers used to interconnect networks have specialized software that forwards packets from one network to another. For example, an interconnecting computer does not run word processing or spreadsheet applications. Instead, such a computer has only special-purpose software that performs tasks related to the job of interconnecting networks.

One more detail helps explain the interconnection technology: the computer systems that provides interconnection uses multiple networks simultaneously. A typical user only connects their device to one network at a given time. If more than one network is available, an operating system chooses one and ignores the other. For example, if a laptop finds both a Wi-Fi network and an Ethernet, the operating system will usually choose the Ethernet (or ask the user to choose). However, it is possible to connect to multiple networks simultaneously, and be able to send and receive packets on any of them.

10.5 Internet Terminology: Routers And Hosts

A specialized computer that interconnects networks has one major task to perform: when a device on one network sends packets to a device on another network, the packets must be sent from the first network to the second. A specialized computer that interconnects networks and performs the task of forwarding packets among them is known as a *router*. We use the term *host* for any other device that attaches to the Internet and is not a router. Thus, each of a user's Internet devices is a host, including desktops, laptops, smart phones, and devices like smart home climate control systems that can be accessed over the Internet. In Figure 10.3, the specialized computer in the center that interconnects networks is a router, and other devices are hosts.

The next chapter provides more details about how routers forward packets, but the idea is straightforward: a router receives a packet sent to it across one network, and the router sends the packet on to its destination across another network. For example, in Figure 10.3, if a device attached to LAN 1 sends a packet to a device attached to LAN 2, the packet is sent through LAN switch 1 to the router that interconnects the networks. The router then forwards the packet by sending it through LAN switch 2 to the destination on LAN 2.

Software on a router needs to know the network to which each computer connects so it can determine where to send packets. In the case of two networks, the decision is straightforward — when a packet arrives over one network, the packet must be sent over the other network. In the case of a router that interconnects three or more networks, however, the decision is more complex because the router must choose the correct network. We use the term *routing* to describe the process of finding all the pos-

sible destinations in the Internet and selecting a path over which to reach them; the name *router* was chosen to emphasize that the device must understand routing. We can summarize:

> *The Internet uses specialized computer systems called* routers *to interconnect networks. A router has multiple network adapters, and runs software dedicated to the task of forwarding packets among the networks.*

10.6 Building A Large Virtual Network

When a user thinks of the Internet, they envision a single, giant network to which many computers attach. Network professionals often draw diagrams in which a cloud denotes a network, and a geometric shape denotes a device attached to the network. Figure 10.4 uses a cloud diagram to illustrate a user's view of the Internet. Remember that a host can be any device, including a laptop, smart phone, desktop, Internet TV, or smart device.

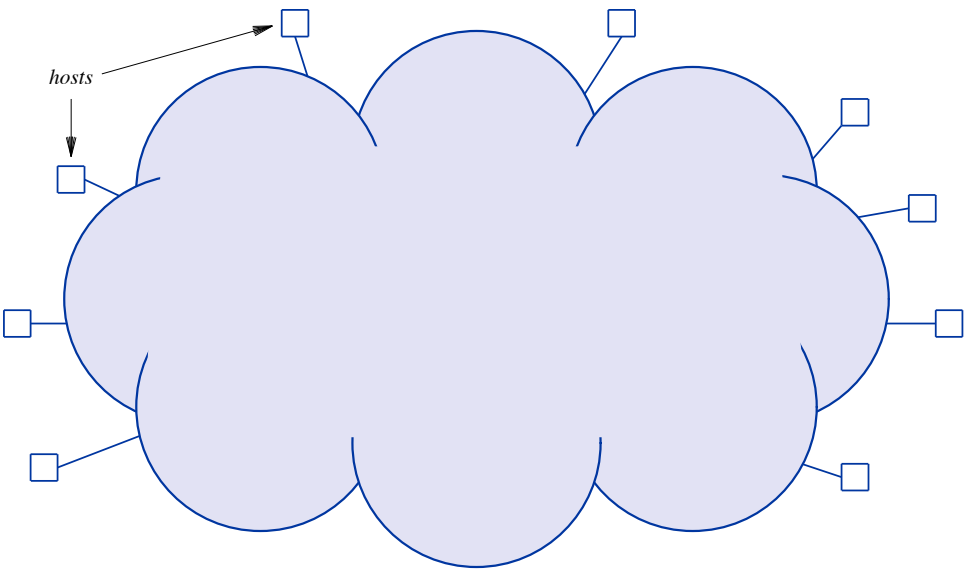


Figure 10.4 The user's view of the Internet: a single large network to which various types of hosts (users' computers and devices) connect.

Although users think of the Internet as a single giant network, it is not. Instead, the Internet uses a *network of networks* approach in which thousands of computer networks are interconnected by routers. Figure 10.5 illustrates the internal structure of the Internet.

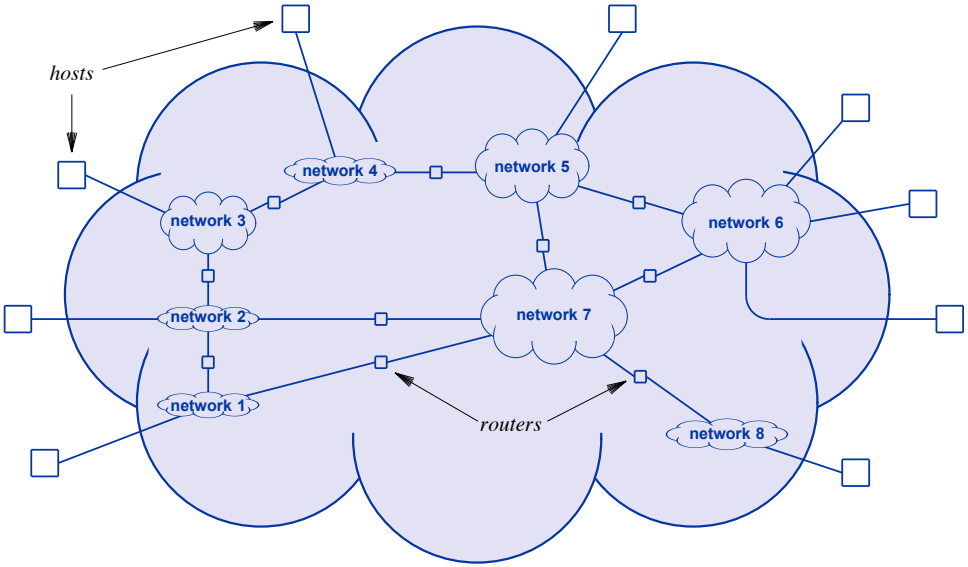


Figure 10.5 Illustration of the internal structure of the Internet with many networks interconnected by routers.

As the figure shows, each host attaches to one of the individual networks. Software on the routers allows a host to exchange packets with any other host. Thus, when a computer on one network communicates with a computer on another network, it sends packets through a router. Software on routers gives the appearance of a single, unified communication system. To summarize:

> *The Internet is not a single computer network. Instead, it consists of hundreds of thousands of computer networks interconnected by routers.*

10.7 The Internet Includes Multiple Types Of Networks

Because a given router can interconnect networks that use different hardware technologies, the router architecture permits the Internet to accommodate multiple types of networks. The figure uses clouds of differing sizes and shapes to illustrate multiple network types.

Whenever a group installs a network, the group can choose a network technology that is best suited to their purpose. The group then uses a router to connect their network to the rest of the Internet. In particular, a given network can be wired or wireless, a LAN or WAN, and can have many hosts attached or a few hosts attached. In fact, a network used merely to connect other networks may not have any hosts attached (e.g., network 7 in the figure).

Connecting multiple types of networks is important for two reasons. First, because the Internet connects many organizations of diverse size, networking needs, and budgets, the organizations use diverse network technologies. Second, and more important, networking technologies continue to change. Many of the hardware technologies currently used in the Internet did not exist when the Internet began, and many technologies in use now will be replaced in the future. For example, Wi-Fi had not been invented when the Internet was created. Using a network-of-networks approach keeps the Internet extremely flexible by allowing any given network to be upgraded at any time. The Internet would not have survived if a single technology had been used.

10.8 Ownership, ISPs, And Transit Traffic

Figure 10.5 gives a simplified view of an Internet composed of individual networks and routers. In practice, entities own and operate groups of networks and routers. For example, a large organization that connects to the Internet, such as a company or a university, may own and operate dozens or even hundreds of networks and routers.

A key part of the Internet structure arises because a set of *Internet Service Providers* (*ISPs*) own and operate networks in the middle of the Internet. Service providers are in the business of providing *transit*. That is, an ISP agrees to accept incoming packets and send the packets on to their destinations. Networking professionals use the term *traffic* to refer to packets moving across the Internet, and say that an ISP handles *transit traffic*. As the term implies, transit traffic moves across an ISP, but the packets neither originate on a host owned by the ISP nor are the packets destined for a computer owned by the ISP. Instead, each packet enters on one side of the ISP and is forwarded out another side. Look again at Figure 10.5. The network labeled *net 7* does not have any host attached to it. Therefore, all the traffic that crosses the network is transit traffic that has been sent from one router to another.

10.9 A Hierarchy Of ISPs

Because an ISP uses the network-of-networks approach, a given ISP can own multiple networks connected by routers. Even the smallest ISPs often use multiple networks. For example, an ISP that provides service to a small town may not run separate wires to each customer. Instead, the ISP may choose to place a box in each neighborhood that contains a network switch, and run wires from each customer to the nearest neighborhood box. The ISP then runs wires from the curb-side boxes to a central location, and uses routers to interconnect the networks.

A larger ISP, informally called a *regional ISP*, may have sites in multiple cities that are connected by a WAN. The ISP may divide each city into separate areas, with one or more networks in each area. The largest ISPs, informally called *backbone providers*, build giant Wide Area Networks (called *backbones*) that span an entire continent or multiple continents. A site on a backbone network typically connects to a set of regional networks.

A small ISP can deliver a packet directly if the packet is sent from one of its customers to another customer. To handle other packets, the small ISP must have a connection to a regional ISP. If two computers in a region send to one another, the regional ISP can handle forwarding. Finally, each regional ISP must have a connection to a backbone provider to handle packets sent to or received from computers outside the region.

Formally, we use the term *Tier 1 ISPs* to refer to the largest ISPs that provide connectivity across continents. Tier 1 providers are said to form the *core* of the Internet. We use the term *Tier 2 ISPs* to refer to an intermediate size ISP that operates a *regional network* covering a geographic region. Finally, smaller ISPs that provide connections to individual businesses or residences within a region are classified as *Tier 3 ISPs*. The network technologies used at each tier provide the capacity appropriate for the tier. Tier 1 providers use the most powerful (and most expensive) networks and routers for their backbone WANs; the equipment used in Tier 2 and 3 is significantly less expensive and significantly less powerful.

How is the Internet really structured? It is not just a set of networks and routers connected in arbitrary ways. Instead, the ISPs form a hierarchy with Tier 1 at the top, Tier 2 ISPs in the middle, and Tier 3 ISPs at the bottom. Some networking professionals add an additional level, Tier 4, to refer to extremely small ISPs that are also called *mom-and-pop* ISPs. Figure 10.6 illustrates the overall hierarchy of ISPs.

In the figure, each ISP is denoted by a cloud to indicate that the ISP may contain multiple networks connected by routers. Boxes at the bottom of the diagram correspond to residential or business subscribers. As the figure indicates, some customers will choose to pay more for a connection to a regional ISP instead of a local ISP. In some cases, a local ISP does not have sufficient capacity to handle the traffic a business generates. The largest business customers may choose to pay for a connection that goes directly to a Tier 1 provider. Such direct connections are extremely expensive, but guarantee that the customer's traffic will go directly to a higher speed ISP network without passing through smaller, intermediate ISPs first.

To summarize:

> *Routers and networks in the Internet are owned and operated by ISPs. Most subscribers connect to a local (Tier 3) ISP, which connects to a regional (Tier 2) ISP, which connects to a backbone (Tier 1) ISP. A business can pay more for a connection to a higher tier.*

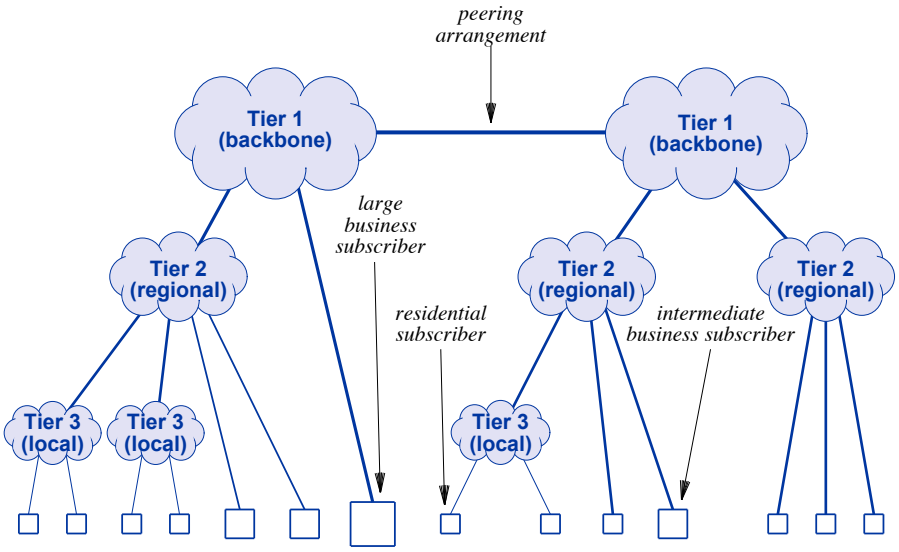


Figure 10.6 Illustration of the hierarchy of ISPs used in the Internet.

10.10 Peering Arrangements At The Center Of The Internet

To guarantee universal service (i.e., that a packet can travel from any source to any destination), all parts of the Internet must be connected. In particular, Tier 1 ISPs that constitute the core of the Internet must also be connected. Technically, the connection between two Tier 1 providers works like the connections between any pair of ISPs: a router connects between a network in one ISP and a network in another. Of course, the routers and communication mechanisms used to connect Tier 1 providers operate at much higher speed than other routers and connections.

Politically and economically, the connection between two Tier 1 providers differs from all other connections. In Chapter 33, we will learn that a Tier 2 provider is a customer of a Tier 1 provider, and a Tier 3 provider is a customer of a Tier 2 provider. However, two Tier 1 providers are said to be *peers* in the sense that they each have approximately the same number and type of customers beneath them. Consequently, the connection between them is called a *peering arrangement*, as the figure illustrates.

Peering among Tier 1 ISPs is reserved for the largest ISPs around the world, and the set of Tier 1 peers are said to form the *core* of the Internet.

10.11 An Example Trip Through The Internet

To understand the hierarchy, consider two subscribers that are geographically distant. Assume that each subscriber connects to a local ISP in his or her area, and consider what happens when they communicate. A packet sent from one to the other travels from the sender's computer across the connection to the sender's local ISP. The packet may go through one or more networks and routers in the local ISP, and then the local ISP forwards the packet to a regional ISP. The regional ISP may have multiple networks that the packet must cross before being forwarded to a Tier 1 provider. If the destination is reachable through the same Tier 1 provider, the packet travels across the backbone to a regional ISP near the destination. If the destination is connected to another Tier 1 provider, the packet passes across a peering exchange point, and then across the second Tier 1 backbone. Once a packet reaches a regional ISP near the destination, it may cross one or more networks in the regional ISP before being forward to the local ISP of the receiver. The local ISP delivers the packet over the connection to the subscriber. So, the packet does indeed make a series of steps, network to router to network to router, and so on. However, because groups of the networks and routers are owned by ISPs, we can say the packet went from ISP to ISP.

The point is:

> *The Internet's network of networks is arranged in a hierarchy with major Tier 1 ISPs that provide backbone networks at the top, regional ISPs in the middle, and local ISPs at lower levels.*

Our description of the ISP hierarchy is somewhat simplified. In practice, a few major ISP companies offer services at more than one tier. For example, a company that offers service to individual residential subscribers may also operate a Tier 1 backbone service. In addition, two ISPs may agree to interconnect directly if they find it economically advantageous, even if the connection does not follow the strict ISP hierarchy.

10.12 The Internet Approach Revolutionized Networking

The idea of using routers to interconnect networks may not seem startling or revolutionary, but it was. Before Internet technology appeared, a company that wanted to use computer networking either needed to choose one network technology for all their computers or had to live with multiple, independent networks that could not be connected. The only way to create a communication system that allowed any computer to communicate with any other computer consisted of choosing a single technology, and most

organizations quickly realized that one size does not fit all. The Internet approach — a network of networks — enables choice. On the one hand, each group can choose a network technology that is appropriate for their needs. On the other hand, routers allow any computer to communicate with any other computer.

The Internet approach also enables technology evolution. Unlike a system in which a single network technology is used, the Internet can evolve slowly, without changing everything at once. In essence, the Internet's ability to accommodate heterogeneous network technologies means new versions of hardware can be phased in — any individual network in the Internet can be upgraded or replaced at any time without affecting other networks.

10.13 Summary

Although to a user it appears to be a single, large network, the Internet consists of thousands of computer networks interconnected by dedicated devices called routers. Because a router can interconnect networks that use different technologies, a router can connect a wired network to a wireless network, a LAN to another LAN, a LAN to a WAN, or a WAN to another WAN. Because it is made up of networks interconnected by routers, we refer to the Internet as a network of networks.

Networks and routers in the Internet are owned by Internet Service Providers (ISPs) that are arranged in a conceptual hierarchy. Tier 1 ISPs that can span continents form the core of the Internet. Tier 2 (regional) ISPs form the next level of the hierarchy, and each Tier 2 network covers one region. At the lowest level, Tier 3 (local) ISPs provide connections to individual subscribers. When two Tier 1 ISPs agree to exchange traffic, they are said to be peers.

Using routers to interconnect networks produced a revolution. The approach permits connections among multiple types of networks, allows each group in an organization to choose a network technology that best suits the group's needs and budget, and allows any network to be upgraded or replaced without changing the rest of the Internet.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

11 Internet Access Using Broadband And Wireless

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

11

Internet Access Using Broadband And Wireless

11.1 Introduction

The previous chapter described the Internet as a network of networks. For most individuals, however, the main question is not how the Internet is built, but how they can connect and use it. This chapter describes technologies used for Internet connections, and explains what each provides.

11.2 Access Technologies For The Last Mile

The question arises, what technology provides the best connection between a customer and an ISP? The answer depends on the physical distance spanned, the data rate the customer desires, and other factors, such as whether the location is rural or metropolitan. Industry professionals use the term *last mile* to refer to such connections, even though the distance is often greater than a mile. To emphasize that they are designed to provide Internet access, technologies used for the last mile are known as *access technologies*, and include:

- Dial-up
- Leased data circuit (copper and optical fiber)
- Digital Subscriber Line (DSL)
- Cable modem
- Wi-Fi
- 4G and 5G wireless

11.3 Dial-up Internet Access

Early Internet Service Providers offered *dial-up* access. The mechanism is straightforward: a *dial-up modem* connects between a computer and an analog voice telephone system. The modem plugs in just like a telephone. The computer can instruct the modem to go *off hook* (i.e., emulate picking up a handset), detect a dial tone, dial a specified number, and wait for the other end to answer. An ISP gives each user software that controls the modem and dials a special phone number at the ISP. The special number doesn't go to telephones, but to a set of modems. Whenever the number is called, one of the modems answers. The user's modem communicates with the modem at the ISP, and to send data, the two modems modulate audible tones. Figure 11.1 illustrates the technology.

Figure 11.1 Illustration of the equipment used for dial-up Internet access. A modem attached to a user's computer calls a modem at the ISP.

11.4 Narrowband And Broadband Access

Dial-up Internet access has the advantage of using relatively inexpensive equipment (a dial-up modem), and being trivial to install (a user simply plugs the modem into a standard landline telephone outlet). However, dial-up is classified as a *narrowband* technology, which means that it transfers data slowly. Early modems operated at 300 bits per second or slower; by the late 1990s, techniques had been invented that enabled transfers of up to 56,000 bits per second. Narrowband technology means a user must wait a long time for even the simplest request (e.g., a web page to appear).

Starting in the 1990s, a set of technologies appeared that could transfer millions of bits per second. The technologies are classified as *broadband* technologies to emphasize that they can transfer data significantly faster than narrowband technologies. Once broadband service became available, users immediately saw the advantage and switched from dial-up to a broadband service. Although still used when broadband is not available, dial-up access has largely been replaced. We can summarize:

> *Early Internet access used narrowband dial-up technologies; most users abandoned dial-up when faster broadband technologies became available.*

11.5 Leased Data Circuit Access

Until the mid-1990s, only one broadband access technology was available: a digital circuit leased from a common carrier. In essence, leasing a circuit means renting unused wires in cables that were originally set up for phone service, and then placing a modem at each end. Leased circuits are classified as *point to point* because the circuit starts at a specified geographic location and runs to another geographic location. Leased circuits are also called *dedicated circuits* because a leased circuit is not shared with other users (i.e., a customer who rents the circuit has exclusive access, even if the circuit crosses a long distance and the path runs through many intermediate cities).

Leased circuits are expensive. A customer must pay an initial fee to have a circuit installed, and then must pay a monthly fee for use of the circuit. The fee depends on the distance and the data rate the circuit supports, but fees are much higher than other broadband access technologies. Consequently, only large business customers or ISPs lease circuits. In terms of data rates, a wide range is available. In the U.S., for example, a *T1* can transfer 1.54 million bits per second; an *OC-192* circuit can transfer 10 billion bits per second.†

11.6 Digital Subscriber Line (DSL) Access

Before the Internet first emerged, the telephone industry had wiring in place to most homes and businesses. Phone companies investigated ways their landline phone wiring could be used to transfer digital data. Researchers found ways to send digital data over the phone wires that were designed to carry analog voice signals, and devised a set of technologies that varied in the data rates available and the distance spanned. Known as *Digital Subscriber Line* (*DSL*) technologies, the technologies are classified as broadband because they offer much higher data rates than dial-up. The actual data rate DSL can deliver depends on the distance between a subscriber and the telephone switching center. However, download data rates of over two million bits per second are common, over thirty times more bits per second than dial-up.

The most interesting aspect of DSL is that when filters are used, data and voice calls can pass over the same wires simultaneously without interfering with one another. Figure 11.2 illustrates the equipment used at the subscriber's site and the phone company.

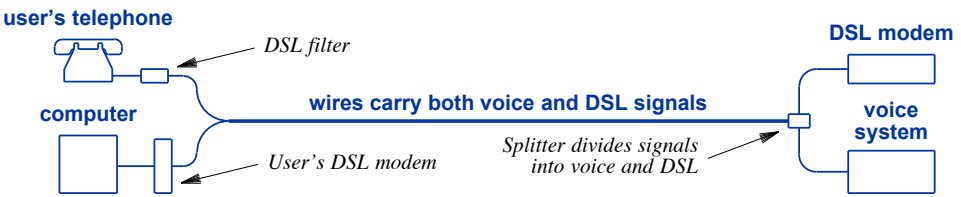


Figure 11.2 Illustration of the equipment used with DSL access.

†Higher speed circuits use optical fiber rather than copper wires.

As Figure 11.2 shows, only one set of wires connects a user with the telephone switching center. Technically, a *DSL modem* sends data using higher frequencies than a voice call. At the switching center, a small device known as a *splitter* separates the signals so that voice signals can be sent to the voice telephone system and DSL signals can be sent to a DSL modem that connects to the Internet. At the user's location, a splitter (also known as a *DSL filter*) prevents a user's telephone from generating random signals that could interfere with DSL. The point is:

> *Although it uses the same wiring as a conventional landline telephone, DSL offers much higher data transfer rates than dial-up access. Furthermore, splitters allow conventional telephone calls and data transfer to proceed simultaneously without interference.*

Several variants of DSL technology are available. Residential Internet customers use a specific form known as *Asymmetric Digital Subscriber Line* (*ADSL*). The asymmetry arises because most users tend to receive more data than they send, and ADSL meets the need because ADSL is designed to transfer more from the ISP to the customer than from the customer to the ISP. In fact, the ADSL variant has become so dominant that phone companies have stopped advertising ADSL service, and simply use the term *DSL*.

11.7 Cable Modem Access

Before the Internet became popular, the cable television industry already had wiring in place to most homes. The coaxial cable the industry used was originally designed to deliver analog video signals, and the hardware was only designed to send information in one direction (from the cable company to the subscriber). Researchers devised a modification of the cable infrastructure that made it possible to transfer Internet data over a cable that simultaneously sends video signals. Although the cable industry has a technical name for the standards, the technology has become known to the public as *cable modem* access technology to emphasize that a user needs a special modem. Figure 11.3 illustrates the equipment used.

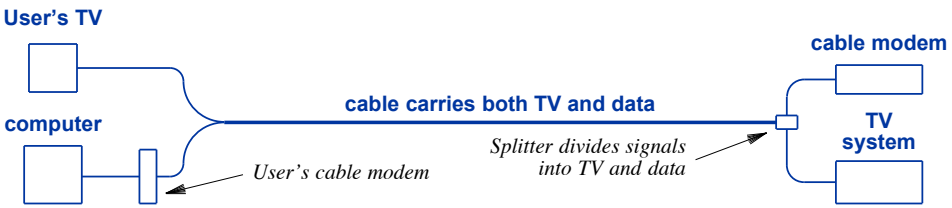


Figure 11.3 Illustration of the equipment used with cable modem access.

At the subscriber's site, a cable modem connects between the user's computer and a cable TV outlet, using Ethernet for the connection between the modem and the computer. The cable company also has a cable modem at its end. The modem at the subscriber's location communicates with the modem at the cable center by sending signals over the coaxial cable that connects the locations. Analogous to DSL modems, cable modems are designed so that data transmission does not interfere with television signals, making it possible to send data over the existing wiring at the same time as cable television signals.

In terms of data transfer, cable modems can transfer more bits per second than even DSL. Initially, cable modem access offered rates of six or eight million bits per second. Now, major providers offer Gigabits per second (billions of bits per second).

11.8 Wireless Access Technologies

It may seem that DSL and cable modem technologies solve the last mile problem completely. However, each has limitations. DSL has a distance limitation that prevents the signals from traveling as far as conventional telephone signals. Thus, it cannot be used to provide service in rural areas. Furthermore, both DSL and cable modem technologies can only be used where physical wires can be run. For example, communication wires may not extend to a cabin in the woods, even if the cabin is only a short distance from a town.

To provide Internet access to such locations, engineers have *wireless access technologies*. Although all wireless networks use radio waves to carry data, a wide variety of technologies have been developed. Some are point to point, meaning that special antennas are used to aim the transmissions in a straight line between two communicating sites (e.g., between a remote residence and an ISP), and some use antennas that broadcast in all directions (e.g., between a transmitter and a set of houses that are close by). Others use a satellite orbiting the earth to relay data between subscribers at arbitrary locations and an ISP. The point is:

> *A variety of wireless access technologies has been developed to meet various needs.*

We have already seen one technology that is used for Internet access: *Wi-Fi*. Recall from Chapter 6 that Wi-Fi is classified as a Local Area Network because Wi-Fi can only reach computers in a small area around an access point (e.g., inside a house or inside a store). Many organizations use Wi-Fi to provide Internet access to customers. For example, hotels, coffee shops, airports, and shopping malls often provide Wi-Fi access, as do high schools, universities, and hospitals.

To provide Wi-Fi Internet access, an owner needs two things; an Internet connection and a Wi-Fi device. The Internet connection can use any of the technologies described above. For example, a national coffee shop chain uses a leased line to con-

nect each coffee shop to the Internet; other small retail stores obtain cable modem service from a local ISP. Once Internet access is in place, only one additional piece of equipment is needed to offer Wi-Fi service to others. The device is known as a *wireless access point*. Vendors who sell such devices to consumers prefer the name *wireless router*; the terminology is appropriate because the device connects the Ethernet network on the modem to the Wi-Fi network, and forwards packets between the two. In essence, the device performs the same function as other Internet routers. Figure 11.4 illustrates the equipment needed to provide Wi-Fi access.

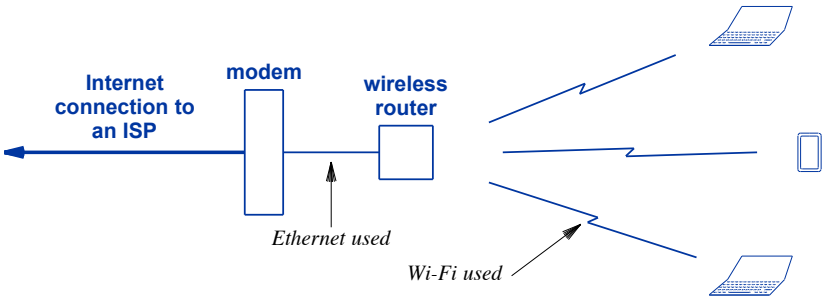


Figure 11.4 Illustration of the equipment used to provide Wi-Fi access.

In the figure, the modem can be a leased line modem, DSL modem, cellular modem, or cable modem, depending on the technology the user selects for their Internet service. The connection between the modem and wireless router typically uses an Ethernet cable, which is the same cable used to connect a modem to a computer. The connection between the wireless router and the user's devices uses the standard Wi-Fi paradigm — the wireless router must be configured to have an SSID, and devices must use the same SSID. In Chapter 17, we will learn that from the ISP's perspective, a wireless router acts like a computer; the ISP does not know it is a router or that the router is using Wi-Fi to provide Internet service to devices.

11.9 Cellular Wireless Access (4G and 5G)

Although it works well inside a coffee shop or a home, Wi-Fi cannot cross a large geographic distance because the technology has limited range. One approach to longer-distance wireless access uses the cellular phone system standards for *4th-Generation* or *5th-Generation* wireless technology (*4G or 5G wireless*). To understand cellular access networks, it is necessary to know that the cellular system is no longer limited to transmission of voice phone calls. Instead, Internet communication is now an integral part of the 4G and 5G systems. Thus, when a user who subscribes to a 4G service powers on a smart phone, the phone contacts a cell tower and is given the ability to transfer packets across the Internet.

As described above, the cellular system provides Internet access to each individual smart phone. However, using cellular as an Internet access technology also means using a cellular connection to provide access to conventional computers, such as desktops and laptops. That is, cellular access replaces DSL or cable modem access.

To use the cellular system to provide Internet access, one merely uses a cellular modem instead of a DSL or cable modem. Two versions are available:

- Mobile broadband modem
- Fixed cellular modem

11.9.1 Mobile Cellular Modems (4G or 5G)

A *mobile broadband modem*, which is available with both 4G and 5G, consists of a small device that is designed to be easy to carry when traveling. For example modems exist that are less than four inches long and a couple of inches wide, small enough to carry in one's pocket. Mobile modems can be battery-powered, which permits their use without requiring a power cord.

A cellular modem contains two basic circuits: one that acts like a cell phone, and one that connects to a device. When powered on, the circuit that acts like a cell phone contacts the nearest cell tower. Assuming the provider recognizes the modem (i.e., the subscriber has signed up for service and the account is in good standing), the cell tower responds and agrees to forward Internet traffic.

The second circuit in a cellular modem provides connections to one or more devices. Most cellular modems use Wi-Fi for device connections. Thus, from the user's point of view, a cellular modem acts like a wireless router. Some vendors advertise their cellular modem as a "portable Wi-Fi hotspot." Figure 11.5 illustrates how a cellular modem provides Internet access to a set of laptops.

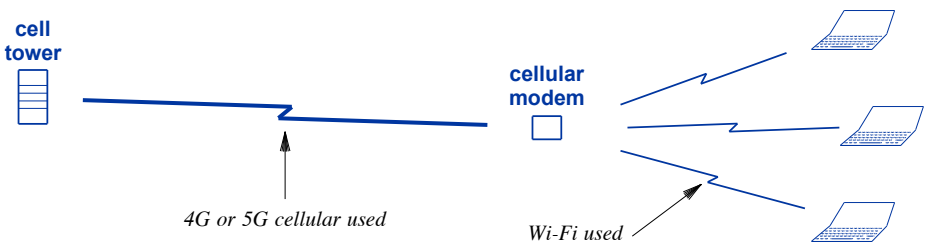


Figure 11.5 Illustration of a portable cellular modem (portable Wi-Fi hotspot) that uses Wi-Fi to connect devices to the Internet.

11.9.2 Fixed Cellular Modems

Recent cellular standards have introduced an additional type of cellular modem intended to be installed permanently. Known as a *fixed cellular modem*, the unit is somewhat larger than the mobile versions, and requires a power cord. The motivation for fixed cellular modems is straightforward: they are intended to replace DSL and cable modems. That is, instead of carrying the modem when traveling, a user installs the modem in their residence, connects one or more devices, and uses the modem for all their Internet access.

In addition to providing Wi-Fi connections for devices, fixed cellular modems offer wired connections using Ethernet. Thus, a fixed cellular modem provides exactly the same connections as a DSL or cable modem. We can summarize:

> *A cellular modem connects to the 4G or 5G cellular service and uses the connection to provide wireless Internet access to computers. Portable units can be carried while traveling. Recent cellular systems offer fixed cellular modems that can be installed in place of a DSL or cable modem.*

11.10 Summary

An access technology provides a connection between a user and an ISP (the so-called "last mile"). Both wired and wireless technologies are available. Early ISPs offered dial-up access, a narrowband service. Newer technologies, such as DSL and cable modems, provide broadband service, which means they transfer data at a higher rate. A wireless router provides local wireless access using Wi-Fi.

Recent standards for the cellular phone system include Internet data transport. Consequently, a cellular modem can be used as an Internet access technology; both small, portable devices and larger fixed devices are available.

EXERCISES

11.1 Take a poll of friends and find out which access technologies they have used: dial-up, DSL, cable modem, mobile broadband, or other.

11.2 ISPs that offer DSL or cable modem service often allow a user to choose a *self-install* option to save money. If an ISP gave you a modem, what would you need to do to install it?

11.3 Contact local ISPs to find out how much their service costs and how many bits per second they will transfer. (Compare their regular rates, ignoring introductory discounts.) Which type of service is the best financially?

11.4 Suppose you have a choice between wired and wireless Internet services (i.e., DSL or cable modem vs. cellular). What advantages does wireless offer, if any?

Chapter Contents

12 Internet Performance

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

12

Internet Performance

12.1 Introduction

Chapter 10 explains that the Internet is a network of networks, created by using routers to interconnect all the networks, and Chapter 11 explains access technologies that can be used to connect a computer to the Internet. This chapter uses the concepts from the two chapters to explain Internet performance. It explains why ISPs advertise *speed*, and answers fundamental questions users often raise about the speed of their Internet service.

12.2 Network Speed

When they advertise their services, ISPs often use the term *network speed* and claim that they are selling a *high-speed* network or a *faster* network. If we think of everyday life, higher speed means something moves along at a higher velocity. For example, a car traveling at a higher speed than another car moves down the highway faster than the other car. For computer networks, the term *speed* is misleading because packets do not move faster — they always travel across wires, optical fibers, and through space at the same rate: the speed of light.† The point is:

> *An ISP cannot make packets travel across a network faster than they do because the ISP cannot change the laws of physics to send signals faster than the speed of light.*

†To be scientifically accurate, we should say *approximately at the speed of light*, but only physicists care about the distinction.

12.3 What Does Speed Mean?

What is an ISP advertising when it advertises higher speed? An ISP is not talking about how fast a packet travels over a wire or through the air. Instead, the ISP is talking about *network capacity*,† and is advertising higher capacity technology. The capacity of a network specifies how many bits can be sent over the network per second. Modern network technologies send over a million bits per second, so ISPs report numbers as multiples of millions or billions. Rather than using English measures, such as *millions of bits per second*, ISPs have adopted the metric prefixes used in the scientific community, and state *Megabits per second* (*Mbps*) or *Gigabits per second* (*Gbps*). Figure 12.1 lists the prefixes and their English equivalent.

Prefix	English	Multiplier	Decimal Places
Kilo	Thousand	1,000	3
Mega	Million	1,000,000	6
Giga	Billion	1,000,000,000	9
Tera	Trillion	1,000,000,000,000	12
Peta	Quadrillion	1,000,000,000,000,000	15

Figure 12.1 Prefixes that ISPs use when referring to network performance and data sizes.

You are only likely to encounter the middle three prefixes. It is easy to remember that *Mega* means *million* because they both start with the letter *m*. Similarly, *Tera* and *trillion* both start with the letter *t*. That leaves *Giga* and *billion*. The main thing to remember is that *Giga* is one thousand times *Mega* in the same way that a billion is one thousand times a million. Therefore, a network that has a capacity of one *Gigabit per second* has a thousand times more capacity than a network with a capacity of one *Megabit per second*.

12.4 Brick Delivery

What does more capacity mean for a user? To understand, think of an analogy. Suppose a construction crew is erecting a large brick building, and suppose the brick factory is a few miles away, with a road leading directly from the factory to the construction site, as Figure 12.2 illustrates. To transport bricks, a truck will start at the brick factory, load the truck, drive down the road, unload the bricks, and return for the next load. How fast can all the bricks be delivered? One way to measure the delivery consists of counting the number of truckloads of bricks that arrive over a given time. For example, let's suppose one truck takes sixty minutes to deliver a load and return to the factory. The delivery rate will be one truckload per hour. If two trucks are used, the delivery rate will be two truckloads per hour, and for five trucks, the rate will be five truckloads per hour.

†ISPs incorrectly use the term *bandwidth* to describe capacity; the correct technical term for capacity is *throughput*, but it will be less confusing if we avoid such terms.

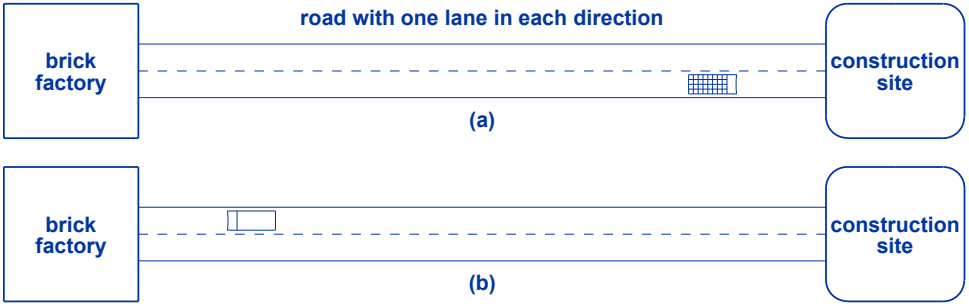


Figure 12.2 Illustration of a road between a brick factory and construction site with (a) a truck carrying bricks as it nears the construction site and (b) the empty truck returning to the factory.

It may seem that adding more trucks will always increase the delivery rate, which will reduce the total time required to complete the delivery. As we have seen, adding trucks does increase the rate in the beginning. Eventually, however, the rate will reach a maximum, and once the maximum has been reached, adding more trucks will not help. For brick delivery, the maximum rate (and minimum total delivery time) occurs when trucks travel down the road bumper-to-bumper at the speed limit, as Figure 12.3 illustrates.

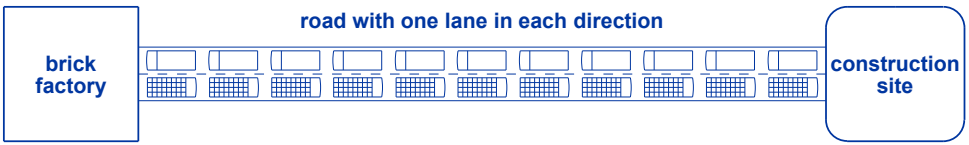


Figure 12.3 Illustration of the road when it is saturated. Trucks are traveling bumper-to-bumper at the speed limit.

Adding more trucks won't have any effect because the roadway is *saturated*. That is, the capacity of the road has been reached. No matter how many additional trucks full of bricks are available, they will simply wait at the factory because there is no space on the road (i.e., it is completely full of trucks).

Is there any way to increase the delivery rate and thereby lower the total time for brick delivery? Yes, we can change the road! For example, consider the effect of adding an extra lane to the road in each direction. Now imagine trucks moving down the road bumper-to-bumper, but with two lanes instead of one. Even though the speed limit remains the same, two trucks can now proceed side-by-side instead of one truck. Now if we count truckloads per hour, the rate will be double what it was for a single-lane road. As a result, it will take half as long to transport all the bricks.

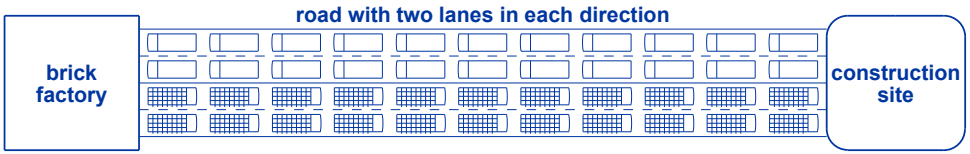


Figure 12.4 Adding a lane in each direction doubles the delivery rate because two trucks can travel side by side.

The idea of adding extra lanes to a highway is a good analogy for how we can increase network capacity. A truck carrying bricks is similar to a packet carrying bits. A computer can generate packets rapidly, but once the network becomes saturated with back-to-back packets, having more packets ready to send does not help. However, if engineers change the network to allow two packets to travel at the same time (like trucks side-by-side in two lanes), the total time to transfer a large data object over the network will be cut in half. In fact, many of the advances in networks have involved discovering ways to send more bits per second. For example, the way DSL technology increased capacity was to define a set of *channels* and then send data over multiple channels at the same time.

Now we can understand the relationship between capacity and "speed": a higher capacity network reduces the total time required to transfer data, making the transfer faster. For example, doubling the capacity of a network will cut the time required to download a movie by half. Therefore, when an ISP says a network is faster, it doesn't mean that packets travel faster than the speed of light. Instead, it means that the total time required to transfer a digital object will be lower.

When an ISP advertises a higher speed *network, the ISP means a network with higher capacity that can deliver more bits per second; the term "speed" is used because increasing the capacity of a network means downloads will complete faster.*

12.5 Transfers Across The Internet

Our description of network capacity accurately describes what happens when the capacity of a single network is increased. However, the Internet isn't a single network; it is a network of networks. When they travel from one device on the Internet to another, packets traverse many networks. When a packet leaves a user's device, the packet crosses an access network. The access network leads to a local ISP. The packet may then pass through a regional ISP, a Tier 1 backbone, another regional ISP, another local ISP, and finally an access network to the destination device. At each ISP, the packet may traverse multiple networks.

The question arises, how many bits per second can two devices transfer across the Internet? Of course, there is no simple answer because the rate depends on the locations of the two computers and the path packets follow when traveling from one to another. Although we cannot arrive at a specific answer, the question brings up an important principle that can help users better understand how the capacity of an access network affects Internet transfers.

Going back to the brick delivery analogy will help. We used a single road to explain transfers across a single network. How can we extend the analogy to explain transfers across the Internet? To simplify things, let's start with a very short path through the Internet, one that only crosses two networks. In terms of brick delivery, imagine two roads joined together running from the factory to the construction site, as Figure 12.5 illustrates.

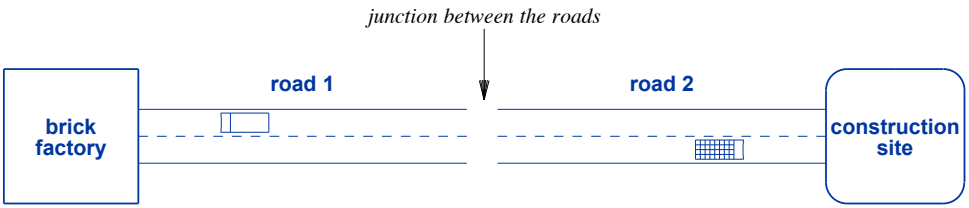


Figure 12.5 Illustration of a path formed from two roads that have been joined together.

To reach the construction site, a truck must traverse both roads. In the figure, both roads have one lane in each direction, and they both have the same speed limit. Consequently, the combination behaves exactly like a single road (assuming a truck can pass from one road to the other without stopping). As with a single network, the delivery rate increases as more trucks are sent across the roads — two trucks doubles the rate, three trucks triples the rate, and so on. Like a single road, the combination of two roads will reach saturation when trucks are moving across both of the roads bumper-to-bumper at the speed limit. The key idea is that because the roads have the same capacity and the same speed limit, they will both become saturated at exactly the same time. The same is true for two networks that have been connected together by a router: two interconnected networks behave like a single network. Furthermore, if the two networks have the same capacity, they will both become saturated at the same time.

12.6 Connecting Heterogeneous Networks

Our analogy is flawed because networks used in the Internet do not all have exactly the same capacity. Let's see what happens when two interconnected networks have different capacities. Suppose, for example, that the brick factory wants to increase their delivery rate, and decides to pay for an extra lane to be added to the road that leads to

the factory. Unfortunately, they cannot afford to upgrade the second road. Figure 12.6 illustrates the new situation.

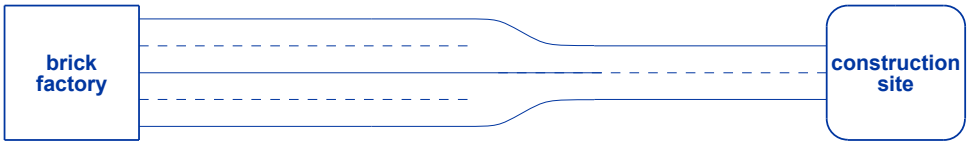


Figure 12.6 Illustration of two interconnected roads in which one has two lanes in each direction, and the other only has one lane in each direction.

Consider what happens when delivery starts. Trucks can leave the factory and proceed down the first road side-by-side. When they reach the second road, however, only one lane is available, and the trucks must proceed one at a time. Using traffic terminology, we say that the two lanes *merge* into one lane.

From experience with traffic, we know what will happen if trucks continue to leave the factory: eventually, the one-lane road becomes saturated, and so does the first road. Figure 12.7 shows what happens as traffic backs up.

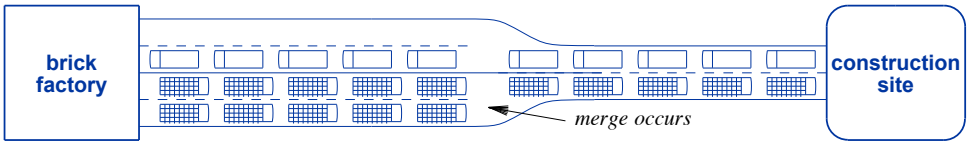


Figure 12.7 Illustration of the roads when the one-lane road reaches saturation.

As the figure shows, trucks can still enter the first road bumper-to-bumper, side-by-side. Once the road fills, however, they will not be able to enter as quickly. Instead, the rate will be controlled by the merge — trucks going toward the construction site are limited by the one-lane road. Trucks cannot enter the first road faster than they can exit to the second road. There are two points to note. First, adding a second lane to one of the two roads had no effect on the delivery rate or the total time required for delivery. Second, the extra lane for returning trucks makes no difference because trucks can only move at the single-lane rate.

We can summarize:

> *If traffic passes across two roads with the same number of lanes, adding lanes to one of them may have no effect on the rate at which vehicles pass.*

Consider extending the example to more roads. Currently, trucks passing from the factory to the construction site have two lanes and then one lane. Suppose we add a third road with four lanes in each direction. Will the rate at which trucks pass across the three roads change? No. The rate is still limited by the one-lane road. Furthermore, it doesn't matter whether the one-lane road comes first, last, or in the middle — the delivery rate along the entire path cannot be higher than the delivery rate on the one-lane road.

The idea applies directly to Internet communication. Every time someone communicates over the Internet, packets flow across a series of networks between the sender and receiver. Data can never flow across the path faster than it flows across the network with least capacity. Scientists use the term *bottleneck* to refer to such networks.

A scientific principle states that along a path through the Internet, a network with least capacity is the bottleneck *for the path, and data cannot flow across the path faster than it flows across the bottleneck network.*

As an example, consider Figure 12.8 which lists the capacities of six networks along a path in the Internet.

Network Type	Capacity
Cable modem connection	75 Mbps
Ethernet	1 Gbps
Ethernet	1 Gbps
Satellite	400 Mbps
Ethernet	100 Mbps
Wi-Fi	20 Mbps

Figure 12.8 The type and capacity of networks along an example path in the Internet.

What is the maximum data rate across the entire path? When looking at the figure, remember that 1 Gbps (Gigabits per second) is 1000 Mbps (Megabits per second). To find the rate of the path, one only needs to find the network with the least capacity. In this case, the bottleneck is the Wi-Fi network.

12.7 The Effect Of Sharing

Our discussion of Internet performance has been optimistic because we have only analyzed traffic sent from one source to one destination. The bottleneck principle, for example, explains the best possible rate at which data can be transferred across an Internet path if there is no other traffic.

In practice, the Internet seldom delivers data at the best possible rate. The reason is that no single user has exclusive use of a path across the Internet. Because the Internet is a shared infrastructure, a given network is likely to have simultaneous traffic from multiple users. Thus, if we could examine a network being shared by two users, we might see a packet that belongs to user 1, then a packet that belongs to user 2, then a packet from user 1, and so on.

How does sharing affect performance? If two users share a network and the sharing mechanism is fair, each user will receive one-half the network capacity. If three users share, each will receive one-third of the capacity. The point is:

For a shared network, the effective capacity *that a given user can receive is the network capacity divided by the number of simultaneous users.*

As an example, consider Figure 12.9, which lists the effective capacity of the networks in Figure 12.8 when each network carries traffic from multiple users.

Network Type	Network Capacity	Number Of Users	Effective Capacity
Cable modem connection	75 Mbps	3	25 Mbps
Ethernet	1 Gbps	500	2 Mbps
Ethernet	1 Gbps	100	10 Mbps
Satellite	400 Mbps	20	20 Mbps
Ethernet	100 Mbps	2	50 Mbps
Wi-Fi	20 Mbps	1	20 Mbps

Figure 12.9 The effective capacity of networks from Figure 12.8 when multiple users share each network. Note that a large number of simultaneous users can reduce the effective capacity of a network dramatically.

What is the best possible data rate a user will experience across the entire path when sharing occurs? To find out, apply the bottleneck principle, but use the effective capacity of each network. In the figure, the network being shared by 500 users becomes the bottleneck because it has the lowest effective capacity, 2 Mbps.

We can restate the bottleneck principle for networks that are shared.

A scientific principle states that along a path through the Internet in which networks are shared, a network with least effective capacity is the bottleneck *for the path, and data cannot flow across the path faster than it flows across the bottleneck network.*

12.8 Delays In The Internet

The discussion above focuses on the rate at which data can be sent. However, another Internet performance measure is relevant in a few cases: the time it takes to send the first packet from one computer to another.†

When does a user care about the delay of a single packet? The only time it matters to a user is when an application allows two users to interact, such as when they can see and hear each other over video and audio. For example, *FaceTime*, *Google Hangouts*, *Skype*, *Tango*, and *Viber* all provide such interaction.

Why is delay important for such apps? Because the human brain expects face-to-face communication with no delay, and becomes slightly confused by delay. In normal conversation, a delay usually means the other person has finished talking. When the Internet introduced an artificial delay, both parties are likely to speak at the same time. One would think that the tiny delays introduced by the Internet would go unnoticed, but humans are quite sensitive. For example, studies of telephone calls have found that humans start to notice delays of a tenth of a second, and delays of more than two-tenths of a second make communication confusing and unpleasant. We can summarize:

> *Studies have shown that when humans talk to one another over a telephone, delays of more than two-tenths of a second cause problems.*

What causes delay in the Internet? We know that electromagnetic signals used in network communication propagate at the speed of light, 186,282 miles per second. The farthest distance between two points on the surface of the earth is about 12,000 miles, and signals can travel that far in less than one-tenth of a second.

There is only one case where the speed of light is significant: satellites. A *geostationary communication satellite* is 22,236 miles above the earth, approximately one-tenth of the way to the moon. At the speed of light, it takes a signal more than one-tenth of a second just to reach the satellite, and more than one-tenth of a second to bounce back. Hikers who use satellite phones know that communication is difficult.

> *Even at the speed of light, communication through a geostationary satellite introduces a delay of over two tenths of a second.*

Fortunately, Internet communication seldom involves a satellite. Unfortunately, users may still experience annoying delays when using an app that sends live audio or video. What causes such delays? Although a path contains multiple routers, the delay introduced by a router is relatively small. The primary source of long delays in the Internet is *network congestion*. Paths in the Internet become congested in the same way that highways become congested: traffic exceeds capacity. Remember that the Internet includes networks of all sizes and types, and look at Figure 12.7 again.‡ Imagine a

†Scientists use the term *latency*; we will use the more intuitive term *delay*.

‡Figure 12.7 can be found on page 136.

high-capacity network connected to a lower-capacity network. When more packets arrive over the high-capacity network than can be sent over the low-capacity network, congestion results. Chapter 14 explains how software helps detect congestion and reduce traffic, but when congestion occurs, it introduces delays that humans notice during live audio and video sessions. The point is:

> *Internet delay is a measure of the time required to send a single packet from one device to another; most long delays arise from congestion.*

12.9 Should You Pay for Higher Speed Internet?

Many users wonder about Internet performance, especially when they sit waiting a long time for a streaming movie to start playing or for a web page to load. They wonder, "Is something wrong with the Internet?" Alternatively, they ask themselves, "Is something wrong with my device?" When a salesperson shows up and offers "higher speed Internet," the offer sounds enticing. Will higher speed Internet solve the problem? We now know enough about the Internet to be able to answer the question. The ISP isn't selling a completely new and improved Internet; they are merely selling a higher capacity access connection between you and the ISP. The bottleneck principle helps us answer the question:

> *Increasing the capacity of a user's access connection will only reduce the transfer time in cases where the access connection is the bottleneck of the path.*

Of course, you probably access services all over the Internet. If access to all of them is slow, your connection is a likely bottleneck. But if some work well and others are slow, your access connection is not the bottleneck. If several devices share your connection (e.g., a family that has laptops, smart phones, games, and Internet TV all sharing Wi-Fi on their cable modem), you can check whether simultaneous transfers make your connection a bottleneck. Browse a web site or download a file, then start multiple transfers and repeat the action. If your access appears to slow down substantially, you may need more capacity to handle simultaneous transfers.

Am I getting what I pay for? If a subscriber pays for an access connection of 25 Mbps, how can the subscriber know the connection is operating at that speed. Fortunately, services are available that measure speed, which means you don't have to trust your ISP's measurements. For example, among others, you might try:

http://www.speedtest.net/

which will measure the effective upload and download rate of transfer across your connection. If you are sharing the connection among several devices, you can repeat the test while other devices are transferring data.

12.10 Summary

Although ISPs advertise higher speed, they are selling a higher capacity access connection. Network capacity is analogous to lanes in a roadway because more network capacity allows more packets to be transported in a given time, just as more lanes allow more trucks to pass down the road in a given time.

Because the Internet is a network of networks, a typical path through the Internet crosses multiple networks. The capacity of the path is limited to the capacity of the bottleneck network. When a network is shared among multiple users, the effective capacity a given user receives is the capacity of the network divided by the number of users.

A second measure of network performance, delay, measures the time it takes one packet to cross a path. Delay is important when using an application that includes live interaction, such as an Internet phone call or a teleconference. Most long delays in the Internet are due to congestion caused when the packet traffic exceeds the capacity of a network.

EXERCISES

12.1 An ISP advertised "Higher speed Internet direct to where you need to go." What is inaccurate about the ad?

12.2 The Ethernet hardware on old computers operated at 100 Mbps, and modern hardware operates at 1 Gbps. How many times faster is the modern hardware?

12.3 A company advertises a lightweight portable device for use in the wilderness, even when there is no cell phone service. The ad shows a business man sitting with the device in front of a tent and a large picture of an office behind him. The caption reads, "With the right backdrop for your teleconference, they'll never know." Will others know? Explain.

12.4 What is Internet congestion, and why does it occur?

12.5 Run a speed test on your Internet connection. Do the upload and download rates match the capacity your ISP advertised?

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

13 IP: Software To Create A Virtual Network

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

13

IP: Software To Create A Virtual Network

13.1 Introduction

Chapter 10 describes the Internet as a network of networks, formed by using special-purpose computers called routers to interconnect networks. Of course, merely connecting hardware together does not make an Internet. Routers and hosts that connect to the Internet need special software before communication is possible. This chapter describes the basic software that makes the Internet appear to be a single, large network.

13.2 Protocol: An Agreement For Communication

It is impossible for two humans to communicate unless they agree to speak a common language. The same holds true for devices — two devices cannot communicate unless they share a common language. A network *communication protocol* is an agreement that specifies a common language two devices will use to exchange messages. The term derives from diplomatic vocabulary, in which a protocol specifies the rules under which a diplomatic exchange occurs.

A computer communication protocol defines communication precisely. For example, a protocol specifies the exact format and meaning of each message that a device can send. It also specifies the conditions under which a device should send a given message, and how a device should respond when a message arrives.

13.3 Basic Functionality: The Internet Protocol

In the Internet, one of the key communication protocols is called, appropriately, the *Internet Protocol*. Usually abbreviated *IP*, the protocol specifies all the details about the packets that are sent across the Internet. IP specifies exactly how a packet must be formed. It then specifies the exact steps a router takes to forward each packet on toward its destination. A device that connects to the Internet must follow the rules of the Internet Protocol, or routers will discard the IP packets the device sends.

13.4 Packets Arrive Unchanged

The Internet Protocol introduced an important principle: packets are not changed as they pass from the sending device to their destination. Thus, when an IP packet arrives at a device, the packet that arrives is an exact copy of the original IP packet that was sent.

Why is the idea of an unchanged packet so significant? After all, that's the way postal systems have worked for centuries. In fact, a letter writer would be appalled if they discovered that their letter had somehow been changed during its trip through the postal system.

Surprisingly, some early computer networks did indeed change messages as the messages were delivered. To understand the motivation for changing messages, think about how we identify the party at the other end of the communication. One method uses the labels *local* and *remote*. My device is local, and the other user's device is remote. Now consider a message sent from device A to device B. When the message is created, the local device is A and the remote device is B. When the message arrives, however, the local device is B and the remote device is A. Thus, if a message specifies "local A" and "remote B," it makes sense for the labels to be swapped in the network, between the time the message leaves the sender and before it arrives at the recipient. Swapping means the recipient will see "local B" and "remote A," which is correct from B's point of view.

The Internet designers realized that changing packets complicates the network and leads to problems. So, instead of the ambiguous concept of local and remote devices, they decided to label each packet with its *source* (the device that originally sent the packet) and its *destination* (the device that will ultimately receive the packet). As the packet travels along, routers always know where the packet is going and which device sent it. The point is:

> *The Internet does not modify a packet as the packet travels. As a result, the ultimate recipient of a packet receives an exact copy of the packet that the original sender sent.*

13.5 Internet Software On Your Device

Computer hardware does not understand IP, and does not know how to send IP packets. Therefore, attaching a device to the Internet does not mean it can use Internet services. To communicate on the Internet, a device needs IP software. Indeed, every device that uses the Internet, including conventional computers, smart phones, and *Internet of Things* (*IoT*) devices, must use IP software for Internet communication.

A typical device has many applications (apps) that communicate over the Internet. Does each app have IP built into it? No. Because IP is required for all Internet services, vendors place a single copy of IP in the operating system, and allow all apps on the device to share the copy. The operating system starts when the device is powered on, and initializes IP. As a result, the device is ready to send and receive packets at all times. The point is:

> *Because all Internet communication uses the Internet Protocol, a device must have IP software before it can access the Internet. Instead of waiting for an application to need IP, the operating system starts IP running automatically so apps can use the Internet at any time.*

13.6 Internet Packets Are Called Datagrams

To distinguish between Internet packets and packets for other networks, we call a packet that follows the IP specification an *IP datagram*. The name was chosen to invoke the idea of a telegram because the Internet packet delivery service handles datagrams in much the same way that a telegraph office handled telegrams. Once the sending device creates a datagram and starts it on a trip through the Internet, the sender is free to resume processing in the same way that an individual is free to leave a telegraph office after handing an operator a message to be sent. A datagram travels across the Internet independent of the sender, analogous to the way operators forward a telegram to its destination independent of the person who sent the telegram. To summarize:

> *Each packet sent across the Internet must follow the format specified by the Internet Protocol. Such packets are called IP datagrams.*

13.7 Providing The Illusion Of A Giant Network

Chapter 10 describes the Internet from two points of view. To users and the software running on hosts that attach to the Internet, the entire Internet appears to be a single giant network. Figure 10.4, which can be found on page 112, illustrates the idea.

Although it defines many communication details, the Internet Protocol has one overriding purpose: provide the illusion of a single large network. Every host and router in the Internet has IP software installed. The software allows any host to create an IP datagram and send the datagram to any other host. In essence, IP transforms a collection of networks and routers into a seamless communication system by making the Internet function like a large, unified network.

Computer scientists use the term *virtual* to describe technologies that present the illusion of larger, more powerful computational facilities than the hardware provides. The Internet is a *virtual network* because it only presents the illusion of a single, large network. In fact, the Internet is a network of networks, and the underlying networks vary in size and type. IP software takes care of the details and allows users to think of "the Internet" as a single entity. Users remain unaware of the Internet's internal structure of networks and routers, just as telegraph users remained oblivious of the underlying system.

The point is:

> *IP software allows the Internet to operate like a single network that connects several billion devices. IP software allows any device to send an IP datagram to any other device.*

13.8 The Internet's Internal Structure

Recall from Chapter 10 that although IP software allows users to view the Internet as a single, large network, the Internet contains a complex internal physical structure that users never see. Hundreds of thousands of routers interconnect networks. Figure 10.5, which can be found on page 112, illustrates the internal structure.

IP software on every router must know how to reach any destination in the Internet. Does that mean every router has a list of all the hosts in the Internet? No. That's the advantage of organizing the Internet into a hierarchy of ISPs — a given router knows about local destinations, and sends all other datagrams up the hierarchy.

The idea of only knowing local destinations has been used in the past. For example, a telegraph office in a small town knew how to deliver a telegram to any address in the town. When someone sent a telegram to another address, the local telegraph office sent it to a big city. Similarly, if a wireless router has two laptops attached, the router only knows how to deliver to the two laptops; it sends all other datagrams across the wired network to a local ISP.

A router in a local ISP only needs to have a list of all the ISP's customers. If a datagram arrives destined for a customer, the router forwards the datagram to the customer. All other datagrams that arrive are forwarded up the hierarchy to a regional ISP. Similarly, routers in a regional ISP must know all the customers in the region, but do not need to know about other regions of the world.

Tier 1 ISPs that form the top of the hierarchy differ from all other ISPs because routers in a Tier 1 ISP must know how to reach every possible destination in the Internet. The routers at a Tier 1 ISP are gigantic. A single router can be over forty feet long and six feet tall. The routers in a Tier 1 ISP must keep track of how to reach Internet destinations as the Internet grows; therefore, the information in each router must be updated constantly.

13.9 Datagrams Travel Inside Network Packets

When a datagram travels across the Internet from one device to another, it follows a path across multiple networks. The Internet includes a wide variety of networks, and each network designer has chosen a packet format and size for their network hardware. How can a datagram be sent across a network if the network hardware does not understand the datagram format? The easiest way to imagine a datagram transfer is to consider how overnight shipping services handle letters. Assume someone has written a letter, placed it in an envelope, and written the name of the intended recipient on the outside. The letter is much like an IP datagram. Suppose the sender asks an overnight shipping service to deliver the letter. The overnight service requires that the letter be placed inside one of their envelopes, and that the name and address of the recipient be written on the outside in the format they specify. The outer envelope is analogous to a *network packet*.

Both the inner and outer envelopes contain a recipient name. Although the names usually agree, they need not be identical. Consider what happens if the sender knows the address of the mail room at the recipient's company, but not the exact office address of the individual to whom the letter is addressed. The sender can mail the overnight parcel to the mail room for delivery. In such cases, the inner address and outer address differ. When the parcel arrives at the address on the outer envelope, the mail room named on the outer envelope opens it and forwards the letter.

Datagram transmission follows the same approach. Each time a datagram must be sent across a network, IP software places the datagram inside a network packet, places the address of the next router in the header of the network packet, and sends the packet. As far as the network hardware is concerned, the entire IP datagram is merely data being carried in a packet the hardware recognizes. When the network packet arrives at the next router, the router "opens" the packet and extracts the datagram. The router examines the destination address on the datagram, and determines the next router along the path. The router forms a network packet that the network hardware understands, encloses the datagram inside the network packet, and sends the packet to the next router. The sending device uses the same technique to send the datagram to the first router along the path, and the last router along the path uses the technique to deliver the datagram to the destination device. Thus, the datagram survives the trip intact, but is enclosed in a different network packet as it crosses each network. Figure 13.3 illustrates how a datagram travels inside a network packet.

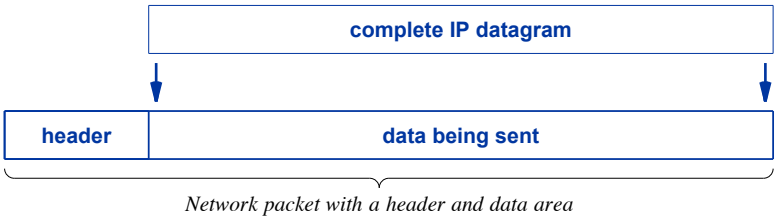


Figure 13.1 Illustration of an IP datagram carried in a network packet, analogous to a letter being carried in an outer envelope.

13.10 Internet Addresses

How does a router know which device should receive a datagram? The datagram contains a label that specifies a destination device. In fact, a datagram resembles a network packet. Like a network packet, a datagram has two parts. One part specifies the device that sent the datagram and the destination to which the datagram has been sent, and the other part contains the data being carried in the datagram. When it creates a datagram, a sending device specifies a destination to which the datagram is being sent.

To make communication possible, every device on the Internet is assigned a unique number known as the device's *IP address*. In a datagram, the destination is specified by giving the IP address assigned to the destination device. Fortunately, users seldom need to type or see IP addresses; most application programs allow humans to enter the alphabetic name of a device when they specify a destination. Chapter 16 describes the format of device names, and explains how each name is translated to an equivalent IP address.

13.11 IPv4 And IPv6

How large is an IP address? Two versions of IP software are used in the Internet, and the size of an IP address depends on which version is being used. The two are known as *IP version 4* (*IPv4*) and *IP version 6* (IPv6); the names reflect the version number that is included in each datagram.† IPv4 and IPv6 offer the same basic functionality, but the datagram format and all the details differ. In particular, each of the two versions defines its own IP addressing scheme. IPv4, which was defined in the 1970s, provides 4,294,967,296 unique addresses, which means over four billion devices can be assigned a permanent address. In the 1990s, several groups predicted the imminent demise of the Internet because they thought all the IPv4 addresses would run out. A committee was formed, and the committee defined IPv6 as a successor to IPv4. As with most engineering by committee, instead of defining a reasonable replacement, the committee made political compromises to satisfy as many groups as possible. The result is a bloated design. In terms of addresses, IPv6 has enough addresses for an absurdly large number of devices:

†IP versions 1 through 3 were early experiments, and version 4 was the first successful design; version number 5 was skipped for political reasons.

340,282,366,920,938,463,463,374,607,431,768,211,456

That's enough addresses so every human on earth can have an Internet all to themselves with as many devices as the current Internet.

The switch to IPv6 did not happen in the 1990s or even in the 2000s. It started after 2010. A political decision was made to accelerate adoption: the remaining IPv4 addresses were allocated without waiting for justifications. Despite the efforts to replace it, IPv4 has lasted and is still used extensively throughout the Internet. To summarize:

The original Internet Protocol, known as IPv4, has enough addresses for four billion devices. Over twenty years ago, a committee designed a second generation of the Internet Protocol known as IPv6 that has an absurdly large number of addresses. Adoption of IPv6 has been slow, and many devices still use IPv4.

13.12 Permanent And Temporary IP Addresses

Before a device can use the Internet, the device must be assigned an address. An address must be assigned for either IPv4 or IPv6. Because IPv6 is still not available everywhere, a device that has software for IPv6 usually also has software for IPv4, and can obtain an IPv4 address as well as an IPv6 address. Having both versions allows a device to connect to a network where routers only understand IPv4. For now, it is only important to know that every device on the Internet is assigned a unique address.

Address assignment can happen two ways:

- A permanent IP address assigned manually
- A temporary IP address assigned automatically.

Permanent IP address. A permanent IP address is assigned manually by a network administrator. The assignment remains in effect until the administrator makes a change. We will learn that large computing systems called *servers* are each assigned a permanent IP address.

Temporary IP address. A temporary IP address is assigned automatically by software, is used for a short time, and then released for another device to use. For example, when a customer walks into a coffee shop and joins the shop's Wi-Fi network, the customer's device is given a temporary IP address. When the customer leaves the coffee shop, the device releases the temporary address, and the same address can be reassigned to a new customer. Chapter 17 explains why temporary addresses are handy, and why the wireless router found in a home uses temporary addresses.

13.13 Summary

The Internet Protocol, IP, specifies the basic rules that a device must follow to communicate across the Internet. IP software on user's devices and routers makes the interconnected set of networks that constitute the Internet operate like a single, large network. To do so, IP defines the format of Internet packets, which are called IP datagrams. IP also defines an address scheme that assigns each device a unique number used in all communication. Two versions of IP exist: version 4 (IPv4) has been used since the 1970s, and version 6 (IPv6) is slowly being adopted.

Before it can communicate over the Internet, a device must be assigned a unique number known as an IP address. The assignment can be permanent (assigned manually by a network administrator) or temporary (assigned automatically by software when needed). Coffee shops and other Wi-Fi hotspots use temporary assignment.

EXERCISES

13.1 Does your smart phone have IP software on it? How can you tell?

13.2 Does the wireless router in someone's house contain a list of every possible device in the Internet? Explain.

13.3 What's an IP datagram, and where is it used?

13.4 An airport offers free Wi-Fi that allows patrons to access the Internet. When a device uses the free Wi-Fi, is the device assigned a permanent IP address or a temporary IP address?

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

14

TCP: Software For Reliable Communication

14.1 Introduction

The previous chapter discusses the Internet Protocol and describes how IP software on hosts and routers gives the illusion of a single large network by making it possible to send an IP datagram from any device on the Internet to any other. This chapter continues the discussion of basic Internet communication software. It examines the second major communication protocol, TCP.

14.2 A Packet Switching System Can Be Overrun

Recall from Chapter 12 that a roadway can become congested if a road connects directly to another road that has fewer lanes. In the Internet, an analogous situation can occur if a router connects a network with higher capacity to a network with lower capacity, as Figure 14.1 illustrates.

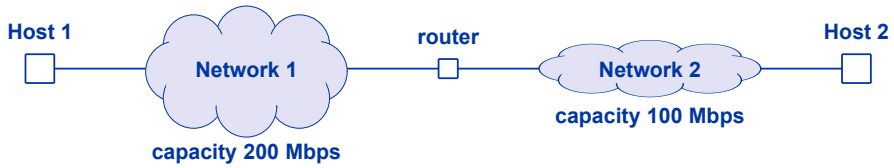


Figure 14.1 An example of networks with different capacities connected by a router. Traffic from Host 1 to host 2 can cause an overrun.

To see how overrun can occur, imagine Host 1 transferring data to Host 2. Host 1 generates datagrams, places each datagram in a packet, and sends the packets over Network 1. Packets enter Network 1 at the rate of 200 Mbps. When a packet reaches the router, the router extracts the datagram from the packet, places the datagram in a new packet that is suitable for Network 2, and sends the packet over Network 2. However, packets can only enter Network 2 at a rate of 100 Mbps, half the rate at which data arrives from Network 1.

When two roads of differing lanes connect, cars must slow down, and a traffic jam ensues. What happens to datagrams? When Network 2 becomes saturated, the router discards them! Of course, each router has a small memory, and can store a few datagrams in memory in case of temporary congestion. However, if datagrams continue to arrive faster than they can leave, the router must discard datagrams until the congestion clears.

14.3 Software To Handle Congestion And Datagram Loss

Because a router will discard datagrams when the router becomes overrun, the researchers who built the Internet knew that additional communication software would be needed. To handle the problem, they invented the *Transmission Control Protocol* (*TCP*). TCP handles the problems of *congestion* (when traffic slows down) and *datagram loss* (when a router becomes overrun and discards one or more datagrams). Hosts that attach to the Internet run TCP software as well as IP software. TCP and IP are designed to work together (which is why the entire set of Internet protocols are known as *TCP/IP*).

To summarize:

> *IP software provides the ability for a host to send a datagram across the Internet to any other host. TCP software handles the problems of congestion and datagram loss that occur when routers become overrun.*

14.4 The Magic Of Recovering Lost Datagrams

Recovering from datagram loss is difficult. More important, how can TCP know that a datagram was discarded? To understand the difficulty, recall that a path through the Internet may involve a long series of networks and routers. Any router along the path can become overrun and discard a datagram. The router that performs the discard could be somewhere in the middle of the path, neither close to the sender nor to the receiver. How can TCP software on the sender or receiver find out that a problem has occurred?

TCP handles problems by arranging for TCP on the sender and TCP on the receiver to coordinate. On the sending side, TCP adds a small amount of extra information to each datagram, including a sequence number. Whenever data arrives at its final destination, TCP software on the receiving host sends an *acknowledgment* back to TCP on the sending host. An acknowledgment is a short message that specifies which data arrived.

The sending TCP is in charge of making sure all the data arrives. Whenever it sends data, TCP software starts a timer using the host's internal clock. The timer works like an alarm clock — when the timer expires, TCP is notified. If an acknowledgment arrives before the timer expires, TCP cancels the timer. If the timer expires before an acknowledgment arrives, TCP assumes the datagram was lost and sends another copy. We use the term *retransmission* to characterize the process of sending a second copy of a datagram.

TCP software on a receiving host sends an acknowledgment when data arrives. If a router discards a datagram, TCP software on a sending host detects the missing acknowledgment and retransmits another copy of the datagram.

14.5 TCP's Sophisticated Retransmission Algorithm

TCP did not invent retransmission — before TCP was invented other computer communication protocols used the scheme of starting a timer and resending data if an acknowledgment failed to arrive before the timer expired. However, TCP's scheme differs from the previous retransmission schemes because earlier retransmission was designed for a single network, where the time required to send data across the network and receive a reply was known in advance. By contrast TCP is designed to work between any two hosts anywhere on the Internet.

The situation TCP faces is that some destination hosts are close to the source and others are far away. TCP must choose how long to wait before retransmitting another copy. If a destination host is close to the source (e.g., in the same building), the time required to send a datagram and receive an acknowledgment is extremely short. If the destination host resides far from the source (e.g., in another country), the time needed to receive an acknowledgment is much longer. In the first case, if a datagram is lost, TCP should retransmit quickly, but in the second case, TCP should wait long enough to see if the first copy arrived before retransmitting or risk clogging the Internet with extra copies that are unnecessary.

The magic of TCP lies in a sophisticated algorithm that automatically chooses how long to wait. As it sends datagrams to a given destination, TCP measures how long it takes for each acknowledgment to arrive. TCP uses the measurements to estimate the current delay to the destination. TCP uses the estimate of delay when it sets the retransmission timer.

From the description above, you may think that the algorithm TCP uses is a straightforward average, but it is not. To understand why, recall that Internet delay — the time it takes to send one packet — depends on congestion. Congestion changes as hosts send or stop sending datagrams. Therefore, TCP must contend with changing delays. A measurement becomes irrelevant in a short time because congestion can appear or disappear quickly. So, the algorithm must continue to take measurements and adjust the timer carefully, not reacting too quickly or too slowly.

TCP's ability to automatically adjust timeout values has contributed much to the success of the Internet. In fact, most Internet applications could not operate without TCP software that adapts to changing conditions. Furthermore, careful measurements and experience have shown that TCP software can adapt to changes in the Internet extremely well — although many scientists have tried to devise improvements, no one has produced a protocol that works better in typical cases. The point is:

> *Because it is designed for the Internet where delay depends on the distance to a destination and delay changes when congestion occurs, TCP uses a sophisticated algorithm to estimate when to retransmit lost datagrams. The algorithm works extremely well.*

14.6 Handling Congestion

In addition to retransmitting datagrams that have been discarded, TCP handles congestion. We already said that congestion affects TCP's retransmission strategy. When many hosts begin to send datagrams and the Internet slows down, TCP increases the time it waits before retransmitting. If conditions change and datagrams begin to flow across the Internet quickly, TCP automatically decreases the retransmission timeout. However, adjusting retransmission only solves part of the problem; TCP adapts to congestion, even before any datagrams are discarded.

Recall that whenever TCP sends data to a destination, the receiving TCP sends an acknowledgment back. Also recall that congestion increases delay. TCP includes an algorithm that uses changes in the time it takes to receive acknowledgments to estimate congestion along the path. When it detects congestion, TCP slows the rate at which it sends data. Perhaps the first network along the path isn't congested at all. That doesn't matter — TCP is measuring the entire path, and when any network in the path becomes congested, TCP slows down. If congestion continues, TCP slows even more. Finally, when congestion decreases, TCP slowly increases the rate.

Does slowing the data rate on one host help? No. However, every device in the Internet that uses TCP follows the same algorithm. Consequently, when a given network becomes congested, *all* hosts that are sending data across the network reduce their rate, which allows congestion to subside. Without TCP, the Internet would quickly experience *congestion collapse*. Congestion collapse is a situation in which hosts continue to send data into a congested network, causing delays to increase until routers start to discard most datagrams, which causes retransmissions. Sending duplicate copies of da-

tagrams into a congested network is a horrible strategy (it's the equivalent of sending twice as many vehicles as usual down a highway after an accident stops traffic).

Once again, our description of TCP's congestion avoidance mechanism makes it sound trivial, but it is not. The algorithm is both sophisticated and efficient, and has proven to work extremely well. The point is:

> *When congestion occurs, TCP automatically reduces the rate at which it sends data. Without TCP software on all hosts, the Internet would experience a phenomenon known as congestion collapse.*

14.7 TCP And IP Work Together

It is not a coincidence that TCP and IP work well together. The two protocols were designed at the same time to work as part of a unified system, and are engineered to cooperate and complement one other. TCP handles problems that IP does not handle without duplicating the work that IP does. The point is:

> *Together, TCP and IP software provide an efficient, reliable communication system. IP provides a way to transfer a packet from its source to its destination, and TCP handles the problems of loss and congestion.*

14.8 Summary

Every host in the Internet needs both IP software and TCP software. IP software provides basic Internet communication and allows a host to send a datagram to any other host. However, like any packet switching system, the Internet can become overrun if many hosts send data at the same time. When hosts send more datagrams than the Internet can handle, IP software in routers must discard some of the incoming datagrams.

TCP software handles the problems that IP does not. TCP on the receiving host returns an acknowledgment when data arrives. TCP on the sending host retransmits data if it fails to receive an acknowledgment. In addition, TCP detects congestion along the path to the receiver, and reduces the rate at which data is sent while congestion is occurring. Because all hosts have TCP software, the Internet does not collapse when congestion occurs.

The algorithms TCP uses to handle retransmission and congestion are both sophisticated and efficient. They adapt to the long or short paths between sender and receiver, and handle changes in delay automatically. TCP and IP software work together to provide a smooth, dependable, and effective communication system.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

15

Clients, Servers, And Internet Services

15.1 Introduction

Previous chapters describe the TCP/IP communication protocols that work together to provide reliable data delivery across the Internet. This chapter describes how application programs use TCP/IP software to provide services across the Internet. It shows that, despite their diversity, all applications on the Internet follow a single organizational model. Later chapters discuss specific examples of services, and show how the model applies in practice.

15.2 All Services Are Outside The Internet

When users think about the Internet, they think about online shopping, following friends on social media, sharing photos, or accessing dozens of other services. The question arises, how do all the services fit into the Internet? The surprising answer is that they are not part of the Internet at all.

As we have seen, the Internet provides a packet service that allows any host to send data to any other host. Where in the Internet are all the web pages, shared photos, and social media sites? They aren't included in the Internet at all. A basic design principle states:

The Internet only provides packet transport; all other services run in hosts that attach to the Internet.

Keeping services separate from the packet delivery mechanism was a stroke of genius. At the time the Internet was designed, the largest communication system in the world, the telephone system, had taken the opposite approach. Telephones were incredibly basic devices. All the intelligence was built into the switches that formed the telephone network. Engineers called the phone "dumb" and switches "smart." Many networking researchers assumed the same design would be used to build computer networks. However, the Internet designers foresaw a better approach in which the network only provided packet delivery and all the intelligence was placed in hosts that connected to the network.

What is the advantage of the Internet approach? The answer is flexibility. If services are built into the network fabric, changing services or adding new services means changing all the network switches. For example, when the *call waiting* service was invented, it took many years for engineers to incorporate it into all the phone switches. In contrast, a new service can be added to the Internet at any time. All one needs to do is change the software in a host, and the service appears.

The Internet's flexibility has enabled new services to appear and old services to disappear without any change to the Internet itself. For example, the World Wide Web wasn't around when the Internet was invented. In fact, those of us who participated in Internet research had used the Internet for ten years before the World Wide Web was invented. Interestingly, the Internet did not need to change to accommodate the Web because web sites consist of software that runs in hosts, not in the network. Similarly, browsers that access web sites run in user's devices, not inside the Internet.

The point is:

Keeping services separate from the Internet itself has stimulated innovation because anyone can deploy a new service at any time without changing the Internet. Services such as the World Wide Web have been created without requiring any change to the Internet.

15.3 Software Provides All Services

Informally, we think of devices communicating, but devices do not. All communication occurs between two computer programs (i.e., apps) that are running in the devices. So, instead of saying, "my device is taking forever to access Facebook," a user should say, "the web browser software running in my device is taking forever to access Facebook." Although such distinctions are usually unimportant, knowing that apps perform all communication will help us understand how Internet services work.

15.4 Services Use Client And Server Apps

The Internet offers an amazing set of services with diverse styles of interaction. In some cases, two humans interact. In other cases, a human interacts with a remote computer program that supplies information. In still others, two computer programs communicate without human intervention. Some interactive services often allow a user to remain connected for hours; other services only need a few milliseconds to supply the requested information. Some services allow users to download (fetch) information, while others allow users to upload (store) information. Still others allows users to update (change) information. Some services involve only two computer programs, and other services, such as games, permit multiple users to interact. Some services exchange audio and video streams; others only use text.

Despite the wide diversity among Internet services and apparent differences among them, the software that implements a service always follows a single approach known as *client-server computing*. The idea behind client-server is straightforward: an app on one device offers a service and an app on another device accesses the service. An app that offers a service is called a *server*, and an app that accesses a service is called a *client*. For example, database software that stores information and makes it available is classified as a server, and the software a user runs to contact a database server and look up information is classified as a database client.

To summarize:

> *Communication across the Internet always occurs between a pair of apps. An app that offers a service is called a* server*, and an app that accesses a service is called a* client.

15.5 A Server Must Always Run

Users typically wait until they need a service before launching a client app that accesses the service. Furthermore, once a user has finished access to a service, the user often stops the app. Thus, client use is unpredictable because it depends on the whim of a user.

A server app does wait to run until it's needed because a client will not be able to access the service unless the server app is already running, waiting for a client to contact it. Consequently, a server is not launched by a user. Instead, a computer that runs servers is configured to launch each server automatically when the computer is powered on. A server remains running, ready for a client to access it. After a client finishes using the server, the server does not stop, but remains running, ready for the next client.

15.6 Multiple Clients Can Access A Server Simultaneously

A device can run multiple apps at the same time. For example, a smart phone can run an app that plays music at the same time a user runs an app that searched the Web. Furthermore, if a user runs two apps that both use the Internet, they can proceed because they do not interfere with one another. The idea of multiple apps using the Internet simultaneously is an important feature of many Internet services.

To understand why multiple apps are important, consider a server. Imagine that a user launches a client that contacts the same server and begins to use the service. Also imagine that while the first client is still using the service, another user launches an app that becomes a client and accesses the server. What happens? They both proceed. Server software is designed so that each time a new client contacts the server, the server creates a new copy to handle that client. The copies each proceed without interfering with one another. Figure 15.1 illustrates two clients accessing a server simultaneously.

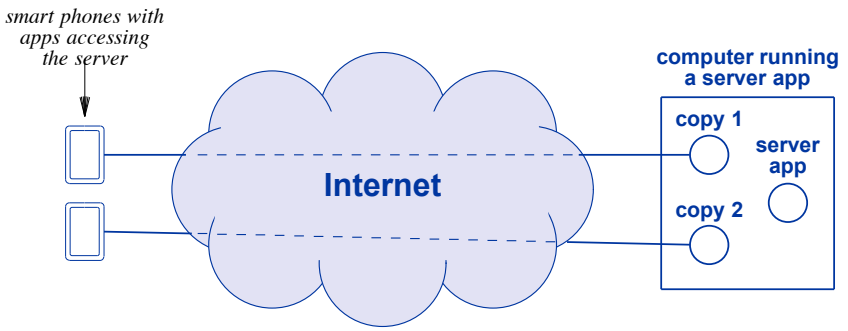


Figure 15.1 Illustration of two client apps running in smart phones contacting a server simultaneously. The server creates a copy of itself for each client.

You do not have to understand the details. Just appreciate that from a user's perspective, multiple clients can appear to access the same server simultaneously without interference. Thus, you and a friend can both access a Facebook server at the same time. We can summarize the key idea that explains how Internet services operate:

A server allows multiple clients to use the service at the same time without interference.

15.7 Ambiguous Terminology

The term *server* has become ambiguous. Server software that handles many clients simultaneously requires a computer with significant amounts of memory and a powerful processor. Therefore, servers are usually run on special computers that have powerful, multi-core processors and especially large memories. Although the term *server* really refers to the app that provides a service, hardware vendors apply the term to any large, powerful computer, even if it is not running server software. The ambiguity can be frustrating when server software and hardware become confused. Fortunately, the meaning is usually clear from the context. If someone says, "we need to purchase six additional servers," they are referring to hardware; if someone asks, "is the server currently handling three clients?" they probably mean the software.

An average user only encounters the ambiguity when shopping for a computer, especially when purchasing a computer to play video games. Like server software, some video games require a computer with many processor cores and a large memory. Although the term "game computer" is often used, vendors also tend to describe their most powerful computers as "server class" computers.

15.8 Summary

The Internet only provides packet delivery; all services run in hosts that are connected to the Internet. The design, which was a departure from previous network designs, offers flexibility because it allows new services to be created and deployed without changing the Internet. Many of the current services, including the World Wide Web, had not been invented when the Internet was created, and the services were added later without requiring changes to the Internet.

The Internet offers a wide variety of services that use many styles of interaction. Despite apparent differences among the available services, all services on the Internet follow the same general paradigm. When apps on two hosts communicate, one acts as a server that offers a service, and the other acts as a client that accesses the service. A client app runs whenever a user accesses an Internet service. A server app must always remain running, so a server starts when the computer is powered on, and remains running until the computer is shut down. Server software allows multiple clients to access the server simultaneously without interfering with one another. A powerful computer designed to run server software is itself called a server.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

16 Names For Computers

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

16

Names For Computers

16.1 Introduction

The previous chapter explains that the Internet only provides packet delivery, and that computers attached to the Internet provide all other services. This chapter describes a key Internet service, one that allows humans to enter names for computers in place of numeric IP addresses. It explains the naming scheme, and describes how software on your device converts an alphabetic name into the computer's IP address.

16.2 Computer Names

Recall that the Internet assigns each host a numeric value called an IP address, and that every packet sent across the Internet contains the IP address of the computer to which it has been sent. When written in decimal, an IP address contains many digits, making it difficult to remember and enter correctly.

To make it easier for humans, a system known as the *Domain Name System* (*DNS*) was invented that allows users to enter a name rather than an IP address. The names are known as *domain names*. To contact a service, a user launches an app, enters the alphabetic domain name of the server, and the app uses DNS to translate the name to the server's IP address. We will discuss how DNS performs the translation after considering how domain names are assigned.

16.3 Computer Names Past And Present

In the 1990s, most office workers had a desktop computer. When a company connected to the Internet, each computer was assigned an IP address and a domain name. Individual employees were allowed to choose a name for their computer, and the names were often chosen to be fun. Many computers had names of movie stars, characters from fiction, terms from popular culture, and even characters from the comics.

By the late 1990s, companies had started to shift employees to less expensive laptop computers. In addition, companies started using the technology described in Chapter 17 that assigns a computer a temporary IP address as needed. Because a laptop does not receive a permanent IP address, the laptop does not have a permanent domain name. Now, the only devices that have names are servers. Consequently, names have become boring, as Figure 16.1 shows.

rank	name	rank	name	rank	name
1	mail	18	vpn	35	server1
2	www	19	io0	36	bb
3	server	20	db	37	b
4	ns2	21	mx2	38	a5
5	ns1	22	exchange	39	a3
6	smtp	23	router	40	c1
7	mail2	24	e0	41	a7
8	remote	25	vps	42	a6
9	host	26	cpe	43	ad
10	gw	27	e1	44	a4
11	mail1	28	a0	45	mail3
12	mx	29	a	46	e
13	webmail	30	bc	47	d1
14	ftp	31	a1	48	da
15	ns	32	gateway	49	de
16	mx1	33	web	50	cc
17	ip1	34	static		

Figure 16.1 The 50 most frequently assigned computer names in 2018.

A computer that runs the company's email server might be named *email*, which is the most popular domain name. Other variants related to email also make the list: *mail1*, *mail2*, *mail3*, and *webmail*. Some of the items on the list that appear to make no sense refer to technical terms. For example, *mx* refers to a *mail exchanger*, a system used with email, and *smtp* is the acronym for the protocol email systems use to transfer mail from one computer to another.

The second most popular name is *www*, chosen for a computer that runs the company's Web server. It may surprise you that *www* is second to email because the Web is so heavily used. However, businesses make extensive use of email. For many years, *www* was the most popular name, but was eventually overtaken by *mail*.

Evidently, many network administrators want to avoid the hassle of choosing names, so they opt for completely generic names. For example, *host*, *router*, and *server* make the list! Note that *server1* also makes the list, as do individual letters, *a*, *b*, *c*, and so on. Interestingly, *d* is less popular than *e*, and is just outside the top 50 at position 53.

16.4 A Computer's Name Must Be Unique

Although humans prefer to use short names, longer names must be used on the Internet to avoid assigning the same name to multiple computers. Two computers with the same name would create a significant problem because communication software could not distinguish between them. The point is:

> *Because apps use names to identify a computer, each computer on the Internet must have a name that differs from the names of all other computers.*

16.5 Using Suffixes To Make Each Name Unique

To make names unique, the Domain Name System extends each name by adding a suffix. Each organization is assigned a string that identifies the organization, and the full name of a computer at the organization consists of the computer's local name followed by a period and the organization's suffix. Currently, the *Internet Corporation for Assigned Names and Numbers* (*ICANN*) approves suffixes, and guarantees that once a suffix has been approved for one organization, no other organization can use an identical suffix. For example, because Harvard University is classified as an *edu*cational institution, it requested and was assigned the suffix:

harvard.edu

Later, when a bookstore in Harvard Square that is named the *Harvard Bookstore* decided to join the Internet, it requested and was assigned the *com*mercial name:

harvard.com

The suffix *harvard.com* clearly distinguishes the company from the university.† If both Harvard University and the Harvard Bookstore each name one of their computers *www*, the suffixes guarantee that the full names of the two computers will differ:

www.harvard.edu

and

www.harvard.com

†Chapter 31 explains how criminals exploit similarities in names to fool users.

The point is:

> *Because a suffix appended to the name of a computer identifies the organization that owns the computer, the full names of any two computers owned by separate organizations are guaranteed to differ from one another.*

16.6 Domain Names With More Than Three Labels

Although the examples above imply that Domain Names always have *labels* (i.e., three parts that correspond to the local computer name, the organization, and the organization type), DNS allows names to contain multiple labels. Once an organization obtains a suffix, the organization is allowed to add additional labels to the names of its computers. Adding labels to names allows computer names to reflect the organization's internal structure, and gives groups inside the organization the ability to assign arbitrary computer names.

An example will clarify the idea. Purdue University follows the pattern used in many universities: each department is assigned a label. The Computer Science Department chose the label *cs*, and the Physics Department chose *physics*. Thus, the names of all computers in the Computer Science Department end with the suffix *cs.purdue.edu*, and the names of all computers in the Physics Department end with *physics.purdue.edu*. The two departments can assign names to their computers without consulting one another because the suffix guarantees the names will not be the same. For example, if both Computer Science and Physics have a computer named *groucho*, the full names of the two computers are:

groucho.cs.purdue.edu

and

groucho.physics.purdue.edu

16.7 Top-Level Domains Before And After ICANN

Initially, the DNS had few top-level suffixes (called *top-level domains*). The top-level domains that were not specific to the US were designated *generic*, and intended to satisfy most situations. Commercial institutions registered under *.com*, and educational institutions registered under *.edu*. Non-for-profit organizations registered under *.org*. In addition, a top-level domain was established for each country, using the international 2-letter country code. For example, the *.uk* domain was assigned to the United Kingdom, *.fr* was assigned to France, *.us* was assigned to the United States, and *.de* was assigned to Germany (from the first two letters of *Deutschland*).

Once it took over ownership of domain names, ICANN expanded the set of top-level domains. Critics claimed that the expansion was unnecessary, and was only done to increase revenue for ICANN and the organizations that ICANN designated to register names. In any case, ICANN has generated a long list of additional top-level domains; Figure 16.2 lists a few of the many new top-level domains.

Domain Name	Meaning
.aero	Air transport industry
.asia	Regional domain for Asia
.biz	Businesses
.cat	Catalan language and culture
.club	Businesses and interest groups
.coop	Cooperative associations
.guru	Individuals and groups offering expertise
.jobs	Human resource management
.museum	Museums
.name	Individuals can register personal names
.pro	Credentialed professionals
.tech	Individuals and groups providing technical support
.travel	Travel industry
.xxx	Internet pornography

Figure 16.2 Some of the top-level domains ICANN added to the Internet Domain Name System.

16.8 Domain Names Outside The US

As we have seen, the naming system provides two ways to organize names: by type and by country. In the United States, most organizations have chosen to register by type. Thus, companies register under *.com* and universities register under *.edu*. However, other countries have chosen to follow alternative schemes. In particular, each domain name ends with the 2-letter country code. Those in Canada end in *ca*, and those in the United Kingdom end in *uk*. In China, Yahoo is named:

yahoo.com.cn

because *cn* is the internationally recognized 2-letter county code identifier for China. Similarly, because domain names used in Germany end in *de*, BMW corporation's web site is named bmw.de. Each country chooses how to further divide domain names. For example, because the United Kingdom has chosen *ac* to denote academic institutions, a computer at the University of York in England has the name:†

minster.cs.york.ac.uk

†The name was chosen because York is the site of *York Minster*, a well-known cathedral.

16.9 Translating A Name To An IP Address

Although humans enter a name, such as *amazon.com*, the underlying communication software must use an IP address to communicate with Amazon. How does an app convert a name to an IP address? It uses an Internet service!

The Domain Name System uses the client-server approach described in the previous chapter. An app running on a user's device acts as a client, and contacts a *domain name server*. The app sends the server the string of characters the user entered (e.g., *amazon.com*), and asks the server,

What is the IP address for this name?

The server looks up the answer, and returns the correct IP address. Once it has the IP address of the computer, the app can send datagrams to the computer and receive replies.

To summarize:

> *Before an app can contact a remote computer, the app must know the computer's IP address. If a user enters a name or clicks on a link that contains a name, an app sends a request to a domain name server, and the server sends a reply that specifies the computer's IP address.*

16.10 Many Domain Name Servers

It would be impossible to build a single server that could answer all domain name questions. There are two reasons. First, with billions of apps looking up domain names, even a supercomputer could not keep up with the rate at which questions arrive. Second, because new computers are added to the Internet constantly and existing computers move to new locations, new names appear and IP addresses are changed. The volume of updates would overwhelm one server.

The Domain Name System uses an interesting approach to solve the problem: thousands of domain name servers. Each organization operates a domain name server that contains the list of all computers in that organization along with their IP addresses.† The organization has authority to change or update the set of names and IP addresses; there is no need to coordinate its changes with a central authority or a central server. For example, Purdue University runs a domain name server for names that end in *.purdue.edu*.

†An individual or small organization that does not have an IT staff can contract with an ISP to operate a server on their behalf.

16.11 Looking Up A Domain Name

The entire set of DNS servers are coordinated. In addition to knowing the organization's computers and their IP addresses, the DNS server at an organization also knows how to find the correct DNS server for other names. When it powers on, a computer only needs to know about one domain name server, a local server. When it needs to obtain an IP address, an app running on the computer sends a request to the local DNS server. The local server either knows the address for the name or knows how to contact a server that can supply the address. In either case, the local server obtains the information, and sends the answer back to the computer.

An example will help clarify the steps. Suppose a laptop user in France opens a browser and enters the name *www.purdue.edu*. Before it can download the web page, the browser must know the IP address of the Purdue computer. The browser sends a request to its local domain name server in France. Although it does not know the answer, the local server knows how to contact the domain name server at Purdue. The local server sends a request to the server at Purdue, obtains the answer, and returns the answer to the browser. After it receives the answer, the browser can contact the web server at Purdue and obtain the requested page for the user. Figure 16.3 illustrates the four steps needed before the browser can contact the web server.

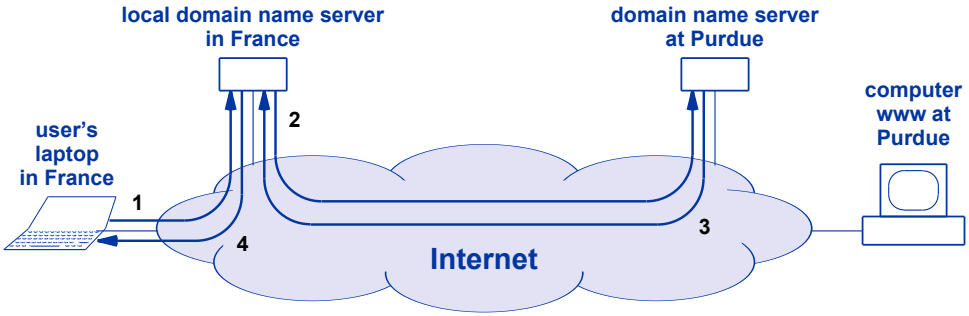


Figure 16.3 The steps taken to look up a domain name *before* a browser can contact the specified web server. The browser starts with its local name server; the local server contacts other name servers as necessary. Numbers indicate the order of the four steps.

The important idea is that each time you launch an app and specify a computer name or click on a link that contains a computer name, your browser must use DNS to look up the IP address of the computer with the specified name. Looking up a name involves sending multiple messages across the Internet, and it may seem like excessive overhead. However, the messages are extremely short, and the total time required to look up a name is trivial. Furthermore, once an app looks up a name, software on your computer remembers the answer for a short time, typically a week. Thus, if you visit the same web site three days in a row, your browser only needs to contact a domain name server one time.

16.12 A Personal Story About A DNS Problem

I was sitting in an airport using my laptop, when two people came up and asked me, "Is the Internet working for you?" I said it was, and they explained that they couldn't use it. After a little experimenting, I identified the problem: the local domain name server had crashed. I explained that the Internet was still working fine, but they could not look up a name and get an IP address. The two stared at me, and wondered how I could be working if they could not.

I explained that I was using a VPN† connection to one of the computers in my lab. Because the computer had a permanent IP address, I had configured the app to use the address instead of the name. I hadn't planned for the DNS server to be down, but I had observed that looking up a name sometimes caused a delay, so I had configured an address. Interestingly, it turned out that skipping the step of using DNS meant I was able to use the Internet that day.

16.13 Summary

The Domain Name System (DNS) assigns a name to each computer, and provides an automated system that apps use to translate a name into an equivalent IP address. A domain name contains multiple parts, called *labels*, that are separated by dots (periods). A computer's primary name comes first, and remaining labels form a suffix that designates the organization that owns a computer.

The DNS consists of many servers that are all coordinated to work together. To look up a name, an app sends a request to the local domain name server. If the local server knows the answer, the server returns the IP address to the app. If it does not know the answer, the local server contacts other domain name servers, as necessary, to obtain the answer, then returns the answer to the app that made the request.

EXERCISES

16.1 Does your computer know a local DNS server? If you have a Mac, select the Apple menu, then *System Preferences*, and *Network*. Click on *Advanced*, and select *DNS* from the list at the top of the page. You will see a panel labeled *DNS Servers* and numbers with dots in them. Each number with dots is the IP address of a local DNS server.

16.2 When a computer powers on, the computer is given the IP address of a domain name server. Some organizations use two servers, and the DNS software in a computer is configured to try the second server if the first doesn't respond. In what circumstances does having a second DNS server help?

†Chapter 32 explains the purpose of Virtual Private Network connections.

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

17

Sharing An Internet Connection (NAT)

17.1 Introduction

In the 1990s, Internet access transitioned from dial-up to broadband (DSL and cable modems). Most households only had one computer, and ISPs designed service for a single computer — if a household had multiple devices using the Internet, they were charged an extra fee for the additional devices. A group of creative engineers invented a technology that allows multiple devices to share one Internet connection and avoid the extra fees.

The technology that allows multiple computers to share an Internet connection is so widely used that now more devices connect to the Internet using the technology than connect in the original way. This chapter explains how the technology works and where you have encountered it.

17.2 Multiple Devices Sharing A Single IP Address

In the original Internet design, each host was assigned an individual IP address that differed from all other IP addresses. When the question first arose asking whether it would be possible to have multiple devices at a residence "share" a single address, the idea seemed preposterous. If two devices had the same address, how could datagrams be sent to the correct device?

A clever design solved the problem: instead of connecting all of a user's devices directly to a DSL or cable modem, insert an additional electronic device. The additional device is designed to connect multiple devices without requiring additional IP addresses.

17.3 Wireless Routers And NAT

The technology that allows devices to share an address is known as *Network Address Translation* (*NAT*). Because NAT is used in many ways, a variety of devices implement NAT technology. Informally, professionals use the generic term *NAT box* to describe such a device. The NAT device most familiar to consumers is known as a *wireless router*; we will use the term in the discussion that follows.

A wireless router connects user's devices to a user's ISP. That is, instead of plugging one device into a DSL or cable modem, a user connects the modem to the wireless router. Then, multiple devices can connect to the wireless router.

Despite its name, a wireless router usually offers both wired and wireless access. The router becomes a Wi-Fi access point to which one or more devices can connect, and also has a small number of Ethernet ports (typically four) that allow devices to plug in. Figure 17.1 illustrates the connections.

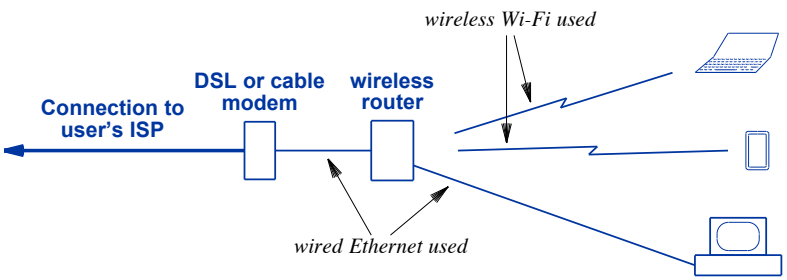


Figure 17.1 Physical connections when a wireless router is used.

17.4 How A Wireless Router Works

The secret of NAT technology arises from a two-part design and the use of *temporary IP addresses*. The two-part design allows a wireless router to act one way on the connection to the user's modem and a completely different way on the connections to local devices.

On the connection to the modem, a wireless router impersonates a single computer. When it powers on, the wireless router communicates with the modem, and is assigned an IP address, exactly as if a computer had connected to the modem. Once it obtains a valid IP address, the wireless router can communicate over the Internet. From the ISP's point of view, the user has connected a single computer to the modem.

On the wired and wireless connections to local devices, the wireless router impersonates an ISP. When a device connects, the wireless router assigns a temporary IP address to the device, exactly as if the device had connected to the modem. From a device's point of view, the device has connected to an ISP, and can communicate over the Internet. Figure 17.2 illustrates the idea.

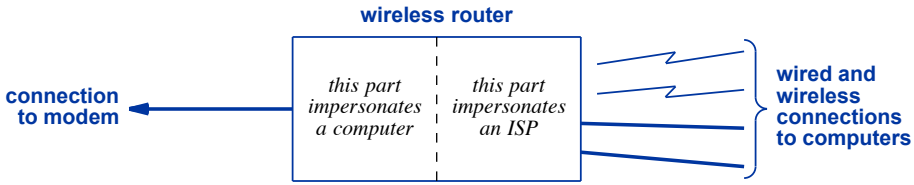


Figure 17.2 Illustration of the two parts of a wireless router that impersonate a computer and an ISP.

The key idea is:

> *By impersonating a computer, a wireless router obtains a valid IP address from the user's ISP. By impersonating an ISP, the wireless router assigns each local device its own temporary IP address.*

17.5 Datagram Modification

The temporary IP address that a wireless router assigns to a device is not valid on the Internet. So how can a device with a temporary address communicate across the Internet? The wireless router modifies each outgoing datagram before sending the datagram on to the Internet. When a local device sends a datagram, the source address in the datagram will be the temporary address the device is using. When the datagram reaches the wireless router, software replaces the source address in the datagram with the valid IP address that the wireless router obtained from the modem. In essence, before a datagram goes out to the Internet, the wireless router replaces the "return" address with the valid IP address that the wireless router obtained when it powered on. When a host on the Internet replies, the reply will return to the wireless router.

If all outgoing datagrams have the same reply address, how can a wireless router know which of the local devices should receive a reply that arrives from the Internet? The wireless router keeps a list of what each device is doing (e.g., which web site is being accessed or which file is being downloaded). When a datagram arrives from the Internet, the wireless router consults its list to determine which device should receive the datagram. The wireless router replaces the "to" address in the datagram with the temporary address that has been assigned to the device.

NAT technology is complex, and you do not need to understand the details. The important point is:

By keeping a record of what each local device is doing and replacing addresses in incoming and outgoing datagrams, a wireless router allows local devices to communicate across the Internet simultaneously, even though the ISP has only assigned one valid IP address.

17.6 Your Device Can Act Like A Wireless Router

Many devices include software that allow the device to fill the role of a wireless router. Popular operating systems, including Windows, Mac OS, and Linux, all include the necessary software. The Microsoft version is known as *Internet Connection Sharing*, and the Mac version is part of the *airport* software. Using the software requires some expertise, but the idea is straightforward:

- Connect your device to an ISP, typically one connects an Ethernet cable from the device to a DSL or cable modem.
- Configure the Wi-Fi on your device to act like a wireless router that has an SSID and allows other devices to connect.
- Configure NAT software to assign each local device a temporary IP address and to modify datagrams that pass between local devices and hosts on the Internet.

We can summarize:

Once it has been configured, the software on your device will behave exactly like a wireless router.

17.7 You Probably Use NAT Every Day

When you use Wi-Fi to connect to "free wireless Internet" at a coffee shop, shopping mall, airport, hotel, or other venue, you are using NAT. In each of those situations, the proprietor has installed a wireless router. The router connects to the Internet, and assigns a temporary IP address to each customer's device. In addition, many ISPs offer modems that include a built-in wireless router. If your modem includes Wi-Fi, it is running NAT.

Surprisingly, cell phone systems also use NAT. When you connect a smart phone to the cellular system (e.g., using 4G or 5G), your phone is assigned a temporary IP address. The phone system modifies datagrams you send exactly the same way a wireless router in your home modifies datagrams. The only difference is that the equipment

used by the cellular phone system is designed to handle many more customers at high data rates, and goes by the name *Carrier NAT*.

The point is:

> *Even if you do not have a separate wireless router at home, you probably use NAT technology daily, either when you access a Wi-Fi hot spot or when you connect a smart phone to the cellular network.*

17.8 Why Internet Size Is Difficult To Estimate

Chapter 8 points out that starting in 2006, estimates of Internet size have become inaccurate. We now understand why. Surveys count permanent IP addresses as a method of counting devices. With NAT, however, multiple devices (especially smart phones, laptops, and tablets) can share a single IP address. As a result, it is impossible to know exactly how many devices are accessing the Internet.

17.9 Summary

A technology known as Network Address Translation (NAT) allows multiple devices to share a single Internet connection. One form of NAT uses a wireless router that connects to an ISP and then allows devices to connect to it. To an ISP, a wireless router appears to be a single computer; to each local device, a wireless router appears to be an ISP. The wireless router keeps a record of what each local device is doing, so the wireless router can modify the addresses in datagrams as they pass from local devices to the Internet and from the Internet back to local devices. Most operating systems allow a user to configure their device to perform the same functions as a wireless router.

NAT technology is widely deployed, and most users encounter NAT every day without knowing it. Venues that offer free Wi-Fi use NAT, as does the cellular phone system.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

18

Why The Internet Works Well

18.1 Introduction

Previous chapters describe the basic Internet technology, including TCP/IP software. This chapter considers reasons for the Internet's success and the lessons that can be learned.

18.2 The Internet Works Extremely Well

The Internet is a marvel of technical accomplishment. The basic idea and the TCP/IP technology has accommodated growth and changes that the original designers did not imagine. The number of computing devices attached to the Internet has grown from a few dozen to billions. Traffic on the Internet has also grown exponentially. Meanwhile, both the basic design and TCP/IP software technology has accommodated the increases. Although modern smart phones operate several thousand times faster than the computers that existed when the Internet was designed, new computers can communicate across the Internet with each other and with older computers.

Why is the Internet and the underlying technology so successful? How could a technology from a research project become the foundation of the world's largest communication system? What lessons have we learned from the Internet project? Obviously, no single technical decision results in the overwhelming success of a complex system like the Internet. However, a poor design choice can ruin an otherwise excellent plan. Remaining sections of this chapter examine some of the best design choices.

18.3 Flexibility To Accommodate Arbitrary Networks

The Internet designers did not attempt to design a new type of network that filled all needs. Instead, the designers assumed that many types of networks would be used, and provided a flexible system that can interconnect a wide range of underlying network hardware. For example, the Internet can accommodate:

- Wide Area Network technologies as well as Local Area Network technologies
- Network hardware that has extremely high capacity as well as network hardware that has extremely low capacity
- Network technologies that guarantee no packet loss as well as best-effort network technologies that do not compensate for packet loss
- Wireless networks that use radio waves for communication as well as wired networks that send signals across copper cables and optical networks that use light to send signals across glass fibers
- Networks that use satellites in orbit around the earth as well as terrestrial networks that send information along the earth's surface

The point is that Internet technology was designed to accommodate almost any type of computer communication technology. More important, by making the design flexible, the Internet was able to accommodate new network technologies when they appeared, even new types that did not exist when the Internet was first designed.

The secret of the Internet's flexibility stems from a tolerant approach. Because it does not demand much from the network hardware, the Internet Protocol tolerates almost any mechanism that can send bits from one location to another.

Because it makes very few demands of the underlying hardware, the Internet accommodates any type of network.

18.4 Flexibility To Accommodate New Apps Quickly

The Internet took a new approach to building a communication system by placing all services outside the network. The idea of using computers to run all services was controversial, but turned out to be ideal. Because the Internet only provides datagram transfer, a new service can be added without modifying software in routers. As a result, new services can be created at any time.

Placing all services in computers attached to the Internet was a brilliant idea that makes new services trivial to deploy and encourages innovation.

Although the Internet designers did not know what new services would be created, they decided that services would use the client-server form of interaction. Now, after many decades of experience building Internet services, the decision seems obviously correct. At the time, however, when networking was so new that even computer scientists had not built applications that used a network, it was not clear that client-server interaction would suffice for any service.

18.5 The Advantage Of Being Open And Vendor Independent

Communication between two computers requires both to agree on the rules for communication. Much of the Internet's success can be credited to a design that is *open* and *vendor independent*. Unlike earlier networking technologies that were designed for one vendor's computers, the Internet was designed to provide communication among arbitrary types of computing devices.

To ensure compatibility among arbitrary devices, the technical specifications for the Internet were written to be completely independent of any specific devices, and were open, meaning anyone could use the specifications to build Internet products and services without paying a fee. The standards documents specify how to send IP datagrams from any type of device over any type of network. Whenever a new network technology appears, a new standard document is written that describes how to use Internet technology with the new hardware. The specifications form an important part of making the Internet work well because they guarantee that all devices and all routers use exactly the same format when sending a datagram across a network. The point is:

> *Because Internet technical documents specify the exact way to send IP datagrams from all types of devices across all types of networks and because the documents are open, devices and routers from multiple vendors always agree on the communication details.*

18.6 An Extremely Efficient Design

TCP and IP form a complementary pair that work together well. IP provides a basic communication system that allows any host on the Internet to send datagrams to any other host. TCP handles all the communication problems that IP does not, including the difficult problems of packet loss and congestion. The result is an efficient, reliable communication system.

In any complex computer system, engineers must choose among a variety of possible designs, the TCP/IP protocols were carefully designed to be efficient. Instead of depending on powerful computers, the designers worked to ensure that Internet software would run well on the slowest, smallest devices. Thus, we now have tiny devices, such as smart home thermostats and appliances connected to the Internet. Each device contains a miniature computer that has a slow processor and small memory. Such devices

are able to communicate over the Internet because the protocol software is designed for extreme conditions.

Because they are designed to be efficient, Internet protocols can run on small, inexpensive devices like lighting and heating system controls used in a smart home.

18.7 Packet Switching Is A Fundamentally Better Idea

Before the Internet, telephone companies around the world had built their networks on the idea of dedicating a pair of wires to each phone call, in essence allowing a caller to lease the wires during the call. The Internet used packet switching, and allowed multiple senders and receivers to share underlying networks. At the time the Internet was designed, packet switching was relatively untried and controversial. Telephone companies denounced the design, claiming that packet switching would never scale. They repeatedly pointed out that they were the only ones who had experience designing a large communication system.

By the 1990s, it became obvious that the Internet approach would indeed handle large scale. The surprise, however, came from economics: the Internet's packet switching technology was much cheaper than the telephone system's technology. Some of the economic savings arose because the telephone system had built all the services into telephone switches. Adding a new service or changing an existing service required waiting years while all the switches were upgraded. In contrast, the Internet design placed services in computers outside the network, allowing services to change quickly and at extremely low cost. By 2000, telephone companies realized that they could replace their expensive design with Internet technology. AT&T announced that it would never buy another telephone switch, but would instead switch to using Internet technology and deploy IP routers inside the telephone network.

The point is:

The Internet's use of packet switching resulted in a communication technology that is much less expensive than the approach used in the original telephone systems. By 2000, Internet technology had edged out the competition, and even the telephone companies — some of the staunchest critics — had switched to using IP routers.

18.8 Can The Success Be Replicated?

Many people who see the success of the Internet pointedly ask, "How can we repeat the success?" Businesses look at the economics. Researchers wonder if a new idea might lead to a revolution as significant as the Internet.

Of course, one never knows whether a potential "breakthrough" will change everything, but it will be difficult to replicate many of the factors that led to the Internet's success.

- *Top People*. Researchers who worked on the Internet project were selected from the very best that DARPA could find. Even among peers — other researchers in Computer Science — the group stood out as especially talented.
- *An Inspiring Problem*. The Internet was a dream that inspired and challenged the best researchers. In the 1970s, many of the goals seemed impossible. Instead of merely combining existing technologies, researchers working on the Internet had to redefine networking and invent entirely new approaches.
- *Tireless Effort*. Most research projects span one to three years. The Internet project started in 1973, and was not taken over by commercial interests until 1989. During that time, researchers persisted against all odds. When something didn't seem to work well, they dug in and discovered new ways to handle the problem.
- *No Economic Constraints*. Unlike the current research environment where both companies and universities look for short-term economic payoffs, DARPA funding allowed Internet researchers to focus on science and engineering without pushing them to start companies or license their inventions. Indeed, Internet research often tried new, innovative technologies that were more expensive than existing commercial equipment, and researchers worked in isolation with no one demanding products.
- *Elegant, Minimalistic Design Instead Of Features*. In an engineering effort, it is always tempting to add more features. However, the Internet researchers asked a much harder question: What is the minimum set of mechanisms needed to solve the problem? The Internet protocols are much smaller and much more elegant than the protocols designed by other groups, and yet they handle more complex problems.
- *Reducing Ideas To Practice*. Many researchers dream up something that might work, and then publish a research paper that describes the idea. Internet researchers insisted that each part of the technology work well in practice. Even after a technical specification was drafted, further testing was mandated before the specification was accepted as a standard. Three independent teams would implement software according to the specification, and then test that all three interoperated correctly.

18.9 Summary

The Internet represents an incredible technical accomplishment. Although careful planning and attention to detail contributed to its success, other factors were important, including: choosing top researchers, using packet switching even though telephone companies insisted that their approach was better, accommodating all types of networks, having a long time for research with no pressure for economic payoffs, and an agreement among researchers to demonstrate all ideas with a practical, working system.

Internet Services

Examples of services along with an explanation of how they work

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

19

Electronic Mail

19.1 Introduction

This chapter begins a discussion of example services available on the Internet. It examines one of the most widely used services: electronic mail. Successive chapters explore other services. In each case, the text explains the underlying mechanism.

19.2 Functionality And Significance

Electronic mail (*email*) was originally designed to allow a pair of individuals to communicate via computer. The first electronic mail software provided only a basic facility: it allowed a person using one computer to type a message and send it across the Internet. Later, the person to whom the mail addressed could access the message.

Current electronic mail systems provide services that permit complex communication and interaction. For example, electronic mail can be used to:

- Send a single message to many recipients (i.e., a *mailing list*)
- Send a message that includes text, audio, video, or graphics
- Have a computer program generate and send a message
- Have a computer program respond to an incoming message

The significance of email arises from its widespread use in the business community. Although text messages and social media have become popular forms of communication, most businesses still use email as the primary communication platform.

19.3 Mailboxes And Email Addresses

To receive email, a user must have a *mailbox* and an *email address*. A user's mailbox consists of a storage area that holds email messages sent to the user until the user accesses them. The mailbox is located on a computer that runs software which accepts incoming messages and places them in the appropriate mailbox.

When email was designed, the question arose: what form should be used for an email address? To understand the design that was chosen, remember that email predates personal computers, and was designed when large computers were shared by multiple users. Consequently, an email address identified two items: a computer to which a message should be sent, and a user on that computer. The syntax used an at sign to separate the two items, resulting in email addresses of the form:

user@computer

19.4 Sending An Email Message Directly

When a user composes an email message, the user specifies one or more recipients. Once the user clicks *Send*, email software on the user's computer sends a copy to each recipient. In the original design, email software would extract the *computer* name from the recipient's email address, contact the email server on the computer, specify the user to whom the message should be delivered, and send a copy of the message.

As an example, consider email sent to:

comer@purdue.edu

Email software on the sending computer extracts the domain name, *purdue.edu*, obtains the IP address, and then contacts the email server on *purdue.edu*. The email software specifies the recipient *comer*, and sends a copy of the message. The server on *purdue.edu* stores the message in the appropriate mailbox.

19.5 Personal Computers And Email Providers

The direct delivery method described above worked well because large, shared computers always remained running. An email server was started when the computer was powered on, and remained running, ready to receive mail at any time. The arrival of inexpensive personal computers in the 1980s and 1990s changed the way people used computers, and forced a change in email. Because individuals would often power down their personal computer when it was not in use, other users could not send email unless the recipient happened to be using their computer.

To accommodate personal computers, a new set of companies emerged known as *email providers*. The idea is straightforward: instead of placing a user's mailbox on a

user's personal computer, the mailbox is placed on a computer run by the email provider. A provider's computer always remains running, and can accept email for the user at any time. Later, when the user wants to access their email, they run an app that contacts the provider's computer and accesses their mailbox.

The new paradigm for email means that an email address no longer refers to a user's computer. Instead, an email address now identifies a mailbox and a provider:

mailbox@provider

19.6 An Example Email Exchange

An example will help explain how an email transfer occurs when a provider is used. Assume Bob is a customer of Provider 1, and has the email address *bob76@provider1.com*. Also assume Alice is a customer of Provider 2, and has the email address *aliceb@provider2.com*. Suppose Bob sends an email message to Alice. Figure 19.1 illustrates the steps taken, which are explained below.

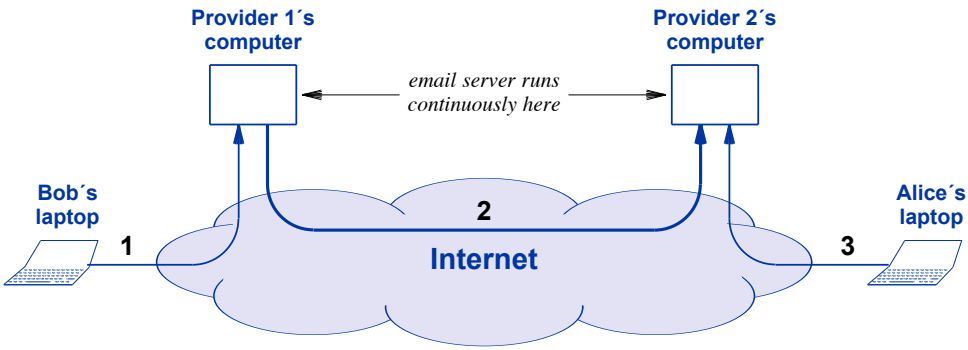


Figure 19.1 Illustration of the steps taken when Bob, who uses Provider 1, sends an email message to Alice, who uses Provider 2.

The three steps shown in the figure are:

1. **Bob creates a message.** To send the message, Bob contacts his provider, either by using an email app or a web browser. Bob types a message, and specifies *aliceb@provider2.com* as the recipient.
2. **The message is transferred.** When Bob clicks *Send*, the email software running on *provider1.com* contacts the email server on *provider2.com*, specifies *aliceb* as the recipient, and transfers the message.
3. **Alice reads the message.** Later, Alice launches an email app or uses her browser to contact her provider. Alice finds the message from Bob in her mailbox, and reads it.

To summarize:

Modern email systems arrange for a user's mailbox to be located on a computer run by an email provider. Because providers run their computers continuously, a provider can accept incoming email at any time, and a user can access their mailbox at any time.

19.7 Email Delays And Retry Attempts

The second step in Figure 19.1 helps explain a common email problem: delay. Suppose Bob and Alice were talking on the phone when Bob sent Alice an email message. Bob tells Alice that he sent the message, and Alice complains that the message has not arrived. How can email take so long to travel across the Internet? As the figure shows, once Bob sends the message, email software on Provider 1's computer must transfer a copy to the email server on Provider 2's computer. Unfortunately, each server has limits on the number of messages it can receive simultaneously. Once the limit has been reached, any further attempts to contact the server will be rejected.

Email software is configured to retry message delivery that fails. So, if a server becomes overwhelmed with requests and rejects contact, the sending email system will retain a copy of the message and retry delivery later. However, instead of retrying immediately, most email systems wait many minutes, perhaps an hour between attempts to give the server time to handle its current backlog. Email systems usually retry for days before they declare that delivery is impossible. The point is:

If the email server that is operated by a recipient's provider becomes busy, a long delay can occur between the time an email message is sent and the time the message is delivered to the recipient.

19.8 Providers, Fees, And Access

The cost to operate an email service is extremely low because email does not take much storage space and handling email does not consume extensive amounts of processor time. Therefore, a provider can use low-cost hardware. Thus, if a provider charges for email service, the monthly fee is usually small. Some email providers offer email accounts for users at no change. The agreement a user signs allows the provider to watch email messages, deduce the user's preferences, and sell the information to advertisers, who then send the user ads.

In terms of access, the steps listed under Figure 19.1 show that both Bob and Alice need to contact their email provider before they can send and receive email. Most email providers arrange access through a web browser. The user enters the provider's URL, logs into their email account by typing a password, and can then process the messages

in their mailbox or create and send a new message. Apps are also available that provide email access; when a user creates an email account, the provider issues instructions on the access method to use.

19.9 Mailing Lists

Most email servers allow the owner to create and use a *mailing list* (*email list*). Each mailing list has a name, and contains one or more email addresses. Many organizations define mailing lists that correspond to subgroups of the organization. For example, a tech company might set up a mailing list named *sales* that includes the email addresses of all employees who work in sales, and a list named *engineering* that includes email addresses of all employees who work in engineering. A mailing list can be *public*, meaning that anyone can send a message to the list, or *private*, meaning that only users on the list can send messages to the list.

When it receives an incoming email message, the server examines the name of the recipient to which the message was sent. If the recipient is the name of one of the user mailboxes, the server adds the message to the mailbox. If the name of the recipient is in error (e.g., the sender has mistyped an email address), the server tells the sender that an error has occurred. Finally, if the name of the recipient matches one of the mailing lists, the server handles the details of sending a copy of the message to each address on the mailing list. Informally, professionals say that the mail software acts as an *exploder*.

19.10 Undisclosed Recipients

Each mail message begins with header lines that specify the sender, recipient(s), and a subject. A *From:* header specifies the sender's email address. Three email headers identify recipients: *To:*, *Cc:*, and *Bcc:*. The names are taken from headers used on office memos before email was invented, when *Cc:* abbreviated *Carbon copy*, and *Bcc:* abbreviate *Blind carbon copy*. The interpretation is:

To: lists the email addresses of the main recipients

Cc: lists email addresses of recipients who should receive a copy

Bcc: lists email addresses of recipients who should receive a copy, but whose identity should be hidden from other recipients

Email software treats addresses in the *To:* and *Cc:* lists the same; the two headers are merely meant to help recipients understand the intent of the sender. However, the *Bcc:* list is hidden from other recipients. That is, when it sends a copy of the email message to a user, the email software omits the *Bcc:* list. Thus, if a user receives a copy of an email message that does not list their email address in either the *To:* or *Cc:* lists, the user can deduce that other recipients will not know they received a copy.

If a sender wants to hide all recipients from one another, the sender can specify all recipients in a *Bcc:* header, leaving the *To:* and *Cc:* headers blank. Each recipient will receive an email message that lists no recipients. To prevent confusion, some email software is configured to fill in the *To:* header of such messages with a phrase, such as *Undisclosed Recipients*. The phrase often appears in spam messages. For example, suppose a spammer sends the following message:

```
To:
Cc:
Bcc: marthas@vinyard.com, george@washington.com, pikes@peak.com
Your email address has been selected at random to win a $5,000,000
lottery.  Please reply to this email with all your bank account info.
```

Each of the three recipients would receive a message of the form:

```
To: Undisclosed Recipients
Cc:
Your email address has been selected at random to win a $5,000,000
lottery.  Please reply to this email with all your bank account info.
```

From a spammer's point of view, the use of *Bcc:* means that a given recipient will not know how to contact other recipients, so it will not be as easy for them to deduce the message is a fraud.

19.11 Summary

Although other Internet services are popular among casual users, email remains the primary communication mechanism used in business. Consequently, the Internet carries a significant amount of email traffic.

In modern email systems, a user's mailbox is stored on a computer operated by the user's email provider. An email address contains two items separated by the "at" (@) sign: a mailbox name and the name of a provider's computer. To send and receive email, a user accesses email software on their provider. Once a message has been sent, software transfers a copy to each recipient, retrying automatically if the recipient's email server is busy. Recipients can be specified in *To:*, *Cc:*, and *Bcc:* headers.

EXERCISES

19.1 Try sending an email message to a friend using only the *Bcc:* header. What does your friend receive?

19.2 Jane complains that an email message from Sue didn't arrive until a day after Sue sent it, and Sue says that Jane must be wrong because "electronic means instantaneous." Could Jane be correct? Explain.

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

20

The World Wide Web: Browsers And Basics

20.1 Introduction

This chapter and the next two explore one of the most widely used Internet services: the World Wide Web. This chapter introduces the concept of hypermedia, and describes how documents are linked together. The next chapters explain web documents and some of the more advanced web technologies.

20.2 Browsers And Web Servers

Like all Internet services, the World Wide Web is not built into the Internet. Instead, the service runs on computers attached to the Internet, and follows the client-server form of interaction explained in Chapter 15. Many web servers attached to the Internet store information. To access the information, a user launches a *web browser*. The browser acts as a client that contacts one or more servers to obtain the requested information, which it then displays for the user.

The information on a web server is divided into *web pages*, and a browser fetches one page at a time. In most cases, a given web server stores a set of related pages, and we use the term *web site* to refer to the entire collection of pages on a server.

20.3 URLs And Their Meaning

How does a browser know where to find information? A user must specify the correct web server and a web page on the server. To do so, a user enters a *Uniform Resource Locator* (*URL*). A URL is a string of characters divided into several parts by punctuation characters. Figure 20.1 illustrates how a URL is divided, and gives the meaning of the three most important parts.

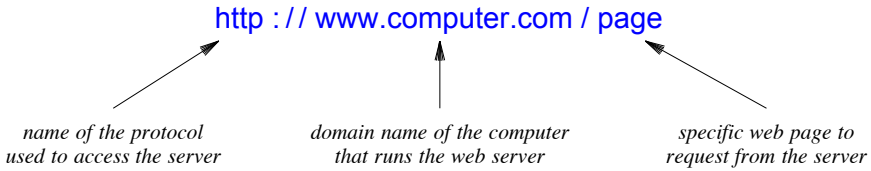


Figure 20.1 The three primary parts of a URL and their meaning.

As the figure shows, the first part of a URL specifies a protocol to use when contacting the web server, the second part specifies the domain name of the server to contact, and the third tells which web page on the server to request. The string :// separates the protocol from the server name, and a slash separates the server name from the name of a specific web page.

20.4 Web Pages With Links To Other Pages

The most interesting aspect of the Web arises from its use of *hyperlinks*. A given web page can contain text with links embedded on the page that point to other web pages, allowing a user to navigate from one page to another by clicking on a link. When it displays a page, a browser highlights any text that corresponds to a link, typically by changing the color and underlining.

As an example, consider a small web site that has six web pages devoted to information about the New York Stock Exchange. Each page contains text with embedded links. The following paragraph shows how a browser might display text that contains two links.

The New York Stock Exchange is a world-renown center of financial activity. Located on Wall Street in downtown New York City, the stock exchange allows stock brokers to buy or sell shares of stock electronically. The exchange provides the current price of stocks as well as the total number of shares that have been used as a measure of financial activity. Many online sites update their list of stock prices continuously.

Figure 20.2 illustrates how links in a web page can point to other web pages. In the figure, all text is covered in gray except the page title and links.

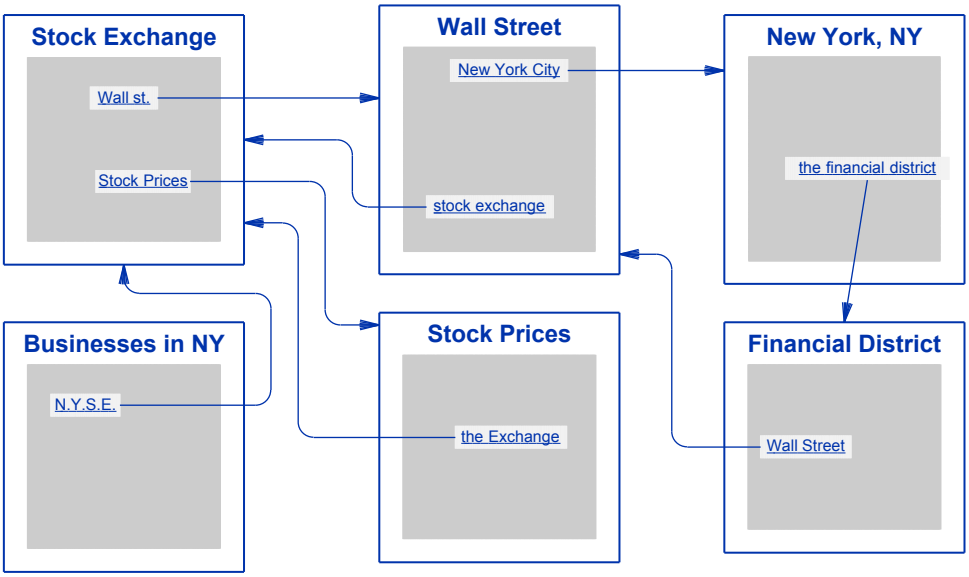


Figure 20.2 An example web site that has six web pages with links pointing to other web pages.

20.5 Linking Across Web Servers

Figure 20.2 illustrates that pages on a web server can contain links that point to other pages on the server. The power of the Web arises from the ability of a hyperlink to span multiple servers — a page on one server can contain a link to a page on another server. We will learn that each link is a URL that can reference a server in an arbitrary computer and can name a web page on the server. Because they only see a clickable item on the screen, users remain unaware of the server to which a given link points. Figure 20.3 illustrates the concept.

Figure 20.3 An illustration of links in a web page on one web server that reference items on other servers.

The key idea is that a user does not need to take any special actions to follow a link from one web server to another. In fact, a user can follow a series of links without knowing whether the web pages come from one server or from many. More important, because a browser can obtain a new web page quickly, a user may not even notice a difference in access delay. Thus, from the perspective of a user:

> *The World Wide Web hides the boundaries among web servers and makes information on a large set of servers appear to be part of a single, integrated system. A user can follow links without knowing or caring which server or servers supply the web pages.*

20.6 Hypermedia

In the 1980s, most computers could only display text. Browsing systems that used hyperlinks among pages on multiple servers were known as *hypertext* systems. In the 1990s, when the Web was invented, computers had gained the ability to display graphics images and to play sounds. Consequently, early web technologies expanded the idea of hyperlinks to include both text and graphics. We use the term *hypermedia* to mean that a web page can reference multiple forms of information (e.g., a link can reference a graphics image on a server). Current web technologies allow links that reference a variety of items, including:

Text	in various fonts, sizes, and colors
Images	of various types, shapes, and sizes
Audio clips	that play automatically or manually
Video clips	that play automatically or manually
Streams	of audio or video that play continuously

To understand hypermedia links, we need to expand our definition of a URL. Instead of specifying a web page on a server, imagine that the URL specifies an arbitrary item. Each of the items listed above can be stored on a server and referenced with a URL. Figure 20.4 illustrates the idea of links that point to non-text items.

Figure 20.4 Illustration of a web page with links that point to non-text items.

20.7 A Page With Multimedia Items

In addition to allowing a link to reference an arbitrary item, the Web permits a single web page to contain multiple media types (e.g., images along with text). Web technologies allow any item on the page to be a link. Furthermore, a link can specify more than another web page. For example, Figure 20.5 illustrates a page that contains both text and graphics, and a link that allows the user to send email. If a user clicks on the link *service@hrocker.com*, the browser will invoke the user's email app.

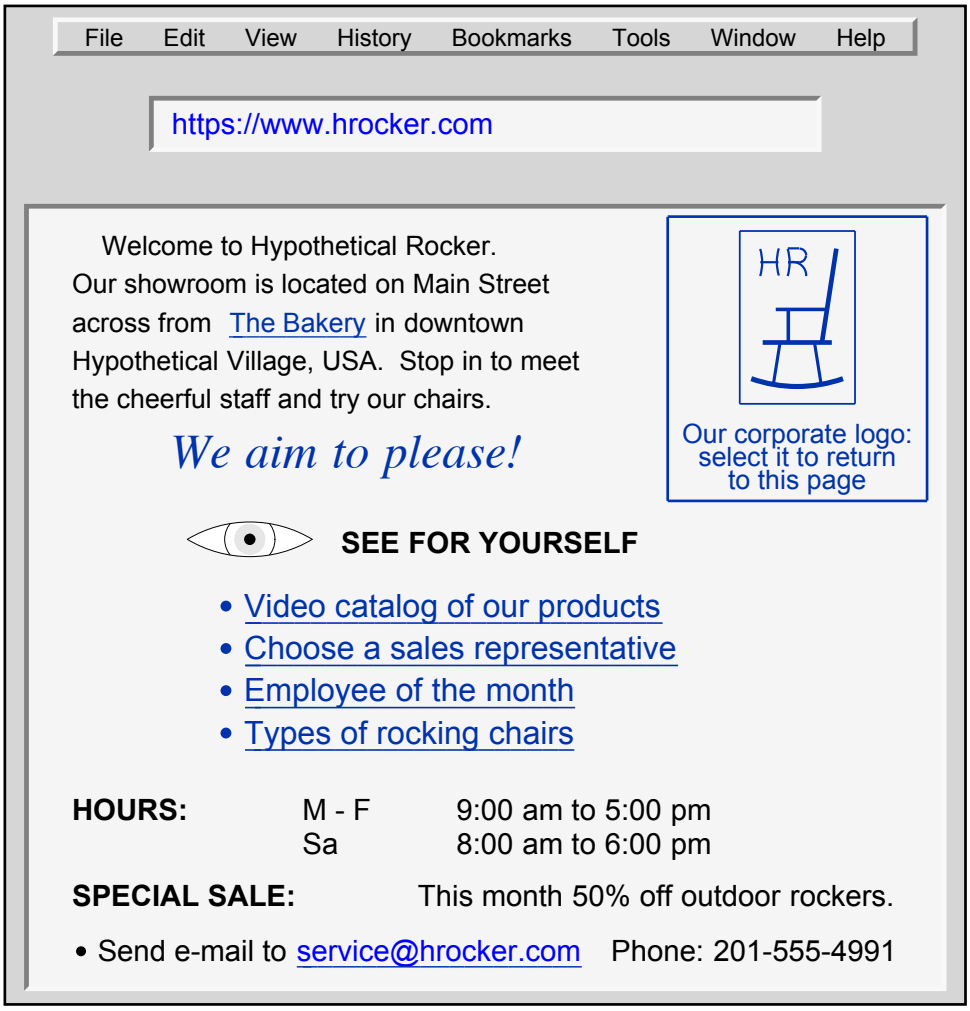


Figure 20.5 Illustration of a web page that contains text, a graphic image and a link that invokes an email app.

20.8 Fetching A Page That Contains Multiple Items

Consider a web page similar to the one in Figure 20.5 that contains a graphics image as well as text. When a user enters a URL (or clicks a link that points to the page), the user's browser contacts the web server and downloads the specified page. Surprisingly, because the page contains non-text items, the downloaded version does not include all the items to be displayed. Instead, the page contains text plus a URL for each multimedia item that specifies the location of the item. Thus, before it can display the page, the browser must fetch each of the other items on the page, assemble them into a single page image, and then display the final result for the user.

The items on a page might all be stored on the same server as the page, but it is common for some items to be stored on another server. For example. if a web page contains both a news story and an advertisement, the ad might be stored on a separate server. A browser needs to contact a server for each item on the page. Figure 20.6 illustrates the steps taken if a page contains an image stored on another server.

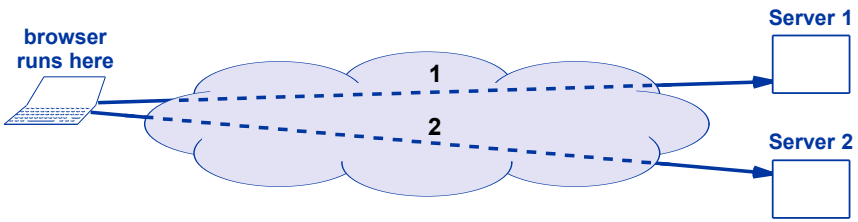


Figure 20.6 Illustration of a browser loading a page from Server 1 that contains an image located on Server 2. The browser takes two steps to load the page.

The next time you look at a web page covered in ads, remember that your browser had to fetch each ad separately. If a given ad also plays audio or contains multiple images, your browser also had to fetch each of them individually. The point is:

> *Although a user thinks of a web page as a single item referenced by a URL, a page often contains multiple items, and a browser must fetch each item individually.*

20.9 Inside A Browser

A browser is a complex piece of software that has many capabilities. For example, a browser understands how to download and display information from a local file on your computer, a remote web server, or a remote file storage server. A browser also understands how to launch an email app that can send email. The first item in a URL

tells the browser what to do. For example, *http* specifies that the browser should use the *HyperText Transfer Protocol* to download a web page, *https* specifies that the browser should use a secure version of http to download a web page, and *ftp* specifies that the browser should use the *File Transfer Protocol* to download a file. Figure 20.7 illustrates that a browser contains the software needed to handle a request.

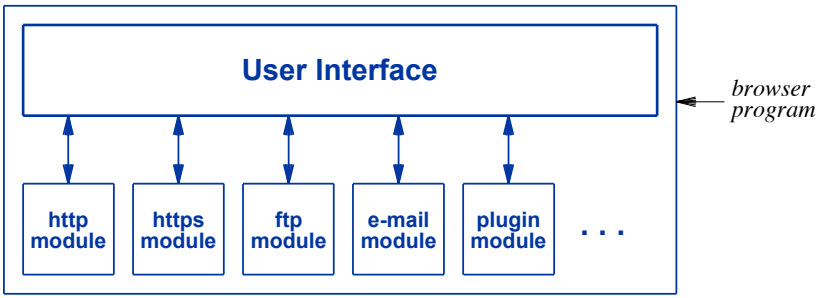


Figure 20.7 The internal structure of a browser with a software module for each type of request.

20.10 Plugins And Other Add-on Software Modules

As Figure 20.7 shows, the *User Interface* constitutes a major piece of the browser that always runs and allows a user to enter a URL or click a link to move to a new page. Other software modules are invoked when needed.

Browsers offer an interesting feature that increases their generality: extensibility. When a new Internet service is invented, the browser software does not need to be rewritten to include a module for the service. Instead, a user can install a module as needed. We use the terms *plugin* or *add-on* to describe additional software that has been added to a browser. Many plugins exist. For example, when animated graphics first appeared, browsers did not have a software module to display the animation. Companies that sold animation software each created a plugin module that could display their animations.

> Plugin (add-on) *technology allows a user to enhance a browser by installing additional software modules that handle new services and new file formats.*

20.11 Historical Notes

Hypertext systems were around before the World Wide Web was invented. One of the earliest services was known as *gopher*.† The first browser for the World Wide Web was developed at the National Center for Supercomputer Applications (NCSA). The browser was named *Mosaic*, and led to *Netscape*, which later became *Firefox*.

20.12 Summary

The World Wide Web, which has become the most widely used Internet service, is a hypermedia system that allows a user to follow links from one web page to another. Each web page is stored on a server, and a link can cross from one server to another. The resulting system appears to be a seamless interconnection of web pages. Each web page is identified by a URL that specifies a web server and a specific item on the server.

A browser is an app that provides web access. A browser contains user interface software plus additional software modules that each handle a specific service. A user can install additional modules called *plugins* to extend a browser when new services or new file types are invented. When a user requests a page that contains multiple items, a browser must fetch each item separately.

†The name was chosen because the system was designed at the University of Minnesota, home of the *golden gophers*, and as a pun because the system was designed to "go for" information.

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

21

The World Wide Web: HTML And Web Pages

21.1 Introduction

The previous chapter describes the World Wide Web, and discusses how a browser fetches web pages. This chapter examines the internal representation used in web pages. It shows the language used to create a web page, and explains how a multimedia document can be created that contains items such as text and graphic images, and how a link can be created that points to another web page.

Why should one learn about the internal representation used for web pages? After all, a browser completely hides the internal details from a user. There are two reasons. First, learning a few basic concepts can help explain the idea of hypermedia and remove much of the mystery from web pages. Second, learning about the internal language will show how much detail a programmer needs to specify when creating a web page.

21.2 Accommodating Display Hardware

The display hardware used with computers varies widely, with the size and resolution of a display depending on cost. Rather than have a version of each web page for each type of display, designers chose to make web pages give general layout guidelines and allow a browser to choose how to display the page on a given computer. Thus, a web page does not give many details. For example, the author of a web page can specify that a group of sentences form a paragraph, but the author cannot specify details such as the exact length of a line or whether to indent the beginning of the paragraph.

Allowing a browser to choose display details has an interesting consequence: a web page may appear differently when viewed through two browsers or on two computers that have dissimilar hardware. If one screen is wider than another, the length of a line of text or the size of images that can be displayed differs. The point is:

> *A web page gives general guidelines about the desired presentation; a browser chooses details when displaying a page. As a result, the same web page can appear slightly different when displayed on two different computers or by different browsers.*

21.3 HTML, A Language Used For Web Documents

Although the computer language used for web pages is a high-level language, it is not a natural language such as English. Instead, each web page is written in the *HyperText Markup Language* (*HTML*). Like other computer languages, HTML has rules of grammar, and uses conventional punctuation symbols in unusual ways. HTML is designed to make it easy for a computer to process the language, but some of the details make it difficult for a human to read or understand. In particular, HTML is unlike a word processor because the specification of a web page does not appear the same to its author as when it is translated and displayed by a browser. For example, HTML allows text to contain extra spaces and to be divided across many lines. Figure 21.1 illustrates the idea by showing part of a web page in HTML and the resulting output when a browser displays the page.

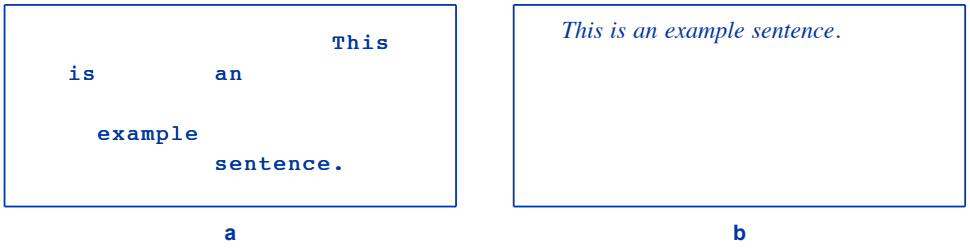


Figure 21.1 (a) Part of a web page in HTML, and (b) the result when a browser displays the page. A browser removes spaces before displaying the text.

We say that HTML uses a *free format* approach. The motivation should be clear: because HTML does not specify exactly how to display the output, a browser has the freedom to choose a form that is appropriate for a given computer.

21.4 Specifying Formatting Guidelines

Although HTML uses free-format input, a web page can contain guidelines that tell a browser how to interpret or display the page. For example, an author can tell a browser to center a specific line of text and to display items in a list. Each formatting instruction consists of a command surrounded by less-than and greater-than symbols. We use the term *tag* to refer to a command. For example, the HTML tag **
** instructs a browser to begin a new line† and the tag **<P>** begins a new paragraph. The convention of using less-than and greater-than symbols to distinguish tags from ordinary text works well because common English syntax does not place these symbols around words.

As an example of using tags, consider Figure 21.2, which shows the HTML for three paragraphs of text and the result when displayed by a browser.

```
This   is   an example of text on a Web page that consists
of three paragraphs. The first paragraph only contains a
couple of sentences(two, in fact) . <P> The
                second
paragraph also contains two sentences.    A
sentence can          span multiple input lines, and can
     contain additional      spaces; a browser ignores such
spacing   when   displaying    the        paragraph.
<P>
   This is the final paragraph.
   In HTML, a tag separates each pair of paragraphs.
   On the screen, however, vertical blank space
   separates paragraphs as in a textbook.
```

a

This is an example of text on a Web page that consists of paragraphs. The first paragraph only contains a couple of sentences (two, in fact).

The second paragraph also contains two sentences. A sentence can span multiple input lines, and can contain additional spaces; a browser ignores such spacing when displaying the paragraph.

This is the final paragraph. In HTML, a tag separates each pair of paragraphs. On the screen, however, vertical blank space separates paragraphs as in a textbook.

b

Figure 21.2 (a) An HTML document that contains three paragraphs of text, and (b) the result when the example section is displayed by a browser.

†The characters *BR* were chosen because the printing industry uses the technical term *line break* to refer to the beginning of a new line.

21.5 A Link Embedded In A Web Page

The previous chapter describes links that allow a user to move from one web page to another. How are such links embedded in a web page? HTML uses a pair of tags to surround each item that forms a link. The tags can surround a single word, a phrase, or other objects (e.g., an image). When a browser finds the tags, it marks the items as a link. Typically, a browser underlines or highlights the marked items.

In HTML terminology, items on a page that correspond to a link are said to be *anchored*. The character **A** was adopted for use to specify an anchor: the anchor begins with the tag **<A>**, and ends with the tag **</A>**.

To specify the web page to which a given link points, the initial tag contains the keyword *HREF*, followed by an equal sign and a URL enclosed in double quotes. Figure 21.3 shows part of an HTML page that contains a link and the corresponding output that a browser produces on the user's screen.

```
On the Web, you can find some of the classic
works of English literature.  For example, the
works of
<A HREF="http://the-tech.mit.edu/Shakespeare">
                  William Shakespeare
</A>
are available.
```

a

On the Web, you can find some of the classic works of English literature. For example, the works of William Shakespeare are available.

b

Figure 21.3 (a) HTML with a link, and (b) what a user sees when a browser displays the HTML.

In the example, the phrase *William Shakespeare* is anchored to the URL:

http://the-tech.mit.edu/Shakespeare

As the figure shows, when it displays anchored text, the browser highlights the link by changing the color and underlining the anchored item. When a user clicks on a highlighted item, the browser loads the new page specified by the URL.

HTML includes mechanisms that can be used to form lists. For example, Figure 21.4 illustrates the HTML used to create an *ordered list*, which is sometimes called a *numbered list*. The pair of tags **<OL>** and **</OL>** surrounds the entire list, while the tag **<LI>** precedes each list item.

```
<OL>    <LI> breakfast
<LI>    lunch
<LI> dinner
            </OL>
```

a

1. breakfast
2. lunch
3. dinner

b

Figure 21.4 (a) HTML for an ordered list, and (b) the result when displayed by a browser.

HTML provides tags for an *unordered list*, commonly called a *bulleted list*, as Figure 21.5 illustrates. The HTML is arranged with one item per line to make it easier for a human to read.

```
<UL>
<LI> cars
<LI> trucks
<LI> buses
</UL>
```

a

- cars
- trucks
- buses

b

Figure 21.5 (a) HTML for an unordered list, and (b) the result when displayed by a browser.

21.6 An Image On A Web Page

The *IMG* tag specifies that an image should be inserted into the web page. A browser recognizes several digital image formats, including the popular *jpeg* standard created by the *Joint Photographic Experts Group*. The IMG tag uses the keyword *SRC*

to specify a URL for the image. For example, the following specifies an image on a server named *somewhere.com*:

```
<IMG SRC="www.somewhere.com/stickman.jpeg">
```

An image on a web page can appear by itself or adjacent to text. In fact, when a browser displays a web page, the browser treats an image like an oversized "word" that appears in the middle of a line of text. Figure 21.6 illustrates how a browser displays a page that specifies an image in a line of text.

```
This example shows
<BR>
a line of text <IMG SRC="stickfig.gif"> with an image in it
<BR>
and other lines around it.
```

a

This example shows

a line of text  with an image in it

and other lines around it.

b

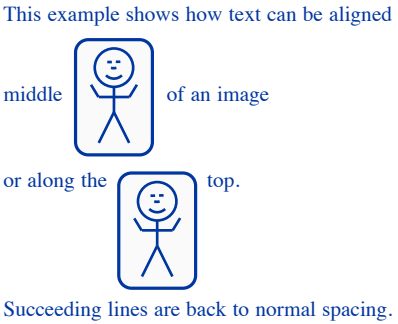
Figure 21.6 (a) HTML that includes an image and text, and (b) the result when a browser displays the page.

HTML includes many ways to control the details of how an image is aligned with surrounding text. If the *IMG* tag specifies *ALIGN=TOP*, the browser will place the image so the top aligns with the surrounding text, and *ALIGN=CENTER* will center the image. Figure 21.7 illustrates alignment specifications, and shows how a browser will display the page.

In addition to alignment, HTML includes a size specification that causes a browser to stretch or shrink an image to fit a specified size. Size specifications are especially useful when an image is much larger than a typical screen size.

```
This example shows how text can be aligned in the
<BR>
middle <IMG SRC="stickfig.gif" ALIGN=CENTER> of an image
<BR>
or along the <IMG SRC="stickfig.gif" ALIGN=TOP> top.
<BR>
Succeeding lines are back to normal spacing.
```

a



b

Figure 21.7 (a) HTML that specifies image alignment, (b) the result when a browser displays the page.

21.7 Point-And-Click Web Page Design

The examples above should make one thing clear:

> *When a browser displays a web page, the final result on the user's screen does not look like the HTML used to specify the page.*

Although browsers require web pages to use HTML, apps are available that allow a user to create a web page without knowing HTML. Known as *web authoring tools*, the apps use a point-and-click approach. For text, the user selects colors and fonts from menus. For graphics, a user can import digital photos by dragging each to a location on the page. Once the user has created a page, the authoring app automatically generates HTML for the page. The point is:

Web authoring apps are available that allow a user to compose a web page without learning HTML.

21.8 Summary

Web pages are written in a computer language known as the *HyperText Markup Language* (*HTML*). Because a browser hides the specification completely when displaying a web page, most users never encounter HTML. Instead, when a user gives a browser a URL, the browser contacts the specified server, obtains a copy of the web page, interprets the HTML, and displays the result.

HTML allows a web page to contain both text and non-text items. Each item on a page can be an anchor that corresponds to a link. Although HTML specifies a URL for each link, the user does not see the URL when the page is displayed.

EXERCISES

21.1 Some browsers allow one to see the HTML for a page (called the *page source*). See if your browser has a menu item that allows you to view the page source.

21.2 If your browser does not permit you to inspect the HTML, use a trick: save the page to a file on your computer, and then open the file with a text editing app.

21.3 You can try writing HTML yourself. Edit a text file, type in some basic HTML (such as an example from this chapter), and then point your browser at the file using the URL file://*filename*.

Chapter Contents

22 The World Wide Web: Web Pages That Change

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

22

The World Wide Web: Web Pages That Change

22.1 Introduction

The previous chapters discuss the World Wide Web and describe the internal language used for conventional web pages. This chapter describes advanced web technologies that make it possible for a web page to change and to interact with the person who views it. The chapter describes how pages can be created on demand as well as technology that allows a page to change after the page has been fetched.

22.2 Conventional Web Pages And Static Content

As the previous chapters describe, web technologies were originally designed to store information that remains stable over long periods of time in the same way a library stores books that do not change quickly. Indeed, the term *digital library* was often used to characterize the early Internet.

Web pages that seldom change are known as *static* pages. Like a page in a book, the content of a static web page is created by its author, and remains unchanged until the author revises it. A static web page behaves exactly as one would expect. Two users who each specify the URL for a static web page will each see exactly the same content. If both users have the same make and model of computer and use the same browser, the page will be formatted identically on the two screens. Finally, each hyperlink on the page will take the two users to the same destination.

Static web documents have another useful property: if an individual visits a static web page repeatedly, the content remains the same. Thus, if a user records the URL for the page one day and then uses the URL to return to the page the next day, the user will see exactly the same information. Of course, the owner of the page might choose to revise the content, in which case a user would see the revised version. Once a change has been made, however, the page will remain unchanged until the owner makes another revision. We can summarize:

If a web page is static, *all users who visit the page see exactly the same contents and the same links, until the page is revised*

22.3 How A Browser Accesses A Static Web Page

A static web page is stored in a file on the computer running the server. When a user enters the URL for the page, the user's browser contacts the server to request the page. The server reads the file that contains the page, and returns the contents to the user's browser. Figure 22.1 illustrates the steps.

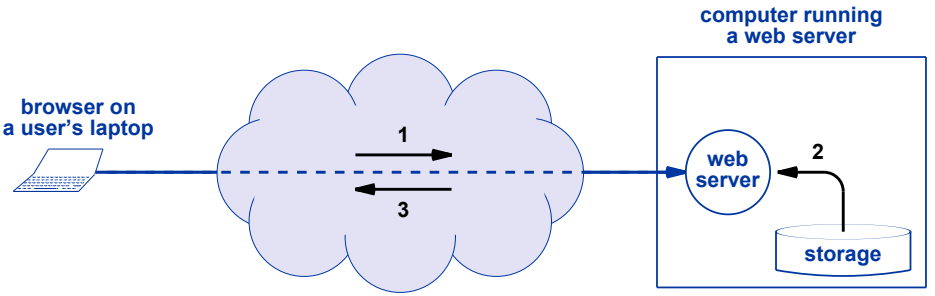


Figure 22.1 The steps taken to fetch a static web page. (1) A browser contacts a web server and requests a page, (2) the server extracts the page from its local storage, and (3) the server sends a copy to the browser.

The point is:

Because a static web page does not change, the page can be placed in a file on the computer that runs the web server.

22.4 Accessing A Page That Has Changeable Content

Static pages are useful for information that either never changes or changes infrequently. For example, the static approach works well for a page that explains the mathematical formula for the area of a circle because the formula will never change. Similarly, a page that contains photos of an event that occurred in the past is usually static.

Many web pages are not static. Instead, the contents of a page either changes each time the page is displayed or changes continuously while the user views the page. Consider, for example, a page that displays the current temperature in Chicago. The page must be updated with the latest temperature each time a user fetches a copy of the page. A key idea is that most of the page remains static, and only the temperature value needs to change.

We use the term *dynamic content* to refer to the values on the page that change. How does a web server handle dynamic content? The short answer is: the server uses a computer program. Instead of merely reading the page from a file, a web server runs a computer program that generates a page. The web server takes the output from the computer program, and sends the output back to the browser. Because the programs used with a web server are much simpler than a conventional app, we call the program a *script*. Figure 22.2 illustrates the steps taken when a user requests a page with dynamic contents.

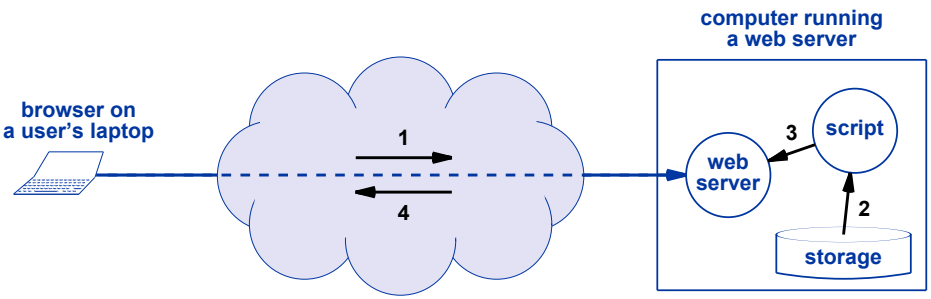


Figure 22.2 (1) A browser requests a page with dynamic content causing the web server to run a script (a computer program), (2) the script reads a page and adds new content, (3) the server receives the modified page from the script, and (4) the server returns the page to the browser.

Because it is a computer program, the script can perform any function an app can perform. The script can access a temperature sensor connected to the computer, and can even contact another computer to obtain dynamic content. The next section describes a feature of HTML that can be used with dynamic content, and a later section explains the relationship between the HTML feature, dynamic content, and advertisements.

22.5 Frames Within A Browser Window

One particular web technology is particularly useful with dynamic content. Known as *frames*, the technology allows a web page to divide the browser window into separate regions, and then fill each region separately. When frames technology is used, the web page first specifies the size and position of a set of rectangular regions (called *frames*). The web page then gives the content to be used in each frame. Figure 21.3 illustrates a web page that divides the browser window into two frames.

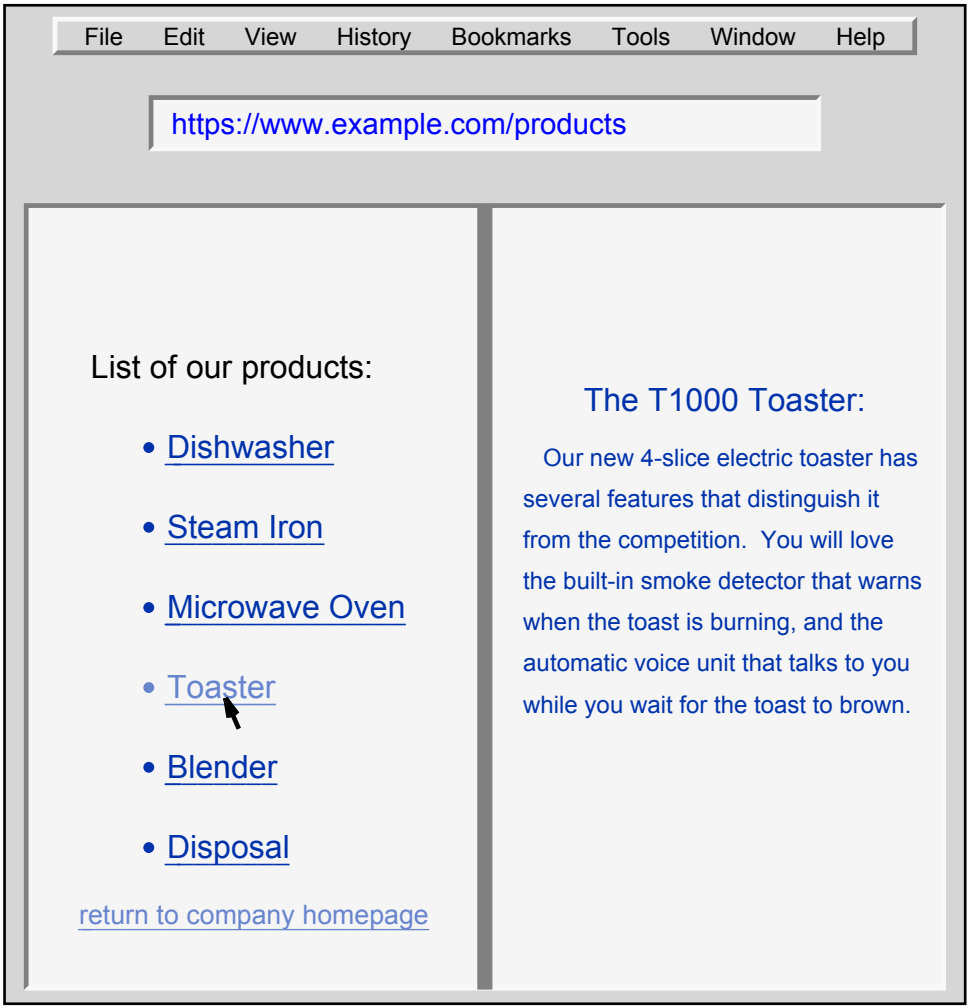


Figure 22.3 Illustration of a web page with two frames. The left frame contains a list of possibilities. When a user selects one of the possibilities, the contents are displayed in the right frame.

In the figure, the user has clicked on *Toaster* in the left frame, and the browser has loaded a description of the toaster in the right frame. If the user clicks *Blender*, the left frame will remain the same, and the right frame will display information about a blender.

22.6 Advertising And Frames

Companies that operate social media and search sites usually provide the service at no charge to users. Such companies make high profits by selling advertising. The company collects information about each user, sells the information to advertisers, and then mixes ads with web page content that the user sees. The advertising is targeted to each user's interests, which means that two users may not receive the same ads.

Companies that mix ads with content often use frame technology. Each web page is divided into frames. One frame, usually the largest, displays the content that the user requested. Additional frames, often located along the sides or bottom of the screen, contain ads.

How does a web server insert ads? The server uses a script. However, a script does not need to have a copy of all possible ads. Instead, the script only needs to have a list of URLs that specify the location of ads. The script selects an ad for each frame. and inserts the URL for the ad into the frame. When a browser displays a page that uses frames, the browser checks each frame. If a frame specifies a URL, the browser uses the URL to fetch the contents for the frame.

Why do some frames take longer to load than others? The reason should be clear. If a browser is given a page that has three ad frames with a URL in each frame, the browser must fetch and display all three ads. In most cases, the ads come from three different web servers (owned by the companies that paid for the ads). If one server is busy and another is not (or the path to one server is congested while the other is not), the time required to download one ad may be much longer than the time required to download another. The point is:

> *When frames are used to display ads, a browser must fetch an ad for each frame. The time to download one ad may differ significantly from the time to download another ad.*

22.7 Personalized Web Pages And Dynamic Content

We said that dynamic web technologies allow advertisers to change the ads that are displayed when a page is downloaded. Interestingly, the technology can also be used to produce *personalized* web page content. That is, a server can compose the content of a page to suit the customer who requests it. For example, a personalized news page can choose top stories based on what a user has selected in the past.

Personalized content has both advantages and disadvantages. The chief advantage arises from the convenience of only receiving items of interest. The disadvantage is the lack of reproducibility — two users who view the same page may not obtain the same information. When a user views news stories, for example, the user will only receive stories that are aligned with their earlier choices. Consequently, personalized content tends to support a user's opinions by avoiding alternatives. In the extreme, one can imagine personalized content in which a vendor estimates a user's income level and adjusts the price of goods and services accordingly. The point is:

> *Dynamic content technologies allow a server to personalize pages by choosing items to display and the order in which to display them. Although it helps focus on items that interest a particular user, personalization means two users who see a page may receive different content.*

22.8 Pop-Ups And Pop-Up Blockers

Another web technology is known as a *pop-up*. As the name implies, a pop-up consists of a new window that appears automatically without any user request. Although pop-ups can be used to prompt a user (e.g., to suggest a related web page), most pop-ups carry advertising. Often, a pop-up window will obscure part of the main window, which interrupts browsing, and requires a user to close the pop-up before continuing.

Because pop-up advertisements are annoying, some browsers have an option that allows a user to disable pop-ups. In addition, third-party software known as a *pop-up blocker* is also available.

The point is:

> *A web page can include* pop-ups *that cause a browser to create a separate window automatically. Because pop-ups carry advertising that disrupts a user, browsers and other software allow a user to block pop-ups.*

22.9 User Interaction With Forms

Although the scripting technology discussed above can be used to create web pages that change, such scripts run only at the server. Thus, a server script cannot interact directly with a user. To make it possible for a user to enter data, another technology was invented. Known as *forms*, the technology permits a web page to contain blank areas in which the user must enter information. After a user fills in the required information, the browser sends the information to the server when requesting another page.

The advantage of forms technology should be clear: instead of merely selecting items from a list, forms make it possible to enter data directly. When a user logs into a web site, the forms technology can be used to prompt the user for a login and password. Similarly, when a user makes a purchase, the user can enter a credit card number and a shipping address. Once the user has filled in a form, the user clicks a button to submit the information, and the browser sends the information to the server. To summarize:

Forms technology provides user interaction by allowing a web page to request information that the browser returns to the web server. Forms are useful for logging in or entering credit card numbers and shipping address information.

22.10 Shopping Carts And Cookies

One particularly useful type of user interaction is known as a *shopping cart*. The idea is straightforward: instead of forcing a shopper to purchase one item at a time, allow the shopper to collect a set of items. A key idea behind the shopping cart technology arises from long-term persistence — a shopping cart can persist across reboots of a user's device. That is, a user can collect items one day, place them in a shopping cart, then close the browser and turn off the device. When they return to the web site the next day, the shopping cart will still have the items that were selected.

Where is a shopping cart stored? On the store's web server. How can a server remember a shopping cart, know about a user, or tell whether the user has visited before? To keep track of previous visits, a web site uses a technology known as *cookies*. A cookie is a number or a short identification string that the web site assigns to a given user. When a user visits a web site, the web site sends a cookie to the user's browser when it sends a web page. The browser stores the cookie along with the web site name. For example, your browser might store:

www.example.com 315826592420

Once a cookie has been assigned, your browser supplies the cookie each time you request a page from the site. The web site uses the value to identify you and your shopping cart.

Note that a cookie itself does not contain information about you, nor does it specify the contents of your shopping cart. Instead, a cookie merely provides the server with an identifier that the server uses to look up your records. Furthermore, a cookie is only meaningful to one web site because only the web site can interpret the meaning. The server must store the list of all items in your shopping cart. The idea of using a short identifier is a powerful technique because it means that only a small amount of information needs to travel between a browser and server.

A cookie *is merely a number that a web site uses to identify you; your browser stores a cookie and uses the cookie to identify you on subsequent visits to the site.*

22.11 Should You Accept Cookies?

A server can pass a cookie to a browser at any time. Most browsers allow a user to configure their browser to reject all cookies, accept all cookies, or prompt the user for each cookie. Thus, a user must decide how to handle cookies.

The advantage of accepting cookies is that they allow servers to keep a history of your visits. The server can then use the history of your past choices along with dynamic content technologies to personalize the web pages you see. The disadvantage is that privacy is lost because the server finds out about your browsing and shopping habits. To summarize:

Accepting cookies means you allow web servers to tailor content and advertising to your tastes; rejecting cookies helps enforce anonymity.

If you choose not to accept cookies, web sites still have the option of assigning an identifier to your order. They simply display the identifier and ask you to write it down. For example the identifier may be a string of letters and digits chosen at random,† such as:

F3C8SBOJQ

When you return to the site, you will be asked to enter the identifier again. In essence, if you configure your browser to refuse cookies, you may be asked to perform the same function manually.

22.12 Animated Web Pages

Several technologies have been developed to allow a web page to include forms of *animation*. There are two basic categories:

- Play a video clip
- Run a script

The easiest animation to understand consists of using a *video clip*. Chapter 26 explains video in more detail; for now, it is sufficient to know that a video clip is a short file that contains digital video. Computer operating systems contain apps that can display a video, and a browser only needs to download the video and then use the apps to display the clip.

†Chapter 31 explains why web sites use random identifiers.

Video can be launched automatically or manually. An automatic video often uses the frames technology described above by making a video the content for one of the frames. When the page is loaded, the browser displays each frame, including the frame that contains the video. A manual approach means that a link on the web page points to a video clip. When the user clicks on the link, the browser downloads and plays the video. The link could be anchored to text, such as:

Click here to see a video.

When a link occurs in text, it means the browser will replace the current web page with the video or will display the video in a separate window. A web page can also allocate a frame for a video, and then display a link in the frame. When the user clicks the link, the video displays in the same frame as the link. One clever scheme fills the frame with an image of a "play" button, and makes the image a link. The image is only the picture of a button, and is not tied directly to an app that displays videos. However, when the user clicks on the image that appears to be a play button, the browser downloads and displays the video, giving the illusion that the video was already loaded and ready to play. Figure 22.4 illustrates how a web page might display an image that links to a video.

Figure 22.4 Illustration of a frame that contains an image that appears to be a *play* button. The image is a link to a video.

22.13 Animation With A Browser Script

When a web page contains a video clip, the result is a high-quality picture with smooth motion. A lower-overhead alternative exists for cases where only basic animation is needed. For example, some web pages repeatedly reverse the background and foreground colors in an ad to make the ad appear to flash. Although the effect can be achieved with a video clip, the same effect can be achieved by running a *script* (i.e., a computer program) that swaps between two versions of the page. Figure 22.5 illustrates the idea.

Figure 22.5 Illustration of two web pages with the background and foreground colors reversed. Repeatedly alternating between the two gives the appearance of flashing.

The script does not need to use two images. Instead, the script uses a small amount of HTML text to tell the browser to paint the frame with one set of colors and a small amount of HTML for the reverse colors. More important, because HTML is downloaded with the rest of page, the display can begin immediately, without waiting to download separate images or a video clip.

22.14 Java, JavaScript, And HTML5

Several technologies have been designed that a browser can use to animate content. One technology uses a programming language named *Java*. Java allows a programmer to create complex animations that can interact with the user and use the display screen in unexpected ways. Java uses the term *applet* for web page programs to imply that they are smaller and less complex than other apps. The terminology has become widespread, and other technologies have adopted the term or chosen to use a minor variation.

Java became popular for five reasons. First, Java was invented before other technologies. Second, Java is easy for professional programmers to use. Third, Java includes mechanisms to handle common tasks, such as tracking a mouse. Fourth, Java handles direct user interaction better than forms, and can control the multiple areas of the screen better than frames. Fifth, Java provides high-quality animations.

Although Java is among the most powerful active document technologies, it is also among the most complex. Creating a Java program that can control the display requires extensive background and training; the language is designed for professional programmers.

Several technologies have been designed as alternatives to Java. One of the most popular alternatives incorporates a few of the basic features of Java, while omitting many of the more complex features. Known as *JavaScript*, the language can be embed-

ded in a standard HTML file. When it encounters a JavaScript section in an HTML document, a browser performs the specified computation, and then displays the results. Thus, although it is not as complex as Java, JavaScript offers much of the same functionality to users. After years of experience with browser scripting systems, a new version of HTML has been designed to incorporate additional functionality, including support for animated web pages. The new version is named *HTML5*, short for HTML version 5). The point is:

> *Technologies, such as Java, JavaScript, and HTML5, are available that allow web programmers to create animated web pages.*

22.15 Summary

A static web page consists of content that does not change. Many pages include dynamic content, which means the server uses a script to form a page whenever a browser makes a request. The frames technology allows a web page to be divided into regions with separate content in each region; advertisers use frame technology to place the main content in one frame and ads in other frames.

Many web pages include ways that a user can interact by sending information back to a web site (e.g., by filling in a form to supply a login ID and password). One popular form of interaction uses a shopping cart to allow a user to select a set of items for purchase. To maintain information across reboots of a browser or a device, web sites use cookies. A cookie is an identifier that the web site uses to associate information saved on one visit with a successive visit.

A variety of technologies, including Java and JavaScript, have been invented to allow a web page to contain animations. In each case, a browser runs a computer program called an *applet* or a *script* to perform animations. Experience with web technologies has led to a new version of HTML named *HTML5*, which is designed to subsume earlier server technologies.

EXERCISES

22.1 Try visiting a web site such as Yahoo or Facebook that displays ads along with content and see if you can detect some ads loading more slowly than others.

22.2 In the previous question, what technology is being used to display ads?

22.3 Joe visits a news web page that has a story about a fire and a separate frame that says "click here for scenes from the fire." After Joe clicks to see the scenes, many seconds pass before a video starts playing. Explain why the video did not start immediately by describing what happened during the pause.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

23

Social Networking And Personal Publishing

23.1 Introduction

Previous chapters describe the World Wide Web and associated technologies that allow access to information. This chapter focuses on social networking services that permit users to share information with others. The chapter explains that although social networking services use the same client-server paradigm as other services, they have changed the way information is distributed.

23.2 The Publish-Subscribe Paradigm Changes

Until the twentieth century, information dissemination followed a *publish-subscribe* paradigm in which a small set of *publishers* selected, reviewed, and edited information, and then published the results. A large set of *subscribers* then paid for access to the published results. For example, a local newspaper selected stories of interest in the town and published them each day. Magazines were published monthly, and book publishers focused on information that lasted years. Because publication and dissemination of information was expensive, an average person could not afford to become a publisher.

By lowering the cost to disseminate information, the Internet caused a shift away from the publish-subscribe paradigm. Instead of publishing on paper and then shipping physical copies to subscribers, the Internet allows information to be stored on a server where subscribers can access it quickly. The cost of computer hardware became so low

that even a small business or an individual could afford the equipment needed to run a server that connected to the Internet. Suddenly, it seemed that anyone could become a publisher.

23.3 The Rise Of Internet Publishing Services

Despite the low cost of equipment and software, one other factor prevents most individuals from becoming Internet publishers: expertise. Even after the World Wide Web was invented, publishing required a staff of Information Technology (IT) professionals that would install, configure, and operate a web server. Consequently, only large organizations had the necessary expertise.

A final shift has occurred that changed the publish-subscribe paradigm: a set of companies emerged that offer *hosting services*. That is, the companies sell the service of running servers for customers. The companies, which are known as *hosting companies*, each own a set of computers, and each handle the details of installing and operating the computers along with the hardware and software that connects the computers to the Internet. A customer of a hosting company does not need IT expertise to use a hosting service — a customer only needs to generate content and then pay a hosting company to make it available on the Internet. Because many customers share the cost of an IT staff, the cost to each customer is low.

23.4 Discussion Forums And Bulletin Boards

Once hosting companies emerged, the question arose, "What Internet services can be created?" An early answer came from a service that was invented while the Internet was being created: discussion forums. A group runs a server that allows anyone to submit a short message on a specific topic, and allows everyone to read the messages. Forums are called *electronic bulletin boards*; one early forum technology used the names *newsgroups* and *network news* (even though the forums contained opinions rather than traditional news).

23.5 Moderated Discussions And Editorial Control

In the beginning, users lauded discussion forums for giving everyone a voice. As the Internet grew, it quickly became apparent that allowing everyone to post opinions on an open forum was not productive. Arguments broke out between users who disagreed, and discussions sometimes deteriorated into name calling and personal attacks. To prevent such behavior, many forums moved back toward a publish-subscribe model by reintroducing a key idea: editorial control. All submissions to a forum are sent to an individual who is known as the forum's *moderator*. The moderator reviews

each submission, and rejects submissions that are off-topic, personal attacks, or advertisements of goods and services.

Interestingly, after experimenting with an open approach, some discussion forums moved back to the publish-subscribe model of requiring users to pay a subscription fee — only users who pay the fee can contribute messages and read the contributions of others.

23.6 Essays And Personal Opinions (Blogs)

Although a discussion forum allows participants to submit opinions, a participant's submission is intermingled with all other submissions. As an alternative, a *blog* consists of an extended opinion written by a single individual. The author of a blog can choose a topic, a style, and a length.

Blogs introduce an interesting reversal of the publish-subscribe paradigm because the owner of the blog pays a company to host a server. Typically, the owner of a blog pays for a web server, and makes the blog available on the Web. Subscribers can access the blog at no charge, which means the underlying paradigm can be described as *pay to publish*. The point is:

> *When an author pays a company to host a web site for their blog, the result is a reversal of the traditional publish-subscribe paradigm because the publisher bears the entire cost and subscribers access the information for free.*

23.7 Cooperative Publishing (Wikis)

The Internet introduced another twist on the traditional publish-subscribe paradigm by allowing a set of individuals to cooperate in authoring and editing information. The technology has become known as a *wiki*. The idea behind a wiki is straightforward: a web site can be configured to allow users to change the information on each page. A group of users then accesses the site. The users cooperatively author and edit information.

A wiki site can be open to the public or restricted to a specific set of users who must log in to obtain access. In fact, permissions to read and change information can be configured separately for each page of a wiki site. Thus, it is possible to grant everyone permission to read a given page, while restricting changes to a few specific users.

The software needed to create a wiki is freely available, and many organizations make it possible for members to edit one or more pages of the organization's web site. *Wikipedia* provides the best-known example of a wiki that is open to the public. The

project started with the assertion that an encyclopedia created and maintained by millions of Internet users will be more extensive and more up-to-date than a traditional encyclopedia. Overall, Wikipedia has become a valued resource. However, the absence of editorial control means users with no expertise on a subject can replace facts with their opinions, and Wikipedia has experienced the problem.

23.8 Personal Web Pages And Social Networking Sites

Sites such as *Facebook*, *Snapchat*, *Instagram*, and *YouTube* extend publishing by allowing each user to have an online presence. The information published by a user constitutes a combination of a personal diary and scrapbook that becomes accessible by friends, family, and others. We use the term *social networking* to characterize the process and the sites.

Social networking sites gain and lose popularity. An early site named *Myspace*, which was once extremely popular, became passé. Teenagers tend to prefer sites other than the sites their parents use.

Social networking changes the traditional publish-subscribe paradigm by making it possible to publish personal information that is not of interest to a large audience. Social networking sites derive revenue from advertising to users. Thus, social networking is a pay-to-publish paradigm in which a user agrees to give up privacy and read ads in exchange for having a company host a page that contains the user's personal information. The point is:

> *Although users think of it as free, social networking follows the pay-to-publish paradigm in which a user agrees to relinquish privacy and read ads in exchange for having their personal information hosted on a web site.*

23.9 Summary

The Internet triggered a shift away from the traditional publish-subscribe paradigm that had been used to disseminate information. A variety of new publication mechanisms have been invented, including moderated and unmoderated discussion forums (also called electronic bulletin boards or newsgroups), blogs, cooperative publishing systems (wikis), and social networking sites.

The driving force for much of the change arises from the low cost of equipment needed to supply information on the Internet. Some mechanisms employ a pay-to-publish paradigm in which a user pays the entire cost of making their information available. Social networking uses the interesting twist of a pay-to-publish model in which a user gives up privacy and agrees to read ads in exchange for having their personal information hosted on a web page.

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

24

The Internet Of Things (IoT)

24.1 Introduction

Previous chapters describe Internet services that humans use to access information or communicate with others. This chapter describes a new use of the Internet that began in the twenty-first century that involves devices connected to the Internet that do not have a human operator. The chapter explains how communication takes place, and provides background for cloud computing (Chapter 29).

24.2 Connected Devices Without Human Operators

Early Internet services focused on allowing a human to access information or allowing two humans to communicate. In each case, a human used a computing system to run an application program that communicated over the Internet. Furthermore, most of the information available was created by humans. For example, each web page required a human to create the page.

Gradually a change occurred in which a new type of device was connected directly to the Internet. Instead of requiring a human operator, the devices operate independently and interact with their surroundings. The devices interact in two ways:

- **Sensors**: devices that measure or sense their surroundings
- **Actuators**: devices that change or control their surroundings

24.3 Sensors

Some of the earliest specialized systems use sensors. As an example, consider a web site that reports the outside temperature. To handle a large volume of requests, such web sites cannot depend on a human to look at a thermometer and enter the current temperature each time a request arrives. Instead, the site places a special stand-alone sensing system outdoors. The system contains a miniature computer, a temperature sensor, a network interface (often a wireless network), and special software. The software repeatedly reads the temperature sensor. When the temperature changes, the system communicates over the network to save the new temperature in a file on the computer that runs the web server. When someone requests the temperature, the web server reads the value from a file and returns the answer. Figure 24.1 illustrates the sequence of steps.

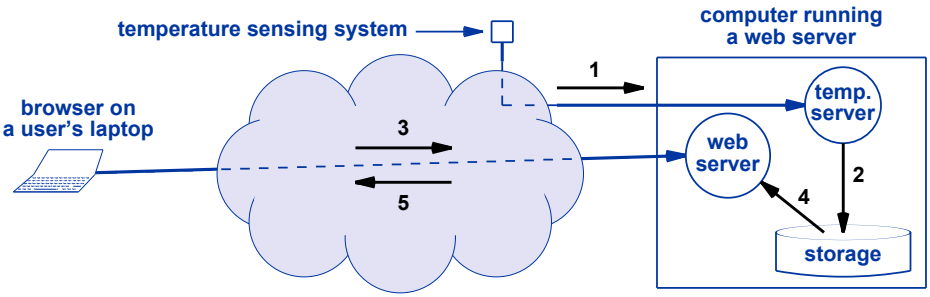


Figure 24.1 Steps used with a temperature sensing system: 1) the system reads the temperature and contacts a server, 2) the server saves the value on storage, 3) a user requests the temperature, 4) the web server fetches the value from storage, and 5) the value is returned to the user.

We say that a sensing system is *autonomous* if it runs without any human intervention. Once it is powered on, an autonomous temperature system starts automatically, forms a connection to a server, and continuously uses the attached thermometer to send updated temperature readings to the server. Thus, the current temperature has already been recorded before a user makes a request.

24.4 Actuators

A device that controls a mechanical or electrical mechanism is known as an *actuator*. For example, when a fire occurs in a large building, actuators are used to close a set of *fire doors* to keep the fire from spreading. Similarly, when a user runs an app that starts their car remotely, a system in the car uses actuators to perform the required actions.

Often, actuators are combined with sensors. For example, consider a remote control surveillance camera that shows the user video and allows the user to turn the camera and control the zoom. Primarily, the camera acts as a sensor that takes video of its environment and delivers the video to the user. However, actuators are used to turn the camera and change the zoom.

24.5 Embedded Computer Systems

At one time, a television consisted of an analog device that received and displayed a broadcast signal. Now, televisions are "smart," which means they can access popular Internet streaming sites. Part of the installation process involves connecting a television to the Internet.

How can a television send datagrams across the Internet? The answer is straightforward: a smart television has a computer built into it. We use the term *embedded system* to capture the idea. Like a laptop or a smart phone, an embedded system has a processor, memory, storage system, and software that allows it to communicate over the Internet. When a user selects an Internet streaming service, the embedded system runs client software that accesses the appropriate server, allows the user to select a video, and then streams the video over the Internet.

Physically, an embedded system can be incredibly small. For example, hearing aids are available that are so tiny they fit into a person's ear. Despite the small size, the devices contain an embedded system that can receive a Wi-Fi audio broadcast at a concert. In terms of computational power, an embedded system can be impressive — a smart phone contains a more powerful processor and larger memory than scientific workstations did in the 1990s.

Embedded systems that can communicate over the Internet are used in many devices, including:

- Cash registers and point-of-sale terminals
- Printers and other office equipment
- Medical, health, and fitness monitoring devices
- Vending machines
- Security systems
- Video games and entertainment systems
- Vehicles
- ATM machines
- Kitchen appliances

24.6 The Internet Of Things

Industry has adopted the term *Internet of Things*† (*IoT*) to refer to devices that use embedded systems to communicate over the Internet. In the twentieth century, the Internet connected conventional desktop and laptop computers. Now, the number of IoT

†The term was created by industry to spark interest among customers; a more accurate term might be *Things on the Internet*, which makes it clear that the things use the existing Internet.

devices is growing rapidly, and industry pundits predict that in the future, far more IoT devices will be using the Internet than smart phones, laptops, and desktops. The point is:

> *Although humans tend to think of the Internet as connecting smart phones, laptops, and desktops, small IoT devices have begun to dominate.*

What are all of the IoT devices doing, and how do they communicate? The next sections survey a few of the many IoT applications, and describe the communication paradigm IoT devices use.

24.7 Gadgets And Wireless Network Connections

Consumers know about IoT devices because vendors advertise such devices in the form of gadgets. For example, a pet food dispensing system allows a user to feed their pet from a remote location. A user fills the dispenser with pet food, and leaves the dispenser in their home, connected to the Internet. The user can run an app on their smart phone to control when the dispenser releases food and how much to release.

Most IoT gadgets use wireless network connections. Smaller, battery-powered IoT devices often use Bluetooth, and rely on an app running on an intermediate system to forward data to its final destination. For example, a wearable medical monitor might use Bluetooth to reach a user's smart phone. The phone runs an app that receives data over the Bluetooth connection, and forwards the data to a hospital database system for the doctor to review.

Larger IoT devices typically use Wi-Fi, and usually connect to a destination without an intermediary. For example, the pet food dispenser described above would probably use Wi-Fi. When a user acquires such a device, the user must configure the network connection by specifying the SSID of a Wi-Fi network.

24.8 Centralized And Mesh IoT Networks In A Home

Many IoT devices use wireless network connections, and the IoT industry has devised two ways to connect devices to the Internet: *centralized* and *mesh*. A centralized system consists of a *wireless router*† plus a set of wireless IoT devices. The wireless router connects to the Internet, typically through a wired connection, and uses a wireless network technology to connect to IoT devices. Figure 24.2 illustrates the idea.

†Some IoT vendors prefer the term *border router* to emphasize that the router forms a border between the Internet and the IoT devices.

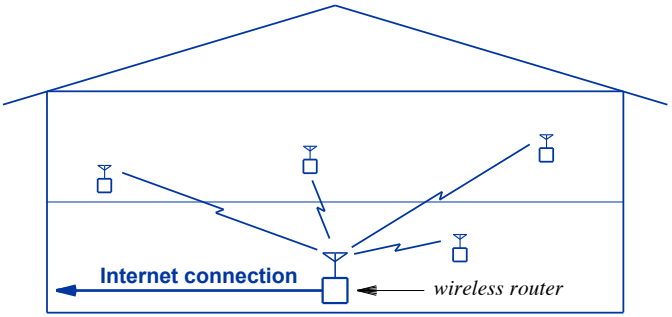


Figure 24.2 Illustration of a home using a centralized wireless router to connect a set of wireless IoT devices.

In a large home, the wireless signal from small battery-powered IoT devices may be too weak to reach a centralized wireless router. In such cases, the home owner can add a *repeater* that acts like a second wireless router. Because it plugs into a power source, a repeater has a stronger signal than a battery-powered IoT device. Figure 24.3 illustrates the use of a repeater.

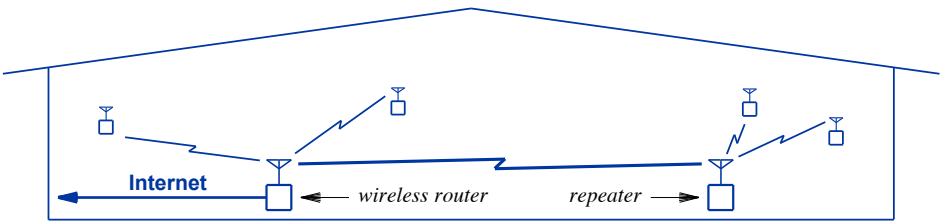


Figure 24.3 Illustration of a large home using a repeater plus a centralized wireless router to span a large distance.

24.9 A Wireless IoT Mesh In A Home

The IoT industry has devised an alternative approach to accommodating a large distance between a wireless router and IoT devices. Known as a *wireless mesh*, the approach uses IoT devices to relay packets on behalf of IoT devices that are farther away.

When it is powered on, a device using the mesh approach must find a path to the Internet. If it is close to a wireless router, the new device communicates directly with the wireless router. If not, the device uses its wireless radio to broadcast a request seeking *neighbors* (i.e., devices that are close by). Each neighbor that receives the request responds, and the new device chooses one of the neighbors to act as an intermediary.

Each time it has a packet to send over the Internet, the new device sends the packet to its chosen neighbor, which forwards the packet on toward the wireless router. Similarly, when a packet comes from the Internet to the device, the packet travels to the wireless router, which forwards the packet across the mesh to its destination. Figure 24.4 illustrates a mesh network.

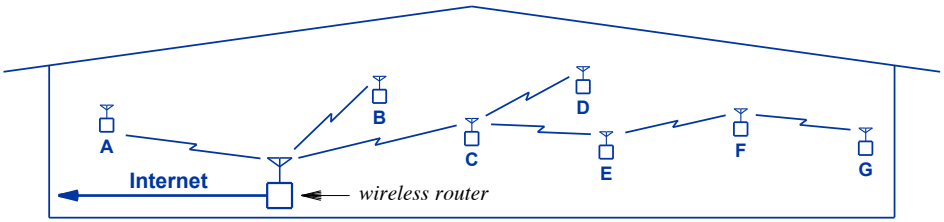


Figure 24.4 Illustration of a mesh network in which each IoT device agrees to forward packets between the wireless router and other devices.

In the figure, only devices labeled *A*, *B*, and *C* can reach the wireless router directly; the remaining devices must use the mesh approach. Device *E* can reach *C*, *F* can reach *E*, and *G* can reach *F*. When *G* sends a packet out to the Internet, the packet travels from *G* to *F*, from *F* to *E*, and from *E* to *C*. Finally, *C* sends the packet to the wireless router, which forwards the packet to the Internet.

The mesh approach has the advantage of being automatic — IoT devices form a mesh without any help from a user. If a user moves a device from one room of a house to another, the device will seek new neighbors, and try to re-establish a path to the Internet. However, the mesh approach has the disadvantage of relying on other devices for connectivity. In Figure 24.4, for example, if a user powers down device *C*, devices *D*, *E*, *F*, and *G* will be cut off from the Internet. We can summarize:

> *In a mesh network, each device agrees to forward packets on behalf of other devices. Although it works automatically without requiring a repeater, the mesh approach has the disadvantage that removing a device can leave other devices cut off from the Internet.*

24.10 Smart Homes, Buildings, And Factories

When they think of IoT devices, most users focus on devices that they can see and touch. One of the main uses of IoT, however, involves the systems built into the infrastructure of homes, buildings, and factories. Industry uses the adjective *smart* to characterize a building in which the facilities can be monitored and controlled over the Internet.

Examples of facilities that IoT devices can control in a smart building include:

- Lights, including both groups of lights as well as individual lights
- Heating Ventilation and Air Conditioning (HVAC) systems
- Security systems, including locks, sensors, and security cameras
- Elevators and stairway access doors

Versions of smart building facilities are available for home use. For example, door locks and garage door openers are available that can be checked or operated from a smart phone, as are security cameras and HVAC systems.

24.11 Civil And Power Infrastructure: Bridges And Grids

Most civil infrastructure (e.g., overpasses, bridges, tunnels, and other structures) must be inspected regularly to detect cracks and wear. The availability of low-cost, low-power IoT devices has enabled civil engineers to attach sensors that measure the infrastructure and report continuously. In many cases, the devices take so little power that a device can operate for many years before a battery must be replaced.

The term *smart grid* refers to an electrical power grid equipped with devices that monitor and control the distribution of power. One aspect of a smart grid involves redistribution of power to handle unexpected outages. Another aspect of the smart grid focuses on control of individual appliances. For example, if all appliances were equipped with smart grid controls, a power company could restrict use of some appliances during peak hours (e.g., restrict kitchen dishwashers to late evenings).

24.12 Summary

We use the term Internet of Things (IoT) to refer to sensors and actuators that can be accessed and controlled over the Internet. In addition to wearable medical and health monitoring systems, IoT encompasses point-of-sale terminals, such as cash registers, ATMs, and vending machines; household appliances and gadgets; building infrastructure components, such as lighting, security, and HVAC systems; and civil infrastructure monitoring, such as bridges and tunnels.

Many IoT devices use wireless networking, especially consumer products designed for use in a home. The IoT industry has taken two approaches to wireless networking. In one approach, a centralized wireless router (possibly with a repeater) connects all IoT devices in the home to the Internet. In the other approach, a centralized wireless router connects to nearby IoT devices, and remaining devices form a mesh in which a device agrees to forward packets on behalf of devices that are farther away from the wireless router.

EXERCISES

24.1 Make a list of IoT devices you own or use. Don't forget infrastructure devices, such as smart thermostats controllable over the Internet.

24.2 Search the Web and find examples of IoT devices that you think are unusual, exciting, or silly.

24.3 Consider a smart home that uses IoT devices to access and control all lights, appliances, and the heating system. Give examples of situations where access and control would be helpful.

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

25

Internet Search (Search Engines)

25.1 Introduction

Earlier chapters describe a variety of Internet services, including the World Wide Web. We learned that a user can enter an initial URL and then follow links to see other web documents.

This chapter considers one of the most widely used Internet services, Internet search. The chapter considers the broad question of how a search company catalogs the information on millions of web pages that change constantly, and how such a site can answer queries quickly.

25.2 Databases And Structured Information

A *database* provides an example of one way that information can be organized to make searching easy. All information in a database is uniform. A company might create a database to store information about its employees. Each entry in an employee database, called a *record*, corresponds to a single employee, and contains a fixed set of items about the employee, which are called *fields*. For example, an entry in an employee database might contain five fields:

- Employee's name
- A picture of the employee
- Employee's badge number
- Department in which the employee works
- Date the employee joined the company

The information in a database is *structured* because each record in the database has the same fields, and each field has a specific meaning; the meaning is set when the database is created. Structure makes it easy to ask precise questions about the records in a database. For example, when database software is used to search the employee database, it allows one to ask for a list of all employees who joined the company in the last two years or a list of all employees who work in the *Finance* department.

25.3 Classification Of Information

To create a database, one first defines a structure for the database by specifying all the fields that will be needed. Once the structure has been defined, records can be inserted that have values filled in for each field. Although structured information is important, the approach does not apply to Internet searches because web pages do not all follow the same structure.

How can information be searched? For centuries, librarians maintained libraries of books, and helped users locate information. The question arose:

Can information in a library be organized to make searching faster and easier?

In 1876, Melvil Dewey proposed a *classification* system as an alternative to relying on human librarians.† Dewey's classification, which has been adopted widely, organizes books by topic, using three digits for major categories and fractional numbers for subcategories. Figure 25.1 lists a few of Dewey's categories.

000	**General works**	**500**	**Pure science**
100	**Philosophy and psychology**	**600**	**Technology**
200	**Religion**	**700**	**Arts & recreation**
300	**Social sciences**	**800**	**Literature**
400	**Language**	**900**	**History & geography**

Figure 25.1 A few of the major categories in the Dewey Decimal system.

The task of classifying all information on the Internet is overwhelming for two reasons. First, the Internet contains vast amounts of information. Second, new types of information appear continuously. Consider, for example, the technologies that appeared after Dewey defined his classification. Do we need subcategories for automobiles, televisions, computers, and social networks? Should the Internet be a new top-level category?

Even if someone devised a classification scheme for all information, searching web pages would not be efficient. To understand why, consider an example. Suppose the classification has a category for schools, and subcategories for elementary schools, high schools, and universities. With such a classification, finding all elementary schools would be easy. If someone wanted to search for all schools in New Jersey, however,

†The formal term *ontology* is used to describe a comprehensive classification scheme.

they would first need to look through elementary schools to find the elementary schools in New Jersey. They would then need to look through high schools to find high schools in New Jersey. Finally, they would need to look through universities to select universities in New Jersey. The point is:

> *Although classifying information appears to aid searching, a given classification scheme makes some searches easier and some more difficult.*

25.4 Searching Unstructured Web Pages

Although database searching works well for a set of records that each have exactly the same fields as other records, the approach does not handle arbitrary items. How can web pages be searched without requiring each page to be classified? Three approaches have been used:

- Text matching
- Pattern matching
- Keyword frequency

Text matching. The simplest way to perform a search consists of matching whatever string of text the user enters as a search request. For example, if a user searches for *apple*, the user will be presented with a list of pages that contain the five letters "apple" in that order. Unfortunately, text matching is inaccurate because a search for *apple* will include pages that contain *applejack*, *crabapple*, *dapple*, *grapple*, and *pineapple*, which may not be what the user intended. Similarly, a search for *pear* will include pages that contain *Shakespeare*, *appear*, *appearance*, *pearl*, *spear*, and *spearmint*.

Another disadvantage of text matching arises from the lack of semantics — although it works with individual letters in a word, a program that uses text matching does not understand the meaning of words or phrases. For example, if a user enters the topic *automobile*, a text matching system will not find pages that contain synonyms or related terms, such as *car* or *vehicle*. Furthermore, if a user misspells a term (e.g., *auotmobile*), a text matching program may not find any matches. The lack of semantics becomes especially pertinent when the meaning depends on an entire sentence. As an example, consider the following:

> *This sentence does not contain any information about biology, money, or foods like butter and milk, and certainly is not about automobile pictures, airline fares, lawyer jokes, opera singers, or library books.*

Such statements confuse text matching systems because the presence of *not* reverses the meaning. Therefore, a text matching system might suggest the page as an answer to a request for information about *money*, *automobile*, *jokes*, *opera*, or *law*.

Pattern matching. Early Internet search services expanded the idea of text matching to provide more complex patterns, thereby, allowing the user to be more specific. For example, a pattern matching system allows a user to exclude pages (e.g., to specify that they are interested in pages containing *pear*, but not *Shakespeare*). Pattern matching can also make inclusion more specific. For example, a user interested in apple pie recipes can request pages that contain the words *apple*, *pie*, and *recipe* in any order.

Keyword frequency. Although they give a user more control than basic text matching, pattern matching systems still rely on pages to contain text that matches specific strings. Clearly, a better search mechanism is needed. The field of *information retrieval* offers a way to solve the problem: instead of simply matching strings of characters or patterns, examine each document to find a set of terms that identify the purpose of the document. Although the analysis uses complex algorithms, the underlying idea is straightforward: compare how often each term appears in a given document with how often the term appears in all other documents. The method is based on the observation that although a term might appear in many pages, it will be much more common in pages that focus on the topic the term describes. Thus, instead of returning all pages where a term appears, a keyword search will only return pages where the specified term has significance.

25.5 A Demonstration Of Keyword Search

To see how well keyword searches work, consider an extreme example. Try searching for the word *and*. Because it is a common conjunction in English, *and* appears many times on millions of web pages. The text matching strategy of returning all pages that contain the word *and* will fail completely.

By using a keyword search (and using a few other analysis techniques), a search algorithm can eliminate most pages and produce a list where *and* is significant. Consequently, a keyword search will only include pages that give a dictionary definition of *and*, provide a description of the *and* operation in Boolean algebra (and its implementation in computer circuits called *and gates*), include a page that describes the *Academy of Nutrition and Dietetics* (*AND*), and list a few other pages where *and* has some significance. The important idea is that a keyword search will *not* return millions of pages on which *and* occurs many times in normal English prose.

25.6 Indexing: How An Internet Search Engine Operates

We use the term *search engine* to refer to a site that provides an Internet search service. A search engine faces a challenge because the Internet changes constantly as new pages appear, old pages disappear, and the information on individual pages is updated. To provide up-to-date answers, a search engine must somehow incorporate all the changes in its answers. However, searching through all the web pages on the Inter-

net takes an incredibly long time (i.e., many hours) — much longer than a user is willing to wait for an answer.

Fortunately, a search engine produces an answer to a user's search request quickly. How can a search engine answer requests without delay if searching the entire Internet takes a long time? The answer lies in gathering the necessary information before a user submits a search request. Then, when a request arrives, the search engine can use information that is already available.

We use the term *indexing* to refer to the process of analyzing a web page and extracting terms that help distinguish the page from other pages. Because web pages change, a search engine must perform indexing repeatedly. To perform indexing, a search engine runs a computer program known as a *web crawler*. Some professionals use the term *spider*,† and say that the program operates as a *bot* because it runs without human interaction. A web crawler systematically accesses web pages, acting like a browser and downloading each page. Instead of displaying a page for a user, a crawler examines the content, indexes the page to extract keywords, and then moves on to the next page.

The web crawler places the indexing information on storage at the search engine site. Later, when a user performs a search, a server at the search engine site accesses the stored indexing information, and forms a list of pages that contain keywords satisfying the user's query. Figure 25.2 illustrates the process.

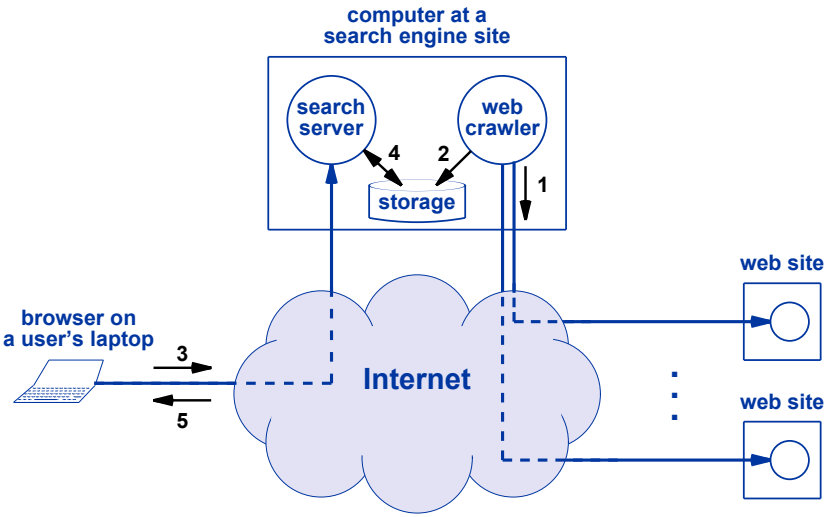


Figure 25.2 Illustration of a search engine where (1) a web crawler searches web pages on all web sites, (2) the crawler places indexing information on storage, (3) a user sends a search query, (4) the search engine server consults the indexing information, and (5) the server returns an answer.

†The term *spider* was chosen to be humorous — if something crawls around a web, it must be a spider.

The important idea is:

> *Because it gathers and stores indexing information before a user sends a search query, a search engine does not need to search web sites when responding to a user's request.*

25.7 Personalized Search Results

When a user requests a search, the search results do not come from a static document. Instead, the search engine creates the result dynamically, formats the results into a list on a web page, and returns the page to the user. Each item on the results page is a link the user can follow. The point is that the server must compute an answer for each request.

Dynamic generation of results offers an advantage: the results a user receives can be tailored to the user. We say that the results are *personalized*. Major search engines gather as much information about an individual as possible, and then use the information to control search results. For example, search engines use cookies and other technologies to watch all the searches a user performs as well as the links a user clicks. The engine uses the data to estimate characteristics, such as the user's age, sex, educational level, and interests. The characteristics are then used when a user makes a search request: the search engine selects a set of web pages that match the user's request, and orders the results according to the user's interests. The ultimate goal is a search algorithm that can match responses to the individual who made the request. For example, consider a user who enters the search term *jaguar*. If the user has previously viewed information about automobiles, the search engine might place pages about Jaguar automobiles higher on the list of results than other sites. In contrast, if the user has previously viewed information about big cats, the search engine might place pages about animals higher on the list.

After it delivers search results to a user, how can the search service know which links, if any, a user clicks? One way involves a small deception — when returning search results, the search engine does not include the correct URL for each link. Instead, the URLs in the search list each point to the search site itself. When the user clicks on a link, the user's browser sends the request to a special server on the search site. The server records the user's selection, and then forwards the request to the actual web site automatically. The user receives the page as expected. Thus, the search site merely acts as a middle man by gathering information about the user's selection before forwarding the request.

To summarize:

> *Search engines keep information about users and the links they click. A search engine uses the information to select and order search results. As a consequence, the personalized search results that a user obtains may differ from the results another user obtains for the same request.*

25.8 Indexing The Entire Web

How can a web crawler find every web site? It cannot. New sites appear constantly, and new pages appear on existing sites. However, crawlers make a valiant attempt. They use the domain name system to find all computers that have been given names, and check each to see if the computer runs a web server. As it looks through a web page, a crawler records all the URLs on the page, and adds them to the list of pages to search.

Search engines use another important optimization that helps limit the amount of information they need to store: instead of indexing all the information on a web page, a spider can use the HTML tags to identify important items. For example, HTML uses the tags *<title>* and *</title>* to identify the title of a page, making it easy to extract keywords from the title. Some search services also recognize the *META* tag, which allows the page to specify keywords that do not appear when a user views the page in a browser. For example, a page about travel might choose to add the following *META* tag to a page:

```
<META NAME="holiday travel" CONTENT="travel, airplanes,
tickets, hotels, motels, rental cars, trains, restaurants,
sightseeing, theaters, shows">
```

In addition to keywords in the title and *META* tags, some search services extract an initial set of words on each page and use the words to compile search keys because the beginning of a page tends to contain keywords that identify the purpose.

25.9 Advertising Pays For Searching

How does a search service generate revenue? Most do not charge a user who requests a search. Instead, the service generates revenue from advertising. When a user requests a topic, the search engine uses information about the user and the topic to choose advertisements to return with the search results. Search engines often assume a user who asks about a topic is interested in making a purchase. As a result, ads often present products and services related to a search. The point is:

> *Internet search appears to be free because a search engine does not charge users for each search. However, the search company sells information about the user to advertisers, and returns ads related to the user's interests or the search results.*

25.10 Summary

Internet search is among the most popular services available. To achieve accurate results, a search engine uses indexing in which a set of keywords are extracted from a page that distinguish the page from others.

To return results without delay, a search engine uses a web crawler that runs before a user requests a search. The crawler indexes as many pages as possible, and stores the information; a search server uses the stored information to find a list of pages pertinent to a given request.

Search engines use a variety of techniques to gather information about users. The engine uses the information in two ways: to provide personalized answers to search queries and to sell ads that are delivered along with search results.

EXERCISES

25.1 Try searching for common words, such as articles (e.g., *a*, *the*), conjunctions (e.g., *and*, *but*, *or*), and prepositions (e.g., *with*, *into*, *from*). Does the search site filter out most of the useless pages?

25.2 Search sites often suggest ways to complete a search. Try searching for *four seasons*. Keep in mind that Vivaldi wrote a classical piece with that name, and a 1960s singing group had the name. What topic does the search site suggest?

25.3 A friend tells you that they must be famous because a search of their name includes a URL for their Facebook page. However, when you perform a search, your friend's page does not appear. Explain why.

Chapter Contents

26 Voice And Video Communication (VoIP)

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

26

Voice And Video Communication (VoIP)

26.1 Introduction

Previous chapters describe a variety of Internet services that can be used to access and exchange data, such as web pages and email messages. This chapter begins with a discussion of services that allow users to send and receive audio and video. The chapter describes video and audio clips as well as live streams. The chapter presents the important concept of buffering, and explains why a packet switching technology uses delayed playback.

26.2 Real-Time Information

We use the term *real-time* to characterize any information presented to the user in exactly the same time sequence that the information was recorded. Audio and video constitute the most common forms of real-time information. To understand the concept, consider an audio recording. If the recording is not played back at exactly the speed it was recorded, sounds are altered. For example, a novelty record named *The Chipmunk Song* (*Christmas Don't Be Late*) became a number one seller in the U.S. To produce high-pitched character voices, the audio was recorded at one speed and then played back at a higher speed. A video must also be played at the speed it was recorded. Playing the video at a higher speed produces a *fast forward* effect, and playing the video at a slower speed produces *slow motion* (*slo-mo*).

26.3 The Two Types Of Real-Time Transfer

Two broad approaches are used to transfer real-time information (i.e., audio and video) across the Internet:

- Complete download
- Live streaming

Complete download. When a user runs an app that transfers a copy of an entire video or audio segment from an Internet site to their device, we say that the user has *downloaded* a copy. Once a download completes, the entire copy resides on the user's device. The user can play the audio or video multiple times, and can pause and restart playback.

Live streaming. We use the term *streaming* to refer to an Internet service in which a user runs an app that requests an Internet site to send audio or video, and then plays the data as it arrives. That is, instead of downloading an entire item, the app starts playback as data arrives. The main advantage of streaming is that a user can see and hear events as they occur (e.g., live coverage of sporting events and news). We will learn that many streaming apps provide additional flexibility by allowing a user to pause playback temporarily, and possibly to rewind and repeat sections. However, the general idea behind streaming is to view the stream once, as it arrives. To summarize:

An entire audio or video segment can be downloaded before playback begins or a segment can be streamed and played as the data arrives.

26.4 Streaming Real-Time Data Over The Internet

Because the Internet uses packet switching, two problems arise when streaming real-time data:

- Insufficient throughput
- Variation in delay

A later section explains a single technology that apps use to solve both problems. We will learn about the technology and see how its output appears to users after we understand the problems.

Insufficient throughput. The problem of insufficient throughput is easiest to understand. As Chapter 12 explains, the throughput (i.e., the capacity) of a path through the Internet specifies how many bits can travel across the path per second.† When streaming audio or video, a problem occurs if the path between the sender and receiver does not have sufficient capacity to keep up with the rate of the data being sent. Audio is not usually a problem because transferring audio does not require many bits per second. However, a video stream can require millions of bits per second, especially when the video is *High Definition* (*HD*) or *4K*‡.

†Recall that providers often use the term bandwidth instead of *throughput* or *capacity*.

‡Even standard-definition video generates approximately 500 times as many bits per second as audio.

If the throughput of a path through the Internet is less than the rate at which real-time video is generated, the video cannot be displayed smoothly. Instead, the data will not arrive fast enough, and the receiving app will "run out" of data to play. The app must stall and wait until more data arrives. Users say that the picture *freezes*, and then restarts.

26.5 Real-Time Streams, Packets, And Jitter

Even if a path through the Internet has sufficient capacity for audio and video, packet switching can cause a problem. To understand why, think of a highway system, and imagine a string of cars entering the highway at a precise interval and traveling at exactly the same speed. For example, suppose a new car enters the highway every ten seconds. If there is no other traffic, one car will leave the other end of the highway every ten seconds. However, most highways are shared — they have ramps that allow other traffic to enter and exit. When other traffic merges in, some of the cars in the stream will experience a slight delay. As a result, cars on a highway tend to "clump" together. When other traffic exits, a small gap appears between clumps.

Because billions of devices share its underlying links, the Internet behaves the same way as a highway. Even if a device sends packets at a steady rate, other traffic on the Internet causes packets to clump together. As a result, we say that packets arrive in *bursts*. Bursts affect packets that carry a stream of real-time data (i.e., audio or video). Although the sender generates packets at a steady rate, the packets will not arrive at the receiver at a steady rate. Figure 26.1 illustrates the idea.

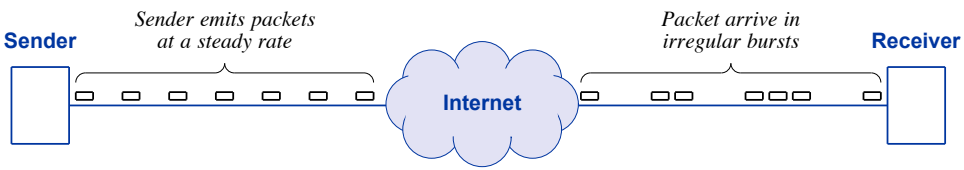


Figure 26.1 Illustration of how packets tend to clump together and arrive in bursts as they travel across the Internet.

An individual packet traveling across the Internet is analogous to an individual car traveling across a highway: the time it takes depends on other traffic. On the Internet, changes in traffic can occur in less than a thousandth of a second, which means that two successive packets may experience slightly different travel times. We use the term *jitter* to describe the variation in delay.

For real-time data, jitter causes a problem. Each packet contains a piece of the data that must be played at a precise time in the sequence. If a receiver attempts to play audio or video as packets arrive, the results are disappointing. Instead of a steady playback, the user experiences minor glitches. Whenever a large gap occurs between pack-

ets, playback must be paused for a short time until the next packet arrives. Consequently, a user listening to audio may hear a pause or a click, and a user watching a video may notice the picture freeze. Later, when a burst arrives with packets too close together, sounds become garbled and a video jumps ahead quickly.

26.6 A Playback Buffer

A clever technology makes it possible to play audio and video that has been streamed over the Internet without any problems. To understand the technique, think of a distribution center for an online smart phone retailer. The manufacturer periodically sends a truck full of phones to the distribution center. When a customer makes a purchase, one of the phones is placed in a box and shipped to the customer. Thus, a continuous stream of phones leaves the distribution center. As long as each truck arrives with enough phones, the delay between shipments will not affect customers — orders can continue to be processed. The supply at the center may dwindle, but the center will not run out because a truck will arrive to restock the inventory.

Apps that play audio or video use an analogous technique. When it first starts, the app on the receiver's device gathers many seconds of the incoming stream, and stores the data in the device's memory. Later, the app starts playing the data at a fixed rate. While playback occurs, additional bursts of packets arrive that are added to the data in the device's memory. The remaining packets will not arrive at a steady rate because they will experience jitter, and some packets will take longer to make the trip than others. A burst of packets arriving is analogous to a truck delivering phones. When a gap occurs between bursts of packets, the app can continue to play the data from the device's memory at the correct rate, analogous to a distribution center shipping phones. Like phones in a distribution center, the number of items in memory may dwindle, but if the app has estimated well, data will not run out before additional packets arrive. To summarize:

> *By gathering a set of packets before playback begins, an app can present the user with a steady playback rate even though successive packets arrive in bursts.*

Computer scientists use the term *buffer* to describe the temporary storage in memory used to hold items, and the term *playback buffer* to describe a buffer that is used for real-time data. Figure 26.2 illustrates the idea by showing a playback buffer used for a movie.

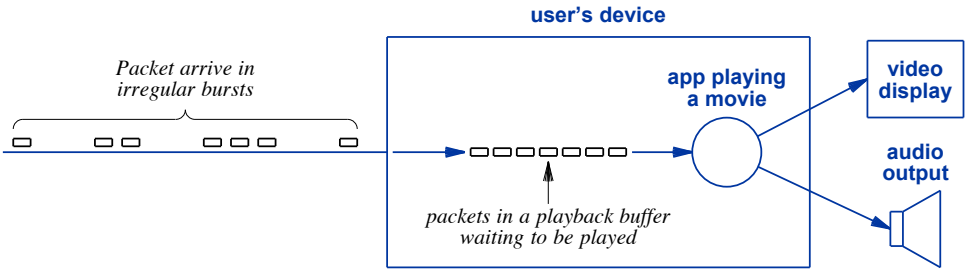


Figure 26.2 Illustration of a playback buffer in a device's memory. By keeping a supply of data, an app can play a movie at a steady rate even though packets arrive in bursts.

26.7 Accommodating Low Throughput

A playback buffer also solves the problem of a low throughput path. Suppose a user streams a movie that takes two hours to play, but the movie will take two hours and twelve minutes to transfer. An app can solve the problem by accumulating twelve minutes of data before starting to play the movie. That is, the app gives the transfer a "head start." When the app starts playing the movie, the transfer will still have two hours remaining, the same time it takes to play the movie. The point is:

> *To handle the problem of a low-throughput Internet path, an app must delay before starting to play a real-time stream until enough data is accumulated in the playback buffer so the remaining data can arrive before it is needed.*

In practice, an app usually waits slightly longer than the bare minimum because traffic on the Internet may change, increasing the transfer time. A later section considers what happens if an app underestimates the transfer time.

26.8 The User's View Of A Playback Buffer

When an app uses a playback buffer, there are two consequences for a user:

- Delayed start
- Visual illustration of the buffer

Delayed start. Before playback begins, an app must accumulate enough data to handle both low throughput and jitter. Therefore, a user will notice a delay between the time a stream is requested and the time playback begins. How long is the delay? To

estimate the amount of data needed, an app starts the stream and measures incoming packets. Gaps between bursts are usually small, so the time an app delays to cover gaps is often so short a user will not notice (e.g., a few tenths of a second). The delay needed to accommodate low throughput may be quite large, and may extend tens of seconds or minutes.

Visual illustration of the buffer. Most apps that play audio or video show the user a visual illustration of the playback buffer. The illustration consists of a long, horizontal rectangle that represents time. For a two-hour movie, the rectangle will have the label *0* on the left indicating the start as zero minutes, and *120* on the right, indicating the total play time of two hours (one hundred twenty minutes). Figure 26.3 shows a playback buffer display for a two-hour movie.

Figure 26.3 Illustration of the display an app uses to show the user a playback buffer. The display represents the time it takes to show the movie (120 minutes).

When an app displays a playback buffer, the app fills in the rectangle to show how much of the movie has been loaded into the buffer in memory. When a user first requests streaming, no data has been placed in the buffer, so the rectangle is empty, as Figure 26.3 shows. As data arrives, the rectangle fills to indicate how much data has been placed in the buffer. Figure 26.4 shows a playback buffer after the first twelve minutes of the movie has arrived and been placed in the buffer.

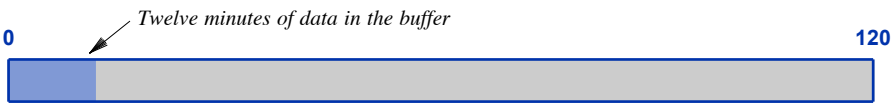


Figure 26.4 An illustration of a playback buffer when it contains the first twelve minutes of a two-hour movie.

We use the term *download point* to refer to the amount of data that has been downloaded. In addition to showing how much data has been placed in the playback buffer, an app also shows how much has already been played. We use the term *playback point* to refer to the current playback position in the movie. For example, Figure 26.5 shows how the display might appear if sixty minutes of a two-hour movie has arrived and been placed in the playback buffer (i.e., half of the data for the movie), and a user has viewed forty-eight minutes of the movie.

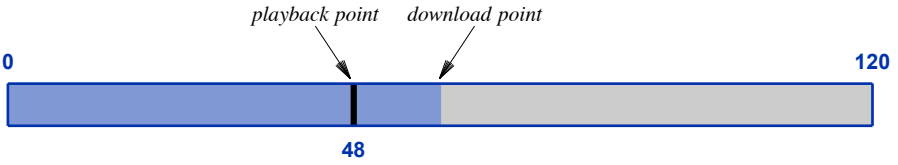


Figure 26.5 An illustration of a playback buffer when the first hour of a two-hour movie has been downloaded, and the user has watched forty-eight minutes of the movie.

26.9 The Effect Of Pausing Playback

A user can only control playback — the app controls receiving packets and placing them in the playback buffer. That is, when a user presses the *pause* button, only the playback freezes. The app continues to accept incoming data and place it in the buffer. For example, suppose the user pauses playback in the situation that Figure 26.5 shows. The playback point will remain at forty-eight minutes, but the transfer will continue and the app will place new data in the buffer. Figure 26.6 illustrates how the display will appear after the movie has been paused for a while.

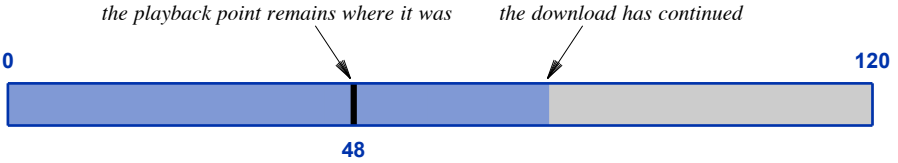


Figure 26.6 The playback buffer after a user has paused playback at forty-eight minutes. Download continues during the pause.

26.10 The Effect Of Network Congestion

When it starts, an app that plays real-time data estimates the rate at which data will arrive, and uses the estimate to calculate how much data to accumulate before starting playback. To obtain an estimate, the app measures the first packets that arrive. If conditions in the Internet remain the same for the entire stream, the estimate will be accurate, and the user will enjoy uninterrupted playback.

Unfortunately, paths in the Internet are shared. Additional traffic can appear and disappear at any time. If traffic begins using any part of the path between a sender and receiver, congestion along the path will mean that the throughput between the sender and receiver will be reduced. In the worst case, a temporary outage can mean throughput drops to zero. In either case, the throughput will be lower than the original estimate, which means data will arrive more slowly than it is being played. Eventually,

the playback point will reach the end of data in the playback buffer, as Figure 26.7 illustrates.

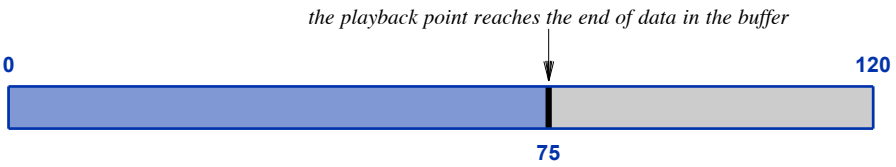


Figure 26.7 A situation in which the playback point reaches the end of data in the playback buffer.

26.11 How To Overcome A Start-Stop Cycle

When an app runs out of data to play, playback must stop temporarily until more data arrives. If the path remains congested, data may continue to arrive in bursts, with a large gap between each burst. Each time data arrives, playback will begin again, but the app will quickly use up the data in the playback buffer and then stop. From a user's point of view, a video will appear to run for a short time (less than a minute), and then freeze.

How can the problem be solved? If videos only exhibit the start-stop behavior when someone else in your home uses the Internet, it suggests that the capacity of your access network is too low. If start-stop behavior only occurs occasionally, the server sending the video may be overloaded or some link along the path between the server and the user may be congested.

A user cannot control the server sending a movie, and cannot control the capacity of the path through the Internet. However, a user does have one way to solve the problem: pausing playback. To understand how pausing playback helps, observe that start-stop behavior occurs when the playback point reaches the download point, as Figure 26.7 illustrates. We learned that even if playback is paused, an app will continue to receive packets and place them in the playback buffer. If a user pauses playback until twenty minutes of the movie have been accumulated, the movie will play for at least twenty minutes before it freezes again.† Figure 26.8 illustrates how the playback display will appear.

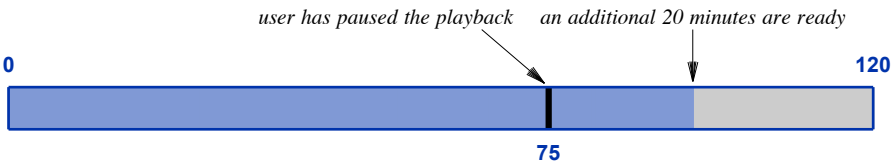


Figure 26.8 An illustration of pausing playback to overcome start-stop behavior.

†In the best case, additional minutes of the movie will be downloaded during the twenty minutes, allowing the movie to play without interruption.

26.12 Teleconferencing Services

A *teleconferencing service* permits a group of users to hold a group discussion that optionally includes video. Each user runs software on their device that connects to the discussion. The software provides two-way communication, sending video and audio from the user to the rest of the group, and playing audio and video from other users in the group.

How can a teleconference work if the Internet introduces jitter? The answer is that teleconferencing software uses the playback buffer technique described above. Unlike the playback buffer used with movies or other streaming video sources, the playback buffer used with a teleconference is small (e.g., one tenth of a second). The reason should be obvious: there is no point in delaying multiple seconds because a long delay interferes with human communication. Keeping the delay short means a user will not notice, and will assume the audio and video are instantaneous (i.e., "live").

For audio, a small playback buffer works well because it handles typical jitter. For video, however, a small playback buffer cannot handle a path with low throughput. Consequently, teleconferencing systems use two additional techniques:

- Low video resolution
- Low video frame rate

Low video resolution. One way to reduce the amount of data being sent involves using low resolution. For a teleconference, where a user sees all the other users' faces, each image is small (i.e., a *thumbnail*). When a video image is displayed in a small area of a screen, the video must be transformed into lower resolution. So, sending a high-resolution video is harmful because it increases the load on the network without achieving any useful purpose. Consequently, teleconferencing software uses low-resolution video.

Low video frame rate. Another way to reduce the amount of data being sent involves a slower frame rate. Video consists of a series of images, called *frames* that are displayed in quick succession. To fool the brain into seeing smooth motion, video typically uses thirty frames per second (higher quality video can use up to sixty frames per second). For a teleconference, most of the video shows faces while users are in front of their devices. Because the video does not contain much motion, a lower frame rate suffices.

If you use teleconferencing software, you can find out whether the two techniques are being used. To check the resolution, click on a thumbnail of a user's face, and blow up the video to full screen. Do some tiny details seem blurry? To check the frame rate, ask another user to move quickly (e.g., wave their hands back and forth in front of the camera). Does the display seem to jump from one position to the next, as if you are looking at a series of still images one after the other?

26.13 Using Internet Technology For Telephone Service

Telephone companies have realized that Internet technology has a distinct advantage over the old analog phone technology: much lower cost. Around 2000, major telephone carriers started using Internet technology to carry telephone calls. Telephone companies do not send calls over the global Internet because they have their own internal networks. However, most long-distance telephone calls now use Internet technology, specifically, the Internet Protocol. The point is:

> *Even if landline telephones are used, chances are high that the call will be sent over Internet technology at some point along the path.*

26.14 VoIP Telephones

The telephone industry uses the term *Voice over IP* (*VoIP*) to describe the use of Internet technology for telephone calls. Instead of using the phrase, they pronounce the acronym by referring to "voyp" technology.

Businesses and individuals are replacing landline phones with telephones that use VoIP technology. We use the term *VoIP telephone* or *IP telephone* to describe such a phone. Instead of a standard analog phone connection, a VoIP phone has an Ethernet connection (i.e., it connects to a computer network). The phone contains a circuit that digitizes the user's voice and sends packets plus a circuit that receives packets containing digitized audio and plays the audio for the user.

Of course, if a VoIP telephone could only place calls to another VoIP telephone, few users would choose VoIP technology. To make a VoIP phone system commercially viable, companies offering the service connect the VoIP system to the conventional phone system. Thus, a VoIP telephone is assigned a conventional phone number, and a VoIP customer can place or receive calls from analog phones, cell phones, or other VoIP phones.

26.15 Summary

Because it uses packet switching, the Internet introduces jitter, which impacts the streaming of real-time information. A technology known as a playback buffer solves both the problem of jitter and the problem of a low-throughput connection. An app that uses a playback buffer accumulates information before playback begins. If all goes well, the playback buffer allows a user to experience uninterrupted playback.

Telephone companies are switching from analog technologies to Voice over IP (VoIP), and most phone calls now use the Internet Protocol for some part of the path. It is possible to obtain a VoIP telephone; a VoIP phone has a conventional phone number, and can communicate with an analog phone, a cell phone, or another VoIP phone.

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

27

File Transfer And Data Sharing

27.1 Introduction

Previous chapters discuss Internet applications that an average user is likely to encounter. This chapter considers additional services related to the exchange of large files.

27.2 File Transfer

Internet services such as email and instant messaging can be especially useful for sending short notes, but they are not designed for sending large volumes of data. The corporate world has adopted email attachments as an easy way to transfer files. However, most email servers are configured to reject extremely large files. Thus, if a user attempts to email a movie, the email system is likely to reject the request and inform the user that the attachment is too large.

How are large files sent across the Internet? A *file transfer* service exists that can be used to transfer a copy of an arbitrarily large file from one computer to another. The service is among the oldest services available, and has been in continuous use since the invention of the Internet. Prior to the invention of the World Wide Web, file transfer accounted for more Internet traffic than any other service; in 1995, Web traffic began to dominate.

The service uses the *File Transfer Protocol* (*FTP*), and most web browsers include facilities to use FTP, which makes transfer convenient. FTP provides the following:

- Complete file copy
- Arbitrary file size
- Ability to list folder contents
- Transfer of arbitrary file types
- Optional authorization
- Transfer in either direction

Complete file copy. FTP can only copy an entire file. If a user transfers a document, the copy will contain all pages; there is no way to request part of the document.

Arbitrary file size. FTP does not limit the file size. Of course, the device that receives a copy of the file must have sufficient storage to hold the file.

Ability to list folder contents. A user can obtain a list of the files that are available for transfer. When the user uses FTP to contact a remote computer, the browser displays a list of available files, and makes each item in the list a link. When the user selects a link, the browser fetches a copy of the file. Note: FTP uses the term *directory* instead of *folder*, but the distinction only matters when the user receives an error message in response to an invalid folder name.

Transfer of arbitrary file types. FTP does not distinguish among file contents, and can transfer arbitrary types of files, including text files, images, documents, audio recordings, and video recordings.

Optional authorization. A site can make files available to all users or can restrict access. To restrict access, the site issues each user a login ID and password, and then limits specific files to specific users.

Transfer in either direction. FTP allows a user to *download* a copy of a file from a remote site to the user's device, or *upload* a copy of a file from the user's device to a remote site (upload usually requires the user to specify a login ID and password).

27.3 An Example File Transfer

Although separate FTP apps are available, most users access FTP through a web browser. In fact, you may already have used FTP without knowing it. When you click on a link labeled *Download*, your browser may use FTP to perform the download.

How does a browser know when to use FTP? The browser examines the URL associated with a link to decide which protocol to use. If the URL begins with the prefix *ftp://*, the browser interprets the rest of the URL as a request for a file transfer. For example, the following URL specifies a file that is available via FTP from a computer at Purdue University:

ftp://ftp.cs.purdue.edu/pub/comer/tib/example.txt

Try typing the URL into a browser. When a user enters the URL, the browser contacts an FTP server on a computer named *ftp.cs.purdue.edu*, and requests a file named *pub/comer/tib/example.txt*. Because it is a text file, once it obtains a copy of the file, the browser displays the file for the user. The point is:

> *The File Transfer Protocol (FTP) can be used to transfer arbitrary files across the Internet. Most users access FTP through a browser; the prefix* ftp:// *on a URL tells a browser to use FTP.*

27.4 An Example URL For Folder Contents

FTP stores files in a conventional file system that has folders that can each contain files and other folders. When a user enters a URL that corresponds to a folder, the browser will contact the specified site, request a list of the available files, and display the list for the user. The display does not include icons or images. Instead, the user sees a list of file names. Thus, FTP sites are only useful if file names make the contents self-explanatory.

An example will clarify the concept. A small FTP folder has been set up that readers of this chapter can explore. To see it, enter the URL:

ftp://ftp.cs.purdue.edu/pub/comer/tib

The folder contains three files:

Bird.pdf
README
example.txt

Note that when a browser displays an FTP folder, a user merely sees a list of file names.

> *Although FTP allows a user to view the contents of a folder, the output merely consists of a list of file names without further explanation of their contents.*

Recall that FTP offers a way to restrict file access. In fact, authentication is not optional — when it uses FTP, a browser must specify a long ID and password. A browser follows an FTP convention known as *anonymous FTP* to access *public* files. An FTP site that contains public file access honors the special login *anonymous* and password *guest*. When a user clicks on a link that corresponds to FTP, the user's browser tries the anonymous login. If anonymous login fails, the browser prompts the user for a login and password. Thus, a login prompt only appears to a user if the requested file is restricted. To summarize,

The FTP service requires the use of a login ID and password. To access public files, a browser uses the login anonymous *and the password* guest*; if the file is restricted, anonymous login will fail, and the browser will prompt the user for a login and password.*

27.5 How FTP Works

Like other Internet services, FTP uses the client-server approach. When the user enters a URL that specifies FTP, the user's browser becomes an FTP client that contacts an FTP server on the computer specified in the URL. When a user selects a file, the browser requests a copy from the server, which sends the data over the Internet. Figure 27.1 illustrates the interaction.

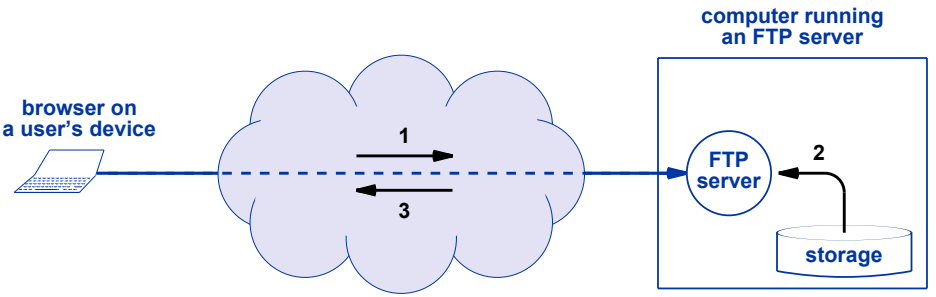


Figure 27.1 Illustration of file transfer in which (1) a browser uses FTP to request a file, (2) the FTP server reads the file from its local storage, and (3) the server returns a copy to the browser.

27.6 File Transfer For An Average User

The chief disadvantage of the FTP approach to file transfer arises because anyone who wants to make files available to others must run an FTP server. An average user cannot run a server easily because the user does not have a permanent IP address, does not keep their devices running constantly, and does not have the expertise needed to configure and operate a server.

How can users transfer large files? Several Internet services have appeared to enable file transfers without requiring users to run servers. Examples include:

- File sharing
- Photo sharing
- Video sharing
- Document sharing

File sharing services offer users the opportunity to exchange files. A user can upload a file and make it available to others for download. Examples include *Dropbox* and *Google Drive*.

Photo sharing services allow users to upload photos that are then available to others. Examples include *Shutterfly*, *Instagram*, *Google Photos*, *Flickr*, and *iCloud*.

Video sharing services allow users to share videos. A user can post a video that other users can view. Examples include *YouTube*, *Vimeo*, *Daily Motion*, *Twitch*, and *Live Leak*.

Document sharing services allow users to share documents. A user can upload a document to the site, and then allow other users to obtain a copy. Some sites (e.g., Google Docs) permit users to edit documents.

27.7 Exchanging Information Without Running A Server

The services described above all follow the same basic approach: the service operates a web site that has a server. A user who wants to share an item with others must contact the site and upload a copy of the item from their device to the site. Later, users who want to obtain a copy of the item contact the site and download a copy to their local device. Figure 27.2 illustrates the steps.

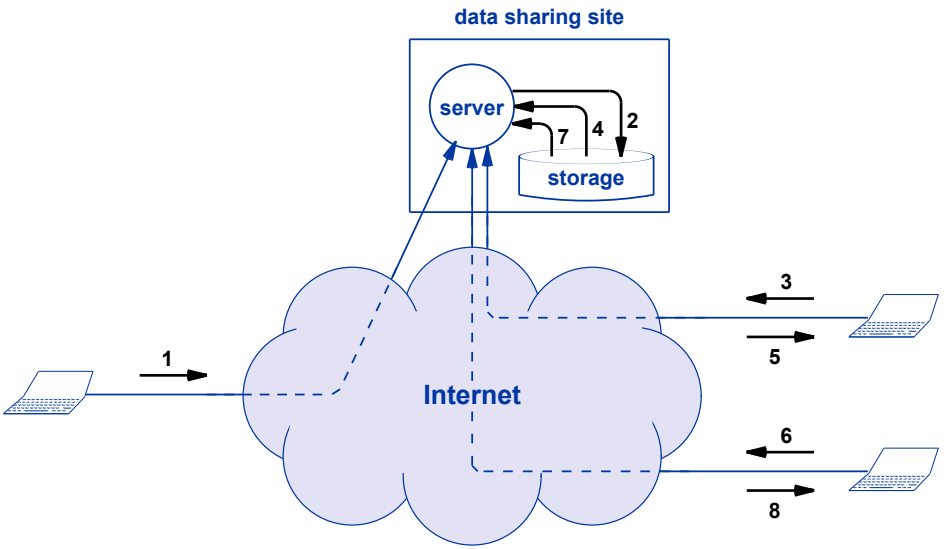


Figure 27.2 The steps taken to share a data item: (1) a user uploads an item, that (2) the server places on storage, (3) another user requests the item, (4) the server obtains a copy from storage, and (5) downloads the item to the user. (6,7,8) Other users also obtain a copy.

27.8 Transfer Vs. Collaborative Work

The basic idea behind many servers can be described as transfer: a user creates an item and makes it available to others. For such services, a sharing site merely serves as a repository — the site holds a copy of the information and provides a way for others to access the information.

As mentioned above, some sharing sites extend the basic paradigm to permit a group of users to collaborate by modifying data items. For example, a *wiki* site allows users to edit a web page. Some document sharing sites apply the same idea by allowing a group of users to edit a document. A site that provides more than a repository foreshadows the discussion of *Cloud Computing* in Chapter 29. For now, we can assume that each service operates its site independently.

27.9 Peer-To-Peer File Sharing

As the twenty-first century began, a new form of sharing became popular on the Internet. Known euphemistically as *file sharing*, the idea started from the observation that if two users each had digital music stored on their device, it would be easy to swap copies. Thus, if one user had songs from artist *A* and another user had songs from artist *B*, it would be trivial to pass copies to each other, giving each user songs from both artists. The system quickly degenerated from a few friends exchanging a few songs to a way for large groups of individuals to make (illegal) copies of copyrighted material. Soon, complete strangers were participating in file sharing schemes, with digital music and videos being the most popular items exchanged.

How did the illegal file sharing scheme work? Unlike a traditional Internet service that requires an expensive, powerful server computer to handle requests, the illegal file sharing used small, slow devices. In essence, an individual participating in the scheme agreed to let their device act as a duplication service — in exchange for the right to obtain copies of files, the user ran software that forwarded copies of files to others. Thus, instead of all users obtaining copies of files from a single server, each user requested files from a nearby device. The file sharing software kept track of which files resided on which devices. When a file request arrived, the receiving device either returned a copy (if the file was available in local storage), or suggested another device to contact.

Because it relied on communication among arbitrary devices, the illegal file sharing was dubbed a *peer-to-peer application*, abbreviated *P2P*. The idea behind peer-to-peer communication is extremely powerful, and groups of users have explored ways to use the peer-to-peer approach for legitimate purposes (e.g., to propagate non-copyrighted material quickly). To summarize:

An illegal music and video sharing service was created in which users exchanged copyrighted material with other users. The scheme is known as peer-to-peer (P2P) file sharing because arbitrary devices participate without the need for a server. Although originally associated with illegal sharing, P2P can be used to propagate legitimate files rapidly.

27.10 Summary

Although email and messaging systems can be used to transfer small data items, a file transfer mechanism is needed to transfer arbitrarily large files. One of the oldest file transfer services uses the *File Transfer Protocol* (*FTP*). FTP allows a user to download files, upload files, and list the contents of folders on a remote computer. Users access FTP through a web browser — when the user supplies a URL that begins with the string *ftp://*, the browser becomes an FTP client and accesses an FTP server.

A variety of data sharing services exist, including file sharing, photo sharing, video sharing, and document sharing. The services permit users to share items without requiring any of the users to own and operate a server. In most services, a user uploads a data item, and other users can then download a copy. Some services permit collaboration in which users can modify an item.

Illegal file sharing was once popular as a way for users to exchange copyrighted materials, such as music, movies, and books. The mechanism is classified as a peer-to-peer application because each participating user agrees to make copies of files available to others in exchange for the right to access the files others have.

EXERCISES

27.1 Use FTP to transfer the example file, and describe what you see:

ftp://ftp.cs.purdue.edu/pub/comer/tib/example.txt

27.2 If you enter the following URL, which file names appear?

ftp://ftp.cs.purdue.edu/pub/comer/tib

27.3 Extend the previous exercise: click on the file name *Bird.pdf* and describe what you see.

27.4 Make a list of popular photo sharing sites (hint: use a search engine).

27.5 Try uploading a photo to a photo sharing site, and ask a friend to download a copy.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

28

Remote Desktop

28.1 Introduction

Previous chapters describe a variety of Internet services. This chapter continues the discussion by focusing on a service that allows a user to access and control a computer from a remote location. The chapter describes both the motivation and the technology. The next chapter continues the discussion by explaining how remote desktop access forms an important component of cloud computing.

28.2 Remote Login

One of the earliest Internet applications consisted of software that allowed a user to access and control a remote computer. To understand the software, one must know about the computers that existed when the Internet was created. At that time, computers had a textual user interface. A user typed on a keyboard and viewed a display screen; the display could only show text (usually 24 rows of 80 characters per row). There were no icons, no mouse, and no graphics. A user logged into a computer by entering a login ID and password. The user then entered a series of commands. For example, to invoke an app, a user entered the name of the app.

Software was created that allowed a user on one computer to log into another computer. A user invoked a *remote login* app. The user was asked to specify the name of a remote computer as well as a login ID and password to access the computer. Once access was established, every keystroke the user entered was sent to the remote computer, and all the output from the remote computer was sent back and displayed on the user's screen. From a user's perspective, the output on the screen was exactly the same as if

the user had physically moved to the computer and logged in. If the remote computer had a keyboard and display attached, neither was used during a remote login session. Figure 28.1 illustrates the idea.

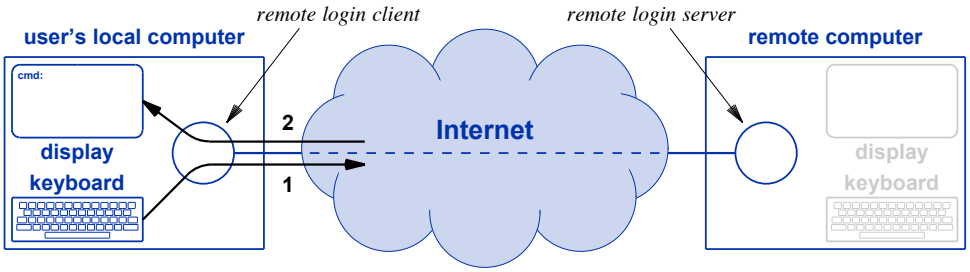


Figure 28.1 Illustration of remote login on an early computer. (1) Each keystroke was sent to the remote computer, and (2) output was sent back to the user's display. The display and keyboard on the remote computer remained inactive.

Once a user finished using the remote computer and logged out, control of the user's keyboard and display were returned to the user's local computer system. That is, keystrokes were once again processed locally, and apps running on the local system displayed output on the screen. The point is:

> *A remote login system used on early computers allowed a user to run a remote login app that connected the user's keyboard and display to a remote computer. Once the user finished using the remote computer, control of the keyboard and display returned to the user's local system.*

28.3 Remote Access With Modern Graphical Devices

The basic idea of remote login has been extended to provide remote access to modern devices. The goal remains the same as with remote login: provide a user with an experience identical to the one users have when they use a device in person.

Modern hardware changes the remote access paradigm. The chief difference between early computers and modern devices lies in the form of user interaction — instead of being limited to letters, digits, and punctuation, a modern device offers users a graphical interface. The main screen displays a *desktop* with *icons* that a user can select to launch apps. Selection either requires a user to move the cursor (with a mouse or other tracking device) or to use their finger with a touch screen. When an app runs, the app can take over the screen or can create its own *window* on the desktop.

The computer industry uses a variety of terms to describe remote access on a modern system, including *remote desktop*, *screen sharing*, and *remote desktop connection*; we will use the term *remote desktop*. To use a remote desktop to access and control a remote system, a user must be able to provide the same inputs as on the remote system. In particular, a user must be able to see a copy of the remote screen, move the cursor on the screen, select icons, and enter text.

Will a user have an identical experience when using a remote desktop system? They will not unless their local device offers essentially the same hardware features that the remote system offers. In particular, the screen size must be approximately the same. To understand why, imagine a user with a small tablet computer trying to access a remote desktop computer that has a large display screen. On the one hand, if the software displays a copy of the entire remote screen on a small display, items will appear much too small to see and read. On the other hand, if the software only displays part of the remote screen, and requires the user to select which part is visible at a given time, the user will spend time trying to navigate around the display. In either case, the experience will not be identical to the one a user receives when using the remote device directly. We can summarize:

> *Remote desktop systems cannot provide an experience close to what a user would experience in person unless the user's local device has approximately the same hardware facilities (especially the same screen size) as the remote system.*

28.4 How Remote Desktop Works

The basic software components needed for remote desktop remain the same as the components used for remote login. The remote device must run a server, and the user runs an app that acts as a client. As with remote login, the app on the user's device takes control of the user's display and keyboard. The app must also take control of the pointing device (e.g., the mouse or trackpad). Every time the user moves the pointing device, selects an icon, or enters a key, the app sends the information to the remote system. Whenever the screen on the remote system changes (including movement of the cursor), the server on the remote system sends an exact copy of the screen to the client app so it can be displayed on the user's screen.

Interestingly, many remote desktop systems allow the remote system to have a display, and keep the display active while the remote desktop session proceeds. When the user moves the mouse, the motion is sent to the remote system, and the cursor moves. When the user enters text, the text appears on the remote screen. Thus, if a human happens to be near the remote screen during a remote desktop session, they will see the cursor move, icons being selected, and apps running. Figure 28.2 illustrates a remote desktop session.

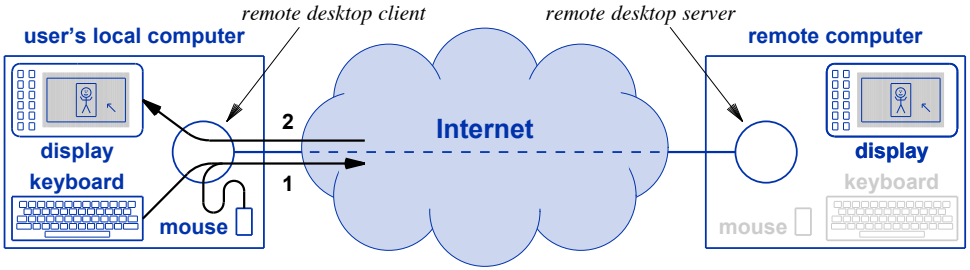


Figure 28.2 Illustration of a remote desktop session in which (1) mouse input and keystrokes are sent to a remote computer, and (2) after the remote display changes, a copy is sent back and shown on the user's display.

28.5 Remote Desktop Software

Vendors who sell desktop and laptop operating systems usually offer remote desktop software to accompany their systems. For example, both Windows and MacOS include remote desktop software. MacOS comes with both a client app (used to access a remote system) and a server app (used to allow others to access your system). The Home Edition of Windows 10 only includes client software; one must purchase the Pro Edition of Windows 10 to obtain server software.

In addition to operating system vendors, third-party vendors also offer remote desktop software. Some versions of remote desktop technology are available without charge; others require a user to purchase a license. For example, a company known as *RealVNC* distributes a free version of *Virtual Network Computing* (*VNC*) software as well as improved versions that require a paid license. A description can be found on the following web page:

http://www.vnc.com

28.6 Assessment Of Remote Login And Remote Desktop

Remote access services are significant for three reasons. First, the services provide mobile users with access to powerful computers without requiring the users to tote heavy hardware devices or heavy batteries. Second, instead of trying to synchronize data on multiple devices, a user can keep their data on a single computer and then access the computer from multiple sites (e.g., from home and work). Third, remote access is a key building block of the cloud computing paradigm described in the next chapter.

Heavy equipment is not the only motivation for using remote access. In addition to higher computational power, the remote computer may have facilities that would be difficult to provide at arbitrary locations. For example, consider a salesperson who

needs to access a large database of customers, warehouses, and inventory. The database changes frequently. The large size and frequent changes make downloading a copy of the database onto the user's laptop impractical. Remote access facilities mean the employer can place the database on a powerful server. The sales staff can then use remote desktop software to access the database as needed.

One of the most impressive aspects of remote desktop services stems from their ability to reproduce even the smallest details. If the path through the Internet between the user and the remote computer does not experience congestion and delay, the software can provide the illusion of "being there."

28.7 Unexpected Results From Remote Access

Although remote access services are convenient, the results can be confusing to a user accustomed to doing all computing on a local device. To understand why, remember that the user sees and touches one device while running applications on another device. In many cases, the local app takes over the display, meaning that once a remote desktop service has been launched, the user will see the remote desktop instead of the local desktop.

To understand how confusion occurs, consider a user who connects to a remote desktop and launches a word processor application. After creating a document, the user chooses to save the document. Although it appears in a window like any other app, the word processing app actually executes on the remote system. If the user saves the document on the desktop, a copy will be placed on the desktop *on the remote computer*. Thus, once the user quits using the remote access app and control returns to the user's device, the document will not be accessible.

Another unexpected consequence of remote access arises from the inability to access local facilities. For example, many apps allow a user to print. However, an app running on a remote system will only have access to printers attached to the remote system. Thus, when using a remote desktop to run an app, the app will not be able to send output to a printer that is connected to the user's local device.

We can summarize:

> *When using a remote access service to run apps, a user must remember that although a desktop and apps appear on the local display, the apps can only access files, printers, and other facilities on the remote system.*

28.8 Summary

A remote access facility permits a user who is using one device to access and control another device. Because early computers used textual interfaces, the first remote access mechanisms, known as remote login mechanisms, sent keystrokes to a remote system and displayed the text that the system returned. More recent services offer a remote desktop capability that shows the user a graphical interface. Client software running on a user's local system sends mouse events and keystrokes to a remote system. When the display on the remote system changes, including cursor motion, the remote desktop server returns a copy of the updated screen, which is displayed for the user to see.

The goal of all remote access systems is straightforward: provide a user with the illusion of working directly on a remote computer. Remote access software does an excellent job of meeting the goal; the illusion is only broken when congestion along the path through the Internet introduces high delay. The next chapter explains that remote access is important for cloud computing.

Remote desktop services have several advantages, including allowing a mobile user to access a powerful computer without carrying heavy hardware or a heavy battery. However, remote access technologies have unexpected consequences: although apps appear on a user's display as if they are running locally, such apps can only access files, printers, and other facilities on the remote system. Thus, a file saved to the desktop will only be saved on the remote system.

Chapter Contents

29 Cloud Services And Cloud Computing

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

29

Cloud Services And Cloud Computing

29.1 Introduction

This chapter describes a computing paradigm that has become known broadly by the name *cloud computing*. The cloud paradigm represents a major shift in the way individuals and corporations use computing, and if the trend continues, the new paradigm will affect just about everyone.

The chapter explains the reason computing is moving to the cloud. It examines the potential benefits for individual users, and describes why cloud computing is required for the Internet of Things devices described in Chapter 24. The chapter also considers why a corporation would choose to adopt the cloud approach, and considers the three forms: public cloud, private cloud, and hybrid cloud.

Finally, the chapter reviews the underlying technologies that enable the cloud approach. It shows how remote access (described in the previous chapter) forms one of the foundations. It also explains cloud data centers and virtualization.

29.2 A Brief History Of Computing

Computing changed dramatically in the late twentieth and early twenty-first centuries. Early experiments in the 1940s led to the first commercial computer companies in the 1950s. Scientists and engineers produced advances in hardware and software technologies, and looked for new ways computers could be used.

Over the years, the physical size of computers shrunk dramatically. Meanwhile, computers' processing power and storage capabilities increased dramatically. Figure 29.1 summarizes some of the historical highlights.

Era	Computing Facilities	Number Of Computers
1960s	Mainframes	one per organization
1970s	Minicomputers	one per department
1980s	Personal computers	one per family
1990s	Laptop computers	multiple per family
2000s	Smart phones and tablets	one or both per individual
2010s	Smart devices	many per individual

Figure 29.1 Major eras in computing and the type of computers used in each.

Mainframes. In the 1960s, a single computer, called a *mainframe*, consisted of many large cabinets, and occupied most of a room. Only a large organization could afford a computer, and the computer served the entire organization.

Minicomputers. By the 1970s, somewhat smaller, less expensive computers appeared. Each department in an organization could afford their own computer, which meant that only fifty to one hundred users needed to share a computer. Organizations used computer networks to connect minicomputers.

Personal computers. The emergence of inexpensive *personal computers* in the 1980s changed computing in a significant way. In the business world, each employee could have a computer on their desk. In addition, each individual family could afford their own computer.

Laptop computers. By the 1990s, computer hardware was so small that portable, battery-powered laptops became available. Businesses gave each employee a laptop.

Smart phones and tablets. The advent of smart phones and tablets further changed computing. Suddenly, every individual, including children, could carry a computing device.

Smart devices. The era of smart devices adds an interesting twist because smart devices outnumber individuals. Surprisingly, a modern smart phone has more computational power and larger storage than an early mainframe that was shared by an entire organization.

When one considers the history above, a trend becomes apparent: computing has moved from highly *centralized* to *distributed*. The trend is away from computers that are shared by many individuals to a situation in which each individual has their own computing device. To summarize:

Computing has moved from a shared, centralized model in which a given computer was shared by many users, to a distributed model in which each user carries their own computing device.

29.3 Maintaining Computers

The shift from centralized to distributed computing has an important downside. When computing followed a centralized paradigm, each organization hired a professional Information Technology (IT) staff to install and operate their computer. To ensure the hardware remained operating correctly, the staff ran hardware diagnostics periodically. Whenever a new version of the operating system appeared, the staff would apply updates. Similarly, when new applications or updates for existing applications became available, the IT staff handled the installation.

In a situation where each user has their own smart phone, laptop, or desktop, the user must assume responsibility for hardware and software maintenance. Each user must choose when to acquire new devices. They must configure their own devices (e.g., specify which Wi-Fi networks to use, choose which apps to install, choose when to install operating system and app updates, and handle the tasks of downloading and applying updates). The point is:

> *When a user has their own devices, the user must act as their own IT staff by configuring the device, installing new software, and updating software.*

29.4 Data Inconsistencies

The lack of an IT staff is only one of the disadvantages of a computing environment in which each user manages their own devices. A second problem arises from *data inconsistencies*. That is, the data on one of the user's devices may differ from the data on other devices.

Even if you have not experienced data inconsistencies yourself, you may have heard others complain. For example, someone might say, "I can't give you Bob's phone number because it's on my other phone," or "I know I loaded that app, but it must be on my tablet instead of my phone."

29.5 Data Synchronization With A Direct Connection

Once users began to acquire multiple devices, software appeared to allow them to *synchronize* data across their devices. For example, Apple provided synchronization for their MP3 player (i.e., iPod). Apple's design used a direct connection between a pair of devices (i.e., a cable plugged into the two devices). Apple's software allowed a user to maintain their music library on a computer, such as a desktop or laptop. The user logged onto their computer, and used the Internet to purchase and download songs, movies, and TV shows. Later, when a user connected a cable between their computer and an iPod, a synchronization app was launched. The app compared the contents of the iPod to the contents of the computer, found a list of items on the computer that were

not on the iPod, and loaded a copy of the items onto the iPod. Once the synchronization app finished, the iPod contained an exact copy of the music on the computer.

Synchronization sends data in both directions. For example, consider synchronizing a smart phone and a laptop. If the user has downloaded new songs on the laptop, synchronization software will place a copy on the smart phone. Similarly, if the user has taken photos with their smart phone, synchronization software will place copies of the photos on the laptop. Figure 29.2 illustrates the idea of two-way synchronization over a direct connection.

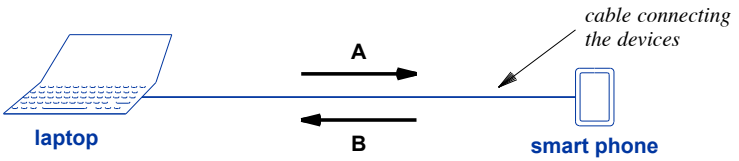


Figure 29.2 Illustration of data synchronization using a cable in which (A) data items on the laptop are copied to the smart phone, and (B) data items on the smart phone are copied to the laptop.

29.6 Selecting Data Items For Synchronization

Which of a user's data items should be synchronized across devices? There is no easy answer because the intended uses of devices may differ. For example, if a user never watches movies on their smart phone, placing copies of movies on the phone uses storage space without any benefit. Even if a user watches movies on a given device, only some formats may be appropriate. To see why, consider a device with a small, low-resolution screen that is incapable of displaying high-definition video. A user may decide not to load copies of high-definition movies onto such a device. The point is:

> *Because intended uses of a device dictate which data items should be stored on the device, most synchronization software allows a user to specify which items to synchronize.*

29.7 Synchronization Problems And Internet Synchronization

A direct connection between a pair of devices has two disadvantages:

- Compatible hardware requirements
- Pairwise synchronization

Compatible hardware requirement. A cable cannot be used to connect a pair of devices unless the devices each have the same interface hardware. For example, if both

devices have a USB port, a USB cable can be used to connect them. However, if one device has only a USB connector and the other device has only a vendor's proprietary connector, the two devices cannot be connected by a cable.

Although a user's devices may have incompatible hardware interfaces, almost every device now has some way to connect to the Internet. Thus, services have been created that use the Internet to synchronize devices. The idea is straightforward: the user runs an app on one device that contacts the service and uploads a copy of data from the device. Then, the user runs an app on another device that downloads a copy of the data. Figure 29.3 illustrates the steps involved.

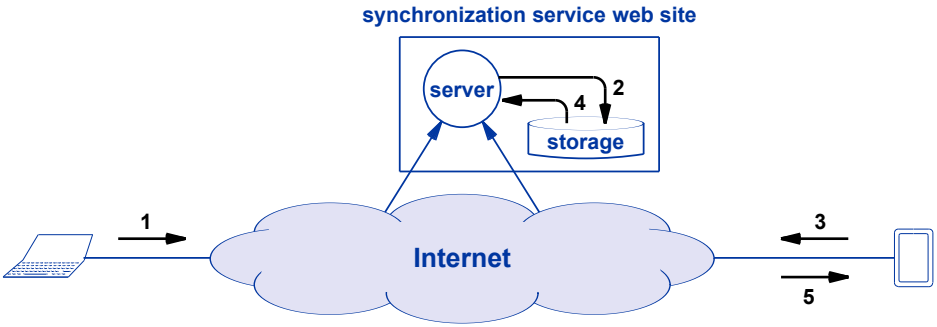


Figure 29.3 The steps when using an Internet synchronization service: (1) a device uploads data, (2) the data is placed on storage, (3) a second device contacts the server, which (4) retrieves and (5) sends the data to the device.

Pairwise synchronization. Using a direct connection means only two devices can be synchronized at a time. Pairwise interconnection may not produce the expected result. For example, consider a user who wants to synchronize data across four devices: a laptop, a tablet, a smart phone, and a media player. It may seem that the user could follow a synchronization plan like the one that Figure 29.4 illustrates. In Step 1, the user connects and synchronizes the laptop and tablet. In Step 2, the user connects and synchronizes the tablet and phone. In Step 3, the user connects and synchronizes the phone and media player.

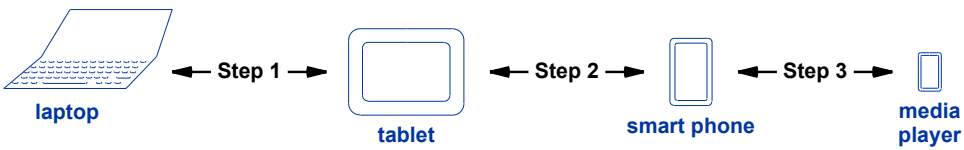


Figure 29.4 An incomplete plan for synchronizing data across four devices.

After the three steps, each device will have participated in synchronization. Does that mean the data on all devices will be the same? No. In Step 1, the laptop and tablet synchronize, which means the data on those two devices will be identical at that point in time. In Step 2, the tablet and phone synchronize. The phone and tablet will have identical data, but the laptop will not receive new items from the phone. Similarly, in Step 3, the media player will receive copies of items on the laptop, tablet, and phone, but only the phone will receive new items from the media player.

Two more steps must be added to the plan to ensure that all four devices receive copies of all data items. Is it obvious what steps are needed, or do you have to think about it? If you have to think a bit, you are like most people. In fact, that's one of the problems with pairwise synchronization:

> *Using pairwise connections to synchronize devices can be tricky because a user must think about which data items have been copied to which devices.*

One possible way to complete the synchronization of Figure 29.4 consists of resynchronizing the phone and tablet, and then resynchronizing the tablet and laptop. That is, a user must reconnect the phone and tablet, and then reconnect the tablet and laptop. Figure 29.5 illustrates the complete series of steps.

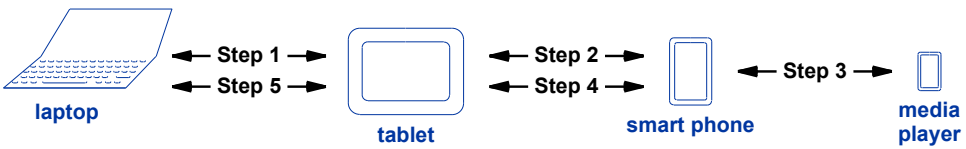


Figure 29.5 Five steps that will completely synchronize data across four devices.

Does using an Internet synchronization service reduce the number of steps? No, the number of steps increases! Instead of connecting a pair of devices, each step connects a device to the synchronization service. Think of two rounds. In the first round, each device contacts the service and uploads a copy of the data from the device. In the second round, each device must connect to the service again to download a copy of all the data that has been collected. The only exception is that the last device to connect in the first round does not need to participate in the second round.

As an example, consider the four devices in Figure 29.5. In the first round, a user might connect the laptop, tablet, smart phone, and then the media player to the synchronization service. At that point, the service will have a copy of the data on all devices, and the media player (the last one to connect) will have a copy of all data from the other devices. In the second round, the user must connect the laptop, tablet, and smart phone to the synchronization service (the order is not important).

29.8 Cloud Terminology

We have used the term *Internet services* throughout the text to describe a service that is accessed over the Internet. Starting in approximately 2010, the marketing departments in companies offering Internet services began changing terminology, and started referring to their offerings as *cloud services*. Furthermore, they claim that such services operate *in the cloud*. For example, a marketing blurb might claim that you can "store your data in the cloud."

The cloud terminology arises because networking professionals use a cloud to depict the Internet, just as figures in the text do. You have already learned enough Internet technology to know that the marketing terminology is inaccurate. Services never run "in the Internet." Instead, all services run in computers attached to the Internet. Thus, it would be much more accurate (but less persuasive) for companies to advertise that their services run *outside* the cloud. We can summarize:

> *Although marketing blurbs imply that* cloud services *are special because they somehow run inside the Internet, all Internet services run on computers attached to the Internet.*

29.9 Types Of Cloud Services

Cloud services have evolved in three stages:

- Cloud storage
- Cloud applications
- Generalized cloud computing

Cloud storage. Some of the first services to use the term *cloud* focused on a new approach to storage. Instead of placing data items on specific devices and then requiring users to synchronize copies of data across all their devices, a *cloud storage service* places data items on a server where they can be accessed at any time by any of the user's devices. That is, cloud storage services make remote storage the primary repository of data items, and then allow a user's devices to *access* the data as needed.

Early cloud storage focused on specific types of data. For example, Chapter 27 discusses photo sharing services in which a user places photos on a remote server where they can be accessed by others. Cloud photo services extend the basic idea by placing the primary copy of photos on a server rather than on a user's device.

Cloud applications. A second step in the use of the cloud occurred when companies began offering *cloud applications*. The distinction between a cloud storage service and a cloud application arises from the location where apps run. With a cloud storage service, only the data is kept on a remote server; the app used to access the data runs on the user's local device. As an example, consider two users working together to create a document. If the two use a cloud storage service, the document resides on a remote cloud server. If a user edits the document, the word processing app runs on the

user's device. The app obtains a copy of the document, makes changes, and sends the changes back to the cloud storage service. If a second user edits the document, the word processing app on the second user's device accesses the document and sends changes back to the cloud storage service. The key idea is that apps always run on local devices.

Cloud applications change the paradigm by running apps on a remote server. For example, consider a cloud application for *collaborative document preparation* (e.g., *Issue*, *Google Docs*, and *Overleaf*). Instead of running a word processing app on a user's device, a cloud application keeps both the document and the word processing app on a remote site. When a user logs into the service and launches a word processing app, the app runs on the remote server — the user's device merely provides a display and a way to enter keystrokes. Interestingly, if multiple users access a service at the same time, the technology propagates changes to each of them rapidly. Thus, when one user changes a document, all other users see the change almost immediately.

29.10 Cloud Applications And The Internet of Things

Three concepts covered in earlier chapters explain how the Internet of Things (IoT) and cloud applications are connected. Chapter 15 describes the client-server paradigm that Internet applications use to communicate. The chapter explains that a server waits for contact, and a client must use the server's IP address to contact the server. Chapter 17 describes wireless routers, and explains that a wireless router provides Internet access to devices by issuing each device a temporary IP address. Finally, Chapter 24 discusses IoT devices found in a home, and explains that the devices can be accessed from remote locations (e.g., by an app running on the user's laptop or smart phone).

The ideas described in the previous paragraph all seem sensible until we consider one additional fact: a user's laptop or smart phone will have a temporary address. To see why having a temporary address is a problem, consider a user sitting at a coffee shop who decides to check whether they left the oven on at home. We'll assume the user's oven is an IoT device that has obtained a temporary IP address from the user's wireless router. In the coffee shop, the user's device has also obtained a temporary IP address. For client-server communication to work, the server must have a permanent IP address. In other words:

> *If two devices on the Internet each have obtained a temporary address, the two devices cannot communicate directly.*

Vendors who sell IoT devices offer a way for a user to communicate with an IoT device, even if both the IoT device and the user's device have temporary IP addresses: a specialized cloud application service. In essence, the vendor obtains a permanent IP address and runs a server. The IoT in the user's home contacts the vendor's server. When a user wants to connect to one of their IoT devices, the user runs an app that also contacts the vendor's server. The two sides must be configured to provide the same

user ID, allowing the vendor's server to match the two sides and pass data between them. Figure 29.6 illustrates how a user sitting in a coffee shop communicates with an IoT device in the user's home through the IoT vendor's cloud server.

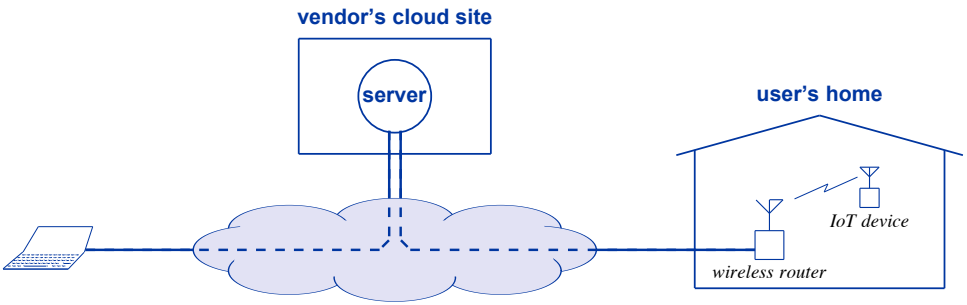


Figure 29.6 Illustration of a user at a remote location communicating with an IoT device at home through a vendor's cloud server. Both sides can have a temporary IP address because both act as clients.

29.11 Generalized Cloud Computing

The third type of cloud service is quite generic. Instead of providing a specific application, a generalized *cloud computing* service moves all applications and data to a cloud server. That is, a *cloud provider* has a large set of computers, and leases computers to customers. A customer can choose which operating systems and which apps to run on the leased computers, and can choose which data to store on the computers.

We will learn that cloud technology is quite sophisticated. Our description implies that a cloud provider has a large set of physical computers, and leases each computer to a customer. A later section explains that the technology is surprisingly flexible, and offers two important features:

- Elastic service
- Pay for use

Elastic service. Cloud technology allows a provider to change the amount of processing dedicated to a given customer. When a customer runs many apps and needs more than one computer, the cloud system can allocate multiple computers to the customer, and spread the customer's apps among them. Later, when the apps finish running and the customer no longer needs multiple computers, the cloud system releases the extra computers for other customers to use. We say that cloud technology provides *elastic service* because the facilities available to a user seem to "stretch" temporarily to accommodate the user's needs.

Pay for use. Although elastic service is convenient, another feature makes it even more desirable: a billing system that only charges for use. That is, instead of charging a user a fee large enough to cover all the user's possible needs, the system keeps records

of the processing dedicated to the user at any time, and only charges the user for the processing actually used. We can summarize:

> *Cloud technology provides an elastic service in which a customer leases computers when needed, and only pays for the amount of processing actually used.*

29.12 Cloud Computing From A Company's Perspective

Many companies are moving some or all of their computing to the cloud. That is, instead of hiring Information Technology (*IT*) staff and paying to install and maintain a set of computers locally, a company signs a contact with a cloud provider. Of course, employees still need a way to access the cloud, but the company does not need to maintain a set of large, server computers.

The move toward cloud computing is driven by economics. Cloud providers advertise two advantages:

- Reduced opex (operational expenditure)
- Reduced capex (capital expenditure)

Reduced opex. A company that uses a cloud provider can reduce the size of its expensive IT staff. Cloud providers argue that their cost for staff is lower because staff expertise is shared across all customers. A company may need to hire an individual with specialized skills, even if the skills are not used all the time; a cloud provider can share such an individual across multiple customers.

Reduced capex. A company that uses a cloud provider does not need to acquire or upgrade server computers. Cloud providers argue that they can acquire hardware at lower cost because they have an economy of scale (i.e., they can negotiate large quantity discounts).

One of the key arguments in favor of using a cloud provider arises because in many businesses computing demand varies. For example, a tax preparation firm will need the most computing cycles when taxes are due. A company that runs beach resorts will have heavy usage during summer months. Cloud providers argue that by having multiple customers (called *tenants*), they can average costs over all tenants, whereas an individual company will need to have sufficient equipment to handle the peak load, even if the equipment is idle part of the time.

To understand why cloud computing appeals to a company, think of how a company uses computers. The company runs a web site, manages employee records and payroll, performs various accounting tasks, and manages a set of internal databases. In each case, the company uses high-power server computers that do not have a display, keyboard, or mouse. Instead, IT staff use tools that allow them to configure and operate server computers without being physically present. Thus, it doesn't matter whether the server computers are located at the company or at a cloud provider site — the same computing tasks can be performed at either location.

29.13 Public, Private, And Hybrid Cloud

The use of cloud computing falls into three broad categories:

- Public cloud
- Private cloud
- Hybrid cloud

Public cloud. The term *public cloud provider* describes a company that sells a cloud computing service. A provider is "public" in the sense that the provider offers cloud service to the general public — any individual or organization can purchase service and begin using leased computers. When a company becomes a customer of a public cloud provider, the company says that it is using the *public cloud*. Despite using the term "public," a public cloud provider never reveals a customer's data to others. Instead, cloud technology keeps each customer's data confidential.

Private cloud. For most companies, the protections offered by public cloud providers suffice. However, some companies have additional requirements for their data. For example, the government imposes regulations on financial institutions that restrict how such institutions can store and share sensitive information. Similarly, defense contractors that handle classified information cannot risk storing it on a remote public cloud server. Even if a company must keep their data "in house," it may be economically advantageous for the company to create its own cloud service internally. The idea works best for larger companies where sharing computational facilities across all divisions can reduce overall costs. We use the term *private cloud* to describe a cloud service that a company creates for use within the company. The private cloud can be configured to enforce extra restrictions on data to ensure the company remains compliant with all regulations.

Hybrid cloud. The largest companies often use a *hybrid cloud* approach in which the company runs its own private cloud facility for sensitive data, and uses a public cloud provider for non-sensitive data. For example, some companies use the public cloud for their corporate web site, online catalog, customer support database, and other public information, such as annual stockholder's reports. The company uses its private cloud for all other company data and processing.

29.14 Cloud Data Centers And Racks Of Computers

We use the term *cloud data center* to refer to a cloud provider's computer site. A cloud provider that offers service in multiple geographic areas may have more than one data center. Each data center contains a large set of high-power server computers along with networking equipment that connects the computers to one another and to the Internet. Some data centers separate storage from computers. That is, instead of placing a disk in each computer, the disks are mounted in separate cabinets in the data centers, and a network connects between computers and disks.

A typical data center has thousands of server computers mounted in tall steel cabinets called *racks* that are placed side-by-side in rows. A rack is approximately six and one-half feet tall, two feet wide, and three and one-half feet deep. Thus, placing forty racks side-by-side produces a row eighty feet long.

Instead of conventional computers, data centers use equipment that mounts in a rack. The height of each piece of equipment is measured in *rack units* (*Us*), where a unit is 1.75 inches. A typical server computer is one unit tall (*1U*), and a rack can hold forty-two units. Figure 29.7 illustrates a row of racks that are each filled with equipment; in the back, cables carrying power and network data connect to equipment in each rack.

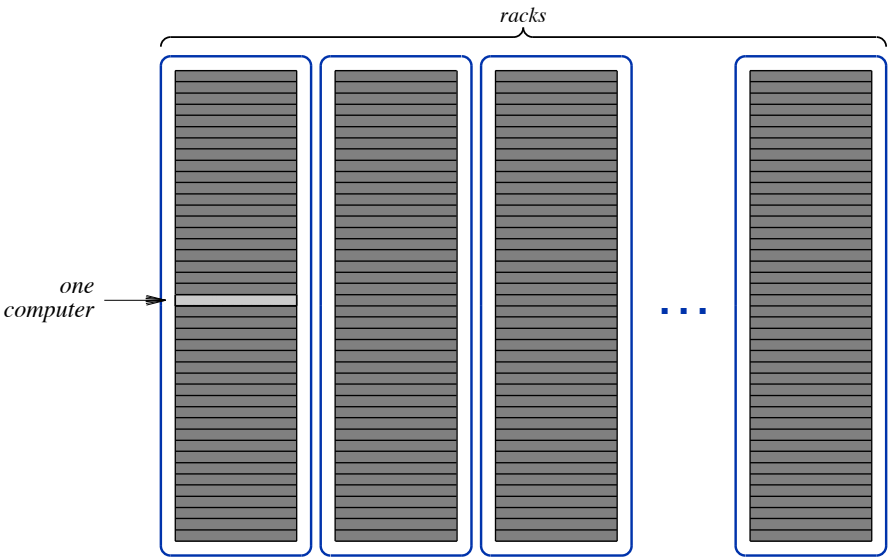


Figure 29.7 Illustration of racks lined up in a data center. Each rack is filled with forty-two pieces of 1U equipment. A single server computer is highlighted.

29.15 Generalized Cloud Computing For An Individual

Most cloud computing focuses on companies, not on individuals. However, it is possible to apply the same idea to individual computing. To use cloud computing, a user's local device only needs two things: remote desktop software that contacts the cloud server and a set of input/output devices, such as a display, tracking device (mouse or trackpad), microphone, speaker, and camera. As Chapter 28 explains, remote desktop means all processing is performed on the cloud server. When the user runs an app, the app runs on the cloud server; when files are saved, the data is stored on the cloud server. The point is:

If an individual uses generalized cloud computing, the user's local device runs remote desktop software, and a cloud server performs computation. Using the cloud eliminates all synchronization problems because a user always sees exactly the same data and apps no matter which device is used to access the cloud.

Imagine how convenient it would be to use a generalized cloud computing service. If a user picks up their smart phone, the user will be able to use the same apps as when they boot a desktop. Differences in screen sizes may mean that items do not appear exactly the same on all devices, but all devices will have the same capabilities and the same data. For example, when the user connects to their cloud server from a desktop computer that has a large screen, the user may see all their icons displayed. When the same user connects to their cloud server from a smart phone, the system may only display some of the icons, and require the user to scroll the screen to see more.

As an example of synchronized data, think of an address book. Once a user enters a friend's name, phone number, and email address, the entry will be available on any of the user's devices. Similarly, if a user saves an attachment from email, the saved file will be visible on all of the user's devices. The point is that a user will not need to remember which data items are stored on which device, nor will the user need to synchronize devices.†

Using generalized cloud computing has another advantage for the average user: a public cloud provider will handle software updates. Of course, a provider charges an additional monthly fee for an update service. When a user signs a contract for service, the user can specify which updates to apply, and the provider will install updates promptly. For example, a user might choose an operating system, and then specify that updates be applied, relieving the user of the task.

29.16 The Disadvantage Of Using The Cloud

From a company's point of view, the primary advantage of cloud computing centers on lower cost, both lower capital cost and lower operational cost. From an individual's point of view, the primary advantages of cloud computing arise from the ability to have a single computing environment available on all devices without the need to worry about synchronization and the ability to have the provider handle software updates. In all cases, however, using the cloud has an important disadvantage: access to computing requires a working Internet connection.

Consider an individual whose entire computing environment depends on continuous cloud access, and imagine the user accessing the Internet through a smart phone. If the user moves out of the range of a cell tower or if congestion on the Internet prevents packets from getting through, all computation will appear to freeze. Because no applications run locally, the user's device will essentially become useless until Internet com-

† Of course, a user will not be able to use the excuse "I'd give you that number, but it's on my other phone."

munication is restored. Imagine a user who happens to be running an app that gives driving directions when the loss of connectivity occurs. We can summarize:

> *Although it has advantages for both companies and individuals, cloud computing has the disadvantage that if a network failure or congestion makes the cloud data center unreachable, access to all computing is cut off.*

Many companies cannot afford to risk being cut off from their cloud computing facilities. Such companies arrange a special, highly reliable connection between the company and their cloud provider's data center. Typically, the company leases a connection from a common carrier (i.e., leases an optical fiber from a router in the company to a router in the provider's data center). Because only data from one company goes across the fiber, traffic from other Internet users cannot cause congestion. Companies that have special requirements for reliability can lease two separate connections between the company and the cloud data center so that if one of them fails, the other will still work.

29.17 Virtualization Technology Used For Cloud Computing

Our description of cloud computing implies that each customer leases physical computers, but that is not the case. To support elastic services, cloud providers use *virtualization*. Virtualization software running on each physical server allows the server to run multiple *virtual machines* (*VMs*). A virtual machine consists of an operating system plus apps. There are two important differences between a VM and a physical computer: multiple VMs can run on a physical computer at one time, and a VM can be *migrated* (i.e., moved) from one physical computer to another. That is, a VM can be stopped, the set of bits that constitute the entire VM can be sent across a network to another physical server, and the VM can then start running again, from where it was stopped.

When a cloud provider leases computers to a customer, the provider leases a set of VMs. If the customer needs additional processing, the provider leases additional VMs. At any time, a provider can migrate VMs across its physical servers to balance the load and avoid having some servers overloaded while others are underutilized.

29.18 Summary

The computing industry has moved from a centralized form of computing, in which an entire organization shared a single, large computer, to a distributed form in which each user has one or more devices. When a user has multiple devices, synchronizing data among them becomes tedious and time-consuming. Vendors offer a variety of cloud services, including data storage and data synchronization services.

Companies have started to move to a generalized cloud computing paradigm in which the company moves its computing to facilities leased from a public cloud provider. The motivation for adopting cloud computing is lower cost; moving to the cloud can lower both operational expenditures (opex) and capital expenditures (capex). If a company has special requirements for data, the company can choose to run a private cloud facility internally; it is also possible to use a hybrid cloud approach in which some of the company's data and processing is moved to a public cloud, and more sensitive data is kept in the company's private cloud.

EXERCISES

29.1 What is the chief advantage of a distributed model in which a user has multiple devices? What is the chief disadvantage?

29.2 Suppose a smart phone provider offers a "cloud synchronization service" for your devices. Explain how the service copies data from one of your devices to another.

29.3 Search the Internet to find a list of public cloud providers. Are you surprised at any of the companies on the list? Explain.

29.4 Search the Internet to find a photo of a cloud data center that shows racks holding computers.

29.5 Suppose a company in New York uses a public cloud provider in California, and places the company web site in the cloud. Explain how packets flow when a customer who lives next-door to the company accesses the company web site.

29.6 Does cloud computing continue the trend toward more distributed computing or represent a move back toward the mainframe model? Explain.

29.7 An IT professional once said that he would never move his company's computing to the cloud until he could find a cloud provider that guaranteed three high-speed connections between the provider's cloud data center and the rest of the Internet. Explain why the professional was so worried about extra connections.

29.8 Suppose you were offered the opportunity to use a public cloud system where all computation was done on a cloud server and all your devices merely used remote desktop software to access the server. Would you choose to use the service or keep computation on your current devices? Explain.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Other Aspects Of Internet Technology

Internet Security And Economics

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

30 Network Security (Encryption And Firewalls)

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

30

Network Security (Encryption And Firewalls)

30.1 Introduction

Previous chapters describe a variety of Internet services and explain how each one works. This chapter and the next two consider the practical matter of *network security*. The chapter begins by considering a fundamental concept: safeguards are needed to make Internet communication and transactions secure. The chapter explains what security means and why it is needed. It then examines two important technologies that help users keep communication confidential and help keep computers safe from unwanted packets. The next chapter continues the discussion by describing ways that attackers fool users into granting them access, and the third chapter on security explains a technology that provides secure, confidential communication between an employee who is traveling and computers in the company for which the employee works.

30.2 Cybercrime And Cyber Security

We use the term *cybercrime* to characterize crimes that involve using digital communication and computational technologies to commit crime. Newspapers frequently contain articles describing cybercrimes. In some cases, perpetrators use the Internet to gain unauthorized access to computers or data in a business. In others, a group will take over hundreds of computers, and then arrange to have the computers bombard a targeted server with so many packets that the server becomes unusable. Sneaking unwanted software onto a computer underlies many security incidents. We use the term

malware to characterize such software. One particularly pernicious form of malware is known as *ransomware*. Once it starts running on a victim's computer, ransomware takes over the computer by blocking the operating system and other apps. The victim is given a choice of paying to remove the ransomware or risking the ransomware erasing all the user's data.

Do cybercrimes occur because the computer and networking industries have failed to make computers and networks safe? Interestingly, security professionals agree that most cybercrimes do not arise from technological weaknesses or from packets that travel across the Internet and invade a computer. Instead, most incidents involve humans. Users may leave their computers and data completely unprotected or may fall prey to scams and trickery that causes them to unwittingly help attackers. Individuals may also succumb to bribes and participate in crime. For example, a famous security incident occurred when criminals broke into the database of a major retail store and stole customers' credit card numbers. It turned out that the criminals did not need to use the Internet to gain access because they bribed an employee who allowed them to enter after hours and access the database. The point is:

> *Most cybercrime does not usually involve geniuses who outsmart security technologies. Instead, most security incidents involve humans who either fall for a scam or decide to abet criminals.*

Although not all cybercrime can be eliminated by using technology, tools exist that can help a user avoid problems. The next sections focus on two key technologies.

30.3 The Unsecure Internet

Many of the networks that constitute the Internet are "shared," which means that multiple computers attach. One of the chief disadvantages of shared networks is a lack of guarantees about *privacy* — an arbitrary computer on the network can eavesdrop on other computers' transmissions. We use the term *unsecure* to characterize such networks. As a whole, the Internet is unsecure because constituent networks may be unsecure.

Chapter 11 describes examples of Internet access technologies, including Wi-Fi used in public areas, such as malls and coffee shops. Most Wi-Fi networks are unsecure in the sense that others are able to "listen in" on conversations. When a device transmits a packet over Wi-Fi, the packet is sent over radio waves. Anyone with a receiver tuned to the appropriate channel can obtain a copy of the packet. More important, no special equipment is needed because the Wi-Fi hardware in most computers already has the necessary capability. Called *promiscuous mode*, and normally only used for network troubleshooting, the feature turns a user's device into a spy system that captures a copy of every packet transmitted within range of the device. That is, a cyber criminal can sit in a coffee shop and run an app that records all the Wi-Fi communication around them without anyone knowing. We can summarize:

Because Wi-Fi uses radio waves to send packets and because the Wi-Fi hardware in most computers can listen to all packets, someone can use a conventional computer to record packets that other users send and receive.

In some cases, lack of Internet security is merely annoying. For example, if a candid conversation between two friends becomes public, statements made in confidence might cause embarrassment. In other cases, however, eavesdropping poses a serious risk. Consider the potential loss that might occur if a third party obtains your credit card number or the password to your bank account.

30.4 Keeping Conversations Confidential

Since ancient times, people have used secret codes to keep messages from being read by outsiders. For example, kings often used coded messages to communicate with their armies. Because each message was written in code, only the sender and recipient could understand the contents. Thus, even if the messenger who carried a message was intercepted, the contents of the message remained safe.

You may enjoy cryptogram puzzles, or may have experimented with coded messages as a child. For example, you might send the following coded message to a friend:

Ij uifsf. Uijt jt b tfdsfu nfttbhf gps zpv.

Your friend knows that the way to decode the message consists of substituting each letter with the previous one in the alphabet. So, *I* becomes *H*, *b* becomes *a*, and so on, resulting in a decoded version of the message.

Hi there. This is a secret message for you.

Modern computer systems use the same basic approach to keep messages private. Before transmitting a message across a network, software on the sending computer encodes the contents of the message. When it arrives on the receiving computer, software decodes the message. Provided the encoding is complex enough, a third party will not be able to decode a message, even if they obtain a copy.

30.5 Computer Encryption And Mathematics

The codes used with modern digital systems differ from the codes used in ancient times because code breaking in the modern world is completely different than code breaking in earlier times. Instead of humans struggling to understand a code, modern codebreakers use computers. A computer can try thousands of combinations per second, and multiple computers can be used at the same time to speed the process.

Thus, to keep a message private, the Internet does not use the same codes that humans use when encoding messages by hand because such schemes are easy for computers to decode. Instead, the Internet uses sophisticated, mathematical encodings that cannot be broken, even when the highest-speed computers are used. We use the term *encryption* to describe the process of transforming a message into a cryptic form that cannot be deciphered by outsiders, and *decryption* to describe the process of transforming an encrypted message back into its original form.

Why does encryption involve mathematics? The answer is simple: inside a digital computer, all information is stored in numbers. Even a sequence of characters such as *abcdef* is represented by numbers. Consequently, encrypting information means manipulating numbers, which involves mathematics.

The mathematicians and computer scientists who study encryption are called *cryptographers*, and the field is known as *cryptography*. The encryption techniques they have produced ensure that outsiders cannot decrypt a message. We can summarize:

> *Data encryption used in the Internet is safe because it uses complex mathematical functions to encrypt data; an outsider cannot decrypt a message, even if the outsider uses many computers.*

30.6 Confidential Web Browsing

How can a user keep their Internet communication confidential? If a URL starts with *https://* instead of *http://*, a browser will encrypt the communication (the "s" specifies using a secure form of http). For example, if a user enters the URL:

https://google.com

the communication with Google will be encrypted. That is, data the user enters on the keyboard is encrypted before being sent to Google, and data arriving from Google is decrypted before being processed by the browser. When communication uses https, a browser displays a closed lock icon, indicating that the connection is secure; the browser displays an open lock icon when communication is not encrypted. Figure 30.1 illustrates encryption and decryption when https is used.

Interestingly, major web sites now use encryption automatically. For example, if a user enters:

http://google.com

The Google web site will instruct the browser to used https instead, and the user will see a new URL displayed.

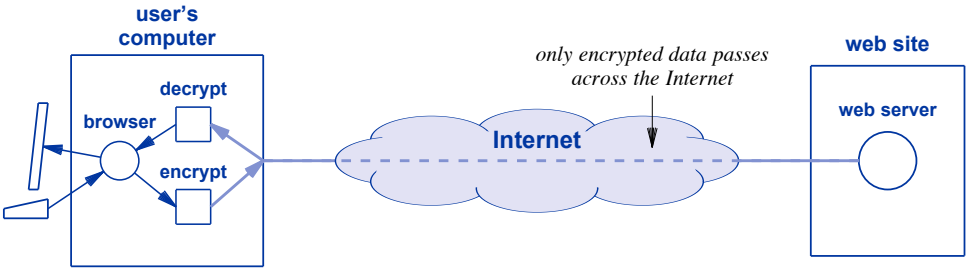


Figure 30.1 Illustration of a browser using https. Data the user enters is encrypted before being sent to the web site, and data the web site sends must be decrypted before the browser can process it.

30.7 No Network Is Absolutely Secure

Does mathematical encryption guarantee absolute security? No. Just as a physical lock cannot provide absolute safety, encryption cannot guarantee confidentiality — if a third party uses enough computers and has enough time, they will be able to break the code and read the message. However, by choosing the encryption method carefully, designers can guarantee that the time required to break the code is so long that the security provided is sufficient. We can summarize:

> *Although no computer network is absolutely secure, modern encryption makes the task of decoding messages so difficult that high-speed computers require years of computing to break a code.*

Keep the principle in mind when thinking about Internet security. When someone asserts that an encryption scheme guarantees security, they mean that although the code can be broken, the effort and time required is great. When someone says that a new encryption scheme is "more secure" than an older scheme, they mean that breaking the new scheme will require a longer time than breaking the old one.

30.8 Encryption Keys

We said that computer software is used to decrypt messages. Suppose someone buys a copy of the decryption software and then obtains a copy of an encrypted message. Will they be able to decrypt the message and understand the contents? No. To understand why, consider an analogous situation in everyday life: although an automobile manufacturer usually makes many copies of each car model, each owner is given a key that only unlocks one vehicle. Therefore, owning a particular car does not mean a person can enter all vehicles of a given model.

Encryption and decryption software uses the same basic idea. Instead of merely encrypting or decrypting a message, the software requires a user to supply a *key* when encrypting or decrypting messages. To "unlock" an encrypted message, an outsider must have the user's key; without the key, one cannot decrypt the message.

Recall that modern encryption schemes use mathematical functions. Consequently, a key consists of a very large number. Apps that choose a key select a number at random, so an outsider will not be able to guess a user's key (and the numbers are so large that an outsider cannot try all possibilities).

30.9 Two Keys Means Never Having To Trust Anyone

The earliest encryption schemes used a *shared key* approach in which the sender and receiver used the same key — the sender used the key to encrypt a message, and the receiver used the same key to decrypt the message. A shared key does not work well for Internet communication. To see why, imagine that you have been issued a key. You can only encrypt messages sent to destinations that have a copy of your key. Before you can encrypt messages sent to Google, you would have to send a copy of your key to Google. The same is true for Facebook, Instagram, YouTube, Amazon, and every other site you visit. Sending a copy of your key would be both tedious and time-consuming.

Ignoring the inconvenience, a shared key scheme has an important weakness: you must trust everyone who has a copy of your key to keep it secret. To understand why lack of trust is an essential ingredient, think of a business with an online catalog. When a customer orders from the catalog, the customer supplies information that must be kept confidential. Of course, major web sites work to keep your key safe, but what about local shops and small startup companies? In many cases, small sites do not have much technical expertise, and may inadvertently expose your key to outsiders. If your key becomes known, you must select a new key. The point is:

> *If a sender and receiver both use the same key to encrypt and decrypt messages, they must trust each other to keep the key secret.*

To avoid having to trust others, cryptographers invented an innovative scheme that has become popular for Internet encryption. Known as *public key encryption*, the scheme assigns each user two keys that are designed to work together. If either key is used to encrypt a message, the other key can be used to decrypt the message. One key is known as the owner's *private key* because it must be kept secret. The second key, which is known as the owner's *public key* can be distributed to anyone. An important mathematical property makes the entire system secure:

The mathematical properties of the keys used in public key encryption are such that knowing someone's public key does not help an outsider guess the private key.

Why has public key encryption become important? It allows an arbitrary person to send a confidential message to an arbitrary recipient, without requiring either party to trust the other party to keep a secret. Each user keeps their private key safe, and never reveals the private key to anyone. For example, suppose a web site uses public key encryption. The site obtains a pair of public and private keys that work together. The site keeps the private keep to itself, but publishes the public key to everyone.† If a user wants to keep their communication with the web site confidential, the user obtains the web site's public key, and uses the key to encrypt messages. Only the web site has the corresponding private key that is needed to decrypt the message.

What about messages sent from the web site back to a user? To keep messages confidential, the web site uses its private key to encrypt the message. The recipient then uses the web site's public key to decrypt the message (remember that public key encryption can work in either direction). Figure 30.2 illustrates the idea.

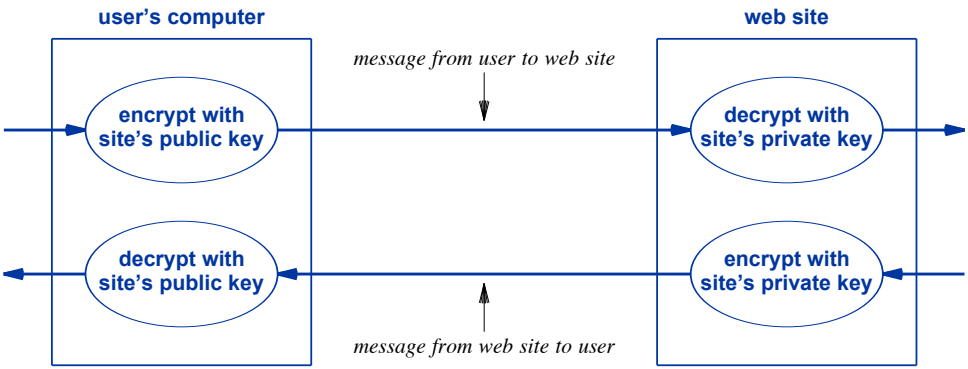


Figure 30.2 Illustration of public key encryption when a user communicates with a web site. Only the public key is known outside the web site.

Each web site has its own pair of public and private keys. A user selects the public key that belongs to the site with which communication is desired. The important point is that public key encryption allows a user to communicate confidentially with an arbitrary web site, without needing to trust others to keep a shared key secret. In other words:

Two keys means never having to trust anyone to keep your secrets.

†Companies exist whose function is to distribute lists of public keys.

30.10 Authentication: User IDs And Passwords

When two parties communicate over the Internet, how can each party know the other is who they claim to be? The question is serious when one party is a bank and the other party claims to be an account holder. The Internet allows any two computers to communicate, so mechanisms are needed that allow an Internet service to know that a communicating party is indeed who they claim to be. We use the term *authentication*, and say that an authentication mechanism provides a way to validate the identity of a communicating party. To summarize:

> *The term* authentication *refers to the process of validating the identity of a communicating party.*

All users are familiar with a common form of authentication: *login IDs* and *passwords*. In essence, a password is a shared secret known to both parties. For example, when a user opens an account at a bank, the bank assigns the user a login ID for the account, and asks the user to choose a password. To access their account, a user must specify the correct login ID and password.

Is sending a password over the Internet safe? We already know the answer: it is only safe if communication is encrypted. Otherwise, eavesdroppers could extract the user ID and password from copies of packets they obtain. We can summarize:

> *Before entering a user ID and password, it is important to check that https is being used and the browser is displaying a closed lock icon to indicate that communication is encrypted.*

30.11 Two-Factor Authentication

Although passwords suffice for most consumer web sites, more sophisticated authentication schemes exist when communicating parties need more security. One mechanism provides an extra check by using a communication channel other than the one on which a password has been entered. Known as *two-factor authentication*, the mechanism arranges to send the user a token that the user must enter along with a password.

The simplest form of two-factor authentication arranges for a web site to use a text message to provide a token. A user begins the login process as usual, by contacting a web site and entering a user ID and password. After it checks the password, the site generates a random string of characters, and sends the generated value to the user's phone in a text message. The site then asks the user to enter the string. Once the user enters the string, login proceeds. Even if someone happens to guess a user's login ID and password, they will not be able to log in because they will not have the user's phone.

30.12 Using Encryption For Authentication

Encryption can also be used to authenticate users. If both communicating parties use public key encryption, they will each have both a public key and a private key. The two can then exchange messages with absolute assurance of the identity of the other party. To send a message, the sender first encrypts the message with the sender's private key. It then encrypts the message again with the recipient's public key, and sends the result. When a message arrives, the recipient first decrypts the message with its private key (only the recipient has the private key, so only the recipient can decrypt it). The recipient then decrypts the message again, using the sender's public key. Only the sender has the sender's private key, so if decrypting the message with the sender's public key results in a valid message, it must have come from the sender. Figure 30.3 illustrates the idea.

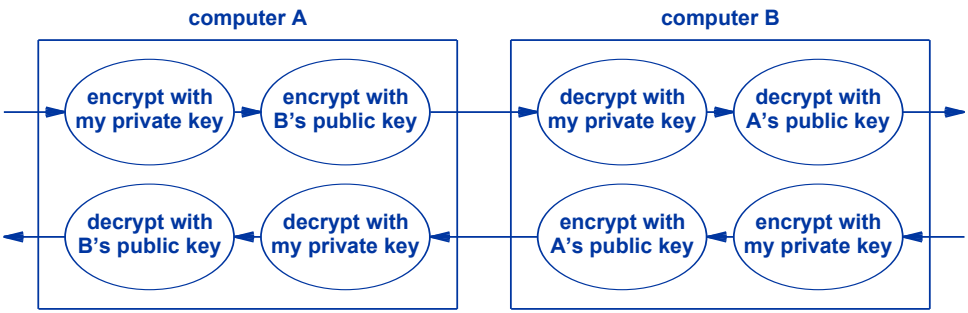


Figure 30.3 Public key encryption used to provide authentication of the sender as well as confidentiality of the message. Each message is encrypted twice.

You do not need to understand the exact details of how or why using encryption twice works. Just appreciate the result: it is possible to send a message across the Internet in a way that the message is confidential (i.e., only the intended recipient can decrypt it) and the authenticity of the sender is guaranteed (i.e., a receiver can know for certain who sent the message).

30.13 Wireless Network Security

As Chapters 17 describes, many homes include a wireless router that uses Wi-Fi to provide Internet access to computers in the home. As we have seen, wireless networks pose a special security threat because they permit eavesdropping. An apartment buildings with many residences in close proximity introduces another problem: even if they do not eavesdrop, neighbors may *piggyback* on a wireless router. That is, they may use their neighbor's wireless router as a way to access the Internet.† The question arises:

†Many ISP service agreements specify that a customer may not use their Internet connection to provide access to others.

how can a wireless router distinguish between the owner's computers, which should be allowed access, and a neighbor's computers, which should be excluded?

There are two ways to exclude outsiders. The first involves hiding the router's SSID. Recall that each wireless router is assigned an SSID, and the router only accepts computers that specify the correct SSID. An owner can choose to hide the SSID or have a router broadcast its SSID periodically. Broadcasting an SSID means every device within range of the router will learn the SSID, and software on the devices will allow users to connect to the network.

The second way to exclude outsiders from a Wi-Fi network consists of using encryption. Wi-Fi encryption has evolved. The original encryption technology was known as *Wired Equivalent Privacy* (*WEP*). Flaws in WEP caused cryptographers to invent *Wi-Fi Protected Access* (*WPA*), which has been replaced by an improved version, *Wi-Fi Protected Access 2* (*WPA2*). Most routers give users a choice of the three standards, and users should choose WPA2.

WPA2 uses a shared key, which means a key must be entered in both the router and devices that connect to the router. If an outsider does not know the key, they will not be able to use the router. Instead of a numeric key, WPA2 allows a user to enter a text string, called a *passphrase*, which is converted to a numeric key. A user should choose a passphrase that is at least sixteen characters long, but one that is easy to remember. For example, a user named *John Doe* who lives at *101 Main Street* might choose the passphrase:

The-101-Doe-Home-Router

Once the correct passphrase has been entered and a device connects to the router, all packets traveling between the device and router are encrypted. Thus, in addition to preventing outsiders from using a router, WPA2 guarantees that no one can eavesdrop on the communication, even if a user fails to use https. To summarize:

Wi-Fi networks offer encryption that keeps communication private and prevents others from using the network. An owner must configure a wireless router to use encryption when the network is installed.

30.14 Network Firewall: Protection From Unwanted Packets

In addition to the security problems described above, a computer or an entire network can be subject to attack from unwanted packets. For example, an attacker can probe a computer to see if the computer has services such as a web server, file sharing server, or a remote desktop server. Once an attacker finds a server running, the attacker can attempt to exploit the server (e.g., guess a login and password).

The chief mechanism used to protect computers from outside attack is known as an *Internet firewall*. The term is taken from physical protection systems where a firewall

consisting of a fire-resistant barrier is placed between two areas to prevent fire from spreading between them. An Internet firewall forms an analogous barrier. That is, an Internet firewall consists of a system that is placed between a computer to be protected and the rest of the Internet; all packets entering or leaving the organization must pass through the firewall. Most large organizations (e.g., companies, schools, hospitals, government sites, and military installations) place a firewall on the link between their site and the Internet to protect all the computers at their site. Figure 30.4 illustrates a firewall used to protect computers in an organization.

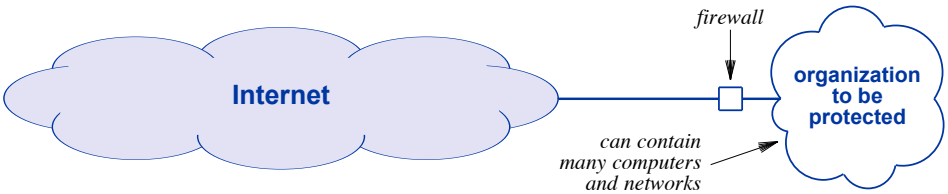


Figure 30.4 Illustration of a firewall used to protect all computers in an organization. A firewall is located on each link between the organization and the Internet.

30.15 Packet Filtering In A Firewall

How does a firewall work? When a firewall is installed, a network administrator configures the firewall according to the desired security policy. The firewall examines each packet, and only allows packets to pass through that satisfy the security constraints. For example, if the site does not have a web site available to outsiders, the firewall will be configured to reject all web requests. Alternatively, a site can configure its firewall to allow web traffic that goes to a specific computer at the site that runs a web server, but forbid web access requests sent to all other computers at the site.

How does a firewall help? A firewall prevents accidental access. For example, if an employee at the company accidentally misconfigures their PC to run a web server, the firewall prevents outsiders from accessing the server. More important, because a firewall is configured to prohibit packets to "unknown" services, the company does not need to worry about a dangerous new Internet service being created — the firewall protects against all access until it is reconfigured to allow access.

30.16 Trojan Horses And Firewall Protection

You may be surprised to learn that a company's firewall does not only restrict incoming packets — most firewalls also restrict outgoing packets as well (i.e., access to Internet sites from inside the company). Why? If it did not restrict access in the re-

verse direction, the company would be vulnerable to a *Trojan horse* attack. How does a Trojan horse arise? The idea is that an attacker tricks an employee inside the company into running an app. In the next chapter, we will learn how attackers trick employees.

Consider an employee's computer. Because the employee's computer attaches to a network inside the company, the computer can reach other computers. The employee may have proprietary files on their computer (e.g., plans for new products or services), may have access to company databases, and may be able to access other computers in the company. A Trojan horse app might be programmed to send a copy of all files on the employee's computer to the attacker. It might access company databases, and send copies of the information to the attacker.

A firewall that restricts outgoing traffic can prevent a Trojan horse app from sending copies of information out to an attacker. Of course, a firewall must allow legitimate access (or employees would not be able to reach common web sites, such as Internet search services). We can summarize:

> *A firewall placed between a company and the Internet restricts access to the company from the outside. To prevent a Trojan horse attack, the firewall also restricts access to the Internet from within the company.*

30.17 Residential And Individual Firewalls

Although we have described using a firewall to protect a company's computers, firewall technology is also available to protect an individual's devices at home. There are two forms:

- A firewall for a residence
- A firewall for an individual device

A firewall for a residence. Most DSL and cable modems have a firewall built into the device, and a user can enable the firewall at no extra charge. Like the firewall that Figure 30.4† illustrates, a firewall in a modem protects all devices at the residence that use the modem. Wireless routers described in Chapter 17 also include a firewall that an owner can enable. The firewall in a wireless router protects all computers that connect to the router.

A firewall for an individual device. As an alternative to a firewall that protects multiple devices, it is possible to configure a firewall that protects a single device. For example, most desktop and laptop computers and smart phones have built-in firewall software that a user can enable. The software examines each packet that arrives or leaves the device, and blocks packets that violate the firewall rules.

Which type of firewall should you use? Both. Enabling the firewall in a cable or DSL modem will help protect all devices that use your Internet connection. Similarly, enabling the firewall on a wireless router protects devices that connect via Wi-Fi. Be-

†Figure 30.4 can be found on page 327.

cause many of a user's devices are portable, however, the devices may connect to networks outside the owner's home. Therefore, it is important to enable the firewall in each device. To summarize:

> *Firewalls can be used to protect devices within a residence. The firewall in a DSL or cable modem can protect devices in the home; enabling the firewall on each device can further protect the device when it connects to a network outside the home.*

30.18 Other Recommended Precautions

We said that using encryption (both using https for web access and using WPA2 on wireless networks) provides confidentiality. We also said that enabling firewalls on modems and individual devices helps prevent unwanted access. A few additional precautions are also recommended:

- Disable or restrict sharing apps
- Disable remote management of wireless routers
- Change default administrator logins and passwords

Disable or restrict sharing apps. One of the ways attackers obtain access to a user's computer or data consists of probing to see if a sharing app is enabled. Disabling sharing apps can prevent such attacks. In cases where the sharing app is needed, it may be possible to restrict access to specific users or specific remote computers.

Disable remote management of wireless routers. Wireless routers offer several ways a user can configure the router. The safest way involves connecting a computer (e.g., a laptop) directly to a port on the router. A wireless router may also allow *remote management* in which a computer connected via Wi-Fi can configure the router. Unfortunately, if remote management remains enabled, an attacker can attempt to gain access by guessing a login and password. Disabling remote management prevents such an attack.

Change default administrator logins and passwords. Modems, wireless routers, and other pieces of network equipment often come from the factory with an administrator login and password preset. Unfortunately, an attacker can discover the administrator login ID and password easily. Therefore, it is a good idea to change both the administrator login ID and password to prevent attacks.

30.19 Summary

Although no network is absolutely secure, encryption technologies exist that provide high levels of assurance against third parties being able to intercept and read messages as they pass across the Internet. Encryption provides the fundamental technology used to make communication confidential. The shared key approach to encryption arranges for both the sender and receiver to have a copy of an encryption key, which they both must keep secret. An innovative type of encryption used widely in the Internet is known as public key encryption. Unlike the shared key approach, public key encryption issues each user two keys: a private key that the user keeps secret, and a public key that the user publishes for others to use. Anyone can use a public key to encrypt a message that only the holder of the corresponding private key can decrypt.

Wireless networks pose special security threats. Encryption mechanisms are available for wireless networks that ensure confidentiality and prevent outsiders from using the network. Currently, WPA2 provides the strongest encryption.

A firewall prevents unwanted traffic from entering a site or a computer; a firewall can also prevent a Trojan Horse app from transmitting outgoing data without the user's knowledge. Most DSL and cable modems include a firewall that a user can enable. Individual devices also include firewall software that can protect the device on any network.

EXERCISES

30.1 Try using the prefix *http://* with several popular web sites. Do they all force the browser to change to *https://*?

30.2 Write down the model number of your modem or wireless router. Then search the web to find the login ID and password that the vendor sets for administrative access.

30.3 A home owner who has two teenage children configures a firewall on the home modem. Will the teenager's smart phones be safe from unwanted packets? (Hint: where do teenagers use smart phones?)

30.4 If you have a wireless router, find out how to enable the firewall.

30.5 A retail web site wants to send a message to a customer in such a way that the customer can be sure the message is not from an imposter. How can the web site use public key encryption to achieve its goal?

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

31

Security Scams: Fooling Users

31.1 Introduction

The previous chapter discusses two key technologies that provide security: encryption that can provide confidentiality and Internet firewalls that can prevent unwanted packets from entering or leaving a site or device. The chapter points out that in many security attacks, the attackers exploit human weakness to gain access.

This chapter continues the discussion by focusing on scams that are designed to trick users into unwittingly helping attackers gain access to computers. The chapter considers a variety of techniques attackers use, and gives guidance on ways to avoid being tricked.

31.2 Traditional Scams And Cybercrime

Dishonesty did not start with the Internet — scams, trickery, and flim-flam have been around as long as humans. In the physical world, scams often involve forged artifacts. At one point, for example, people were tricked into buying land in Florida by sellers who produced fraudulent documents that looked like land deeds. Similarly, scammers have sold forged paintings and fake stock certificates. In most cases, scams prey on greed by guaranteeing a quick way to make money without working hard. For example, criminals who use Ponzi schemes entice additional investors by giving initial investors quick profits. The initial investors tell others, who tell others, until the scam eventually collapses.

How does cybercrime differ from traditional scams? First, in many cyber crimes, no direct human interaction is needed. Consequently, unlike scams that depend on a charismatic individual who charms intended victims, cybercriminals do not need special charm. Second, because the Internet and digital technologies are relatively new, most users do not understand even the basics. Thus, it is easy for cybercriminals to dupe victims by taking advantage of their naiveté. Third, cybercrime spans political boundaries and legal jurisdictions. Consider someone in one country who uses the Internet to break into a computer in another country. If the laws in the two countries differ, the break-in may be a crime in the country where it happened, but the perpetrator may not have committed a crime in their home country. The span of the Internet makes prosecuting such crimes extremely difficult or impossible.

The next sections review some of the scams used on the Internet. The sections give pointers on how to avoid such scams.

31.3 The Foreign Bank Scam

One of the most widely known Internet scams consists of an email message that offers to pay handsomely for help with a financial transaction. The writer claims to be an important person from a foreign country (typically, a prince) who has millions of dollars he wishes to transfer to your country. In some versions of the scam, the victim is asked to send their name and bank account number. In other versions, the victim must make a small payment to enable the transfer. Once a victim pays or sends their bank account information, the scammer takes the money and disappears.

31.4 Phishing

Who would fall for the foreign bank scam? Apparently many people do. Security experts use the term *phishing*† to describe email messages that attempt to lure victims into sending money or revealing personal information that should be kept confidential (e.g., credit card numbers and expiration dates, ATM PINs, and bank account numbers). Other phishing scams claim a distant relative has died and left a large inheritance or say the victim has been selected as a random lottery winner. Unfortunately, the cost to send phishing email is extremely low, so if only a few recipients respond, cybercriminals can make a profit.

The lesson to be learned is:

> *To avoid phishing scams, never respond to random email that solicits money or personal information.*

†The term is pronounced "fishing," and was chosen to imply that the scammers are fishing for personal information.

31.5 The Software Update Scam

One of the most deadly scams involves tricking the user into allowing a cybercriminal to install software on the user's device. The point of the scam is to have the user type their password, giving the cybercriminal complete access to the user's data and device.

One way to trick a user into typing a password involves a fake *software update* message. For example, the user receives email with a link to a web page that contains a news story, cartoon, or other items that might entice the user to view it. As the user starts to view the web page, a window appears to pop up blocking part of the page. The window looks exactly like the vendor's software update window. The message says a new version of the software is ready to be installed, and asks the user to enter their password. Once they have password access, cybercriminals can install *Trojan horse* software that allows the criminal to access your device at any time. The criminal can obtain all the data on your device, including saved passwords. Alternatively, the criminal may use your device at any time to commit cybercrimes (law enforcement will trace the crime back to you instead of the cybercriminal).

Tesla car owners fell prey to a software update scam. A malware app popped up a fake software update screen, asked the user to enter a password, and then took over the vehicle. The lesson to be learned is:

> *To avoid a software update scam, configure your device to notify you when software updates are ready, but do not arrange for software updates to start automatically. If your device has not been configured for automatic updates, any request that asks you to approve installing an update is a scam and can be ignored.*

31.6 Password Change Scam

One form of phishing involves tricking a user into revealing their password. Email arrives that tells a user that their password has expired. The email announces that changing the password will keep the user's account more secure. The scam may specify that the user's bank password has expired; other versions specify the password on the user's device has expired. The email contains a link to a web site. When the user follows the link, the user is asked to enter their old password (for verification) along with a new password. Of course, cybercriminals are only interested in the user's current password, which can be used to gain access.

> *Never trust a random email message that contains a link to reset a password. Instead, login directly to whatever site is mentioned and see if a change is needed.*

31.7 Misleading SSID Scam

Consider how a user connects to a Wi-Fi network. On most devices, the device listens as wireless routers broadcast their SSIDs. The device displays a list of SSIDs, and the user chooses one from the list. Once the user selects an SSID, the user's device connects to the network, and begins to use the network to send packets to Internet sites. An attacker who wants to fool a user can create a Wi-Fi network by running a wireless router. In fact, no extra hardware is needed — an attacker can create a Wi-Fi network merely by running software on their device (e.g., a laptop).

How does an attacker trick a user into selecting their network instead of the intended network? One method involves a slight misspelling or other change to an existing SSID. For example, suppose a coffee shop offers free Wi-Fi using the SSID:

joes-coffee-shop

Because an SSID is case-sensitive, an attacker could choose the SSID:

Joes-Coffee-Shop

An unsuspecting user might easily be tricked into selecting the attacker's network.

The misleading SSID scam becomes especially easy if a business does not choose an obvious SSID. For example, when they first set up free Wi-Fi for customers, Starbucks, Inc. used AT&T as a provider, which used the SSID *attwifi*. Therefore, customers in Starbucks had to know to select *attwifi*. To trick customers into using their networks, attackers merely needed to advertise an SSID that *appeared* to be legitimate, such as *Starbucks*.† The point is:

> *To avoid connecting to an attacker's network, never select an SSID unless you know it is legitimate.*

31.8 Man-In-The-Middle Attacks

What does an attacker do after they trick a victim, and the victim connects to their network? The attacker can try to impersonate web sites that the user visits, and ask the user to log in and give a password. More important, it may be possible for an attacker to mimic the user's requests on the real web site. Doing so means the attacker can obtain a page from the web site, and send the page back to the user. Thus, the user will see pages that appear to be legitimate. Once the attacker has collected personal information, such as a credit card number, the attacker can shut down their Wi-Fi network. A user becomes disconnected from the network, but has no clue that they were the victim of a scam until the attacker uses the stolen information.

†When Google became the provider in some locations, the SSID became **Google starbucks**, which does contain *starbucks*, but still made it easy for attackers to trick victims.

Security professionals use the term *man-in-the-middle* to characterize schemes in which an attacker somehow inserts themselves between two communicating parties. Fortunately, there are ways to avoid some man-in-the-middle attacks. For example, if someone uses the public key of a web site to encrypt communication, an attacker will not be able to decrypt the messages. If a browser detects that the other party is not who they claim to be, the browser will display a message warning the user. The lesson to be learned is:

To avoid man-in-the-middle attacks, encrypt all communications with https, and if a browser warns that a web site does not appear to be legitimate, stop using the connection.

31.9 Misleading Email Addresses And Web Site URLs

Many email providers allow each customer to choose an email address to use when sending and receiving email. As a joke, some customers choose the name of a famous person. Thus, one might receive email that appears to come from the richest man in the world, the chief justice of the Supreme Court, or the president of the United States.

Although a fake email address can be humorous, being able to verify the identity of the sender is a serious issue. An attacker can choose a company name, and write an email message that appears to come from the company. Even if the user checks the email address, they will find the company name embedded in it.

More sophisticated attackers register misleading domain names. For example, suppose the XYZQR Company sends email using the domain name:

xyzqrcompany.com

An attacker might register:

xyzqrecompany.com

and then use it in fake email messages. At first glance, the mail appears to come from the XYZQR company. The point is:

Before acting on a message, check the email address carefully.

Another scam involving variations in domain names involves typing errors. When entering the URL for a web site, a user is likely to make typing mistakes. For example, a user might reverse two letters or accidentally hit a nearby key. To trick a victim into accessing a fake web site, an attacker chooses a target, usually a well-known web site that users are likely to visit. The attacker registers a set of domain names that represent common mistypings of the targeted URL. Then, the attacker sets up web servers for

each of the fake URLs. The servers are arranged to display a page that looks like the target page. If a user accidentally mistypes the URL, instead of reporting an error, the user's browser will contact an attacker's web site, which the user will assume is the real site.

Be careful when entering a URL because typing errors may lead you to an attacker's web site.

31.10 Malware In Email Attachments

One of the most common ways that cybercriminals trick users is by sending email attachments that contain malware. An email attachment can contain a *virus* — a piece of malware that takes over your device and uses your contact list to send copies of itself to your contacts. Often, the attachment will contain funny photos, a short video clip, or other content to keep the user distracted while malware is installed on the user's device. The lesson is:

Never open an email attachment unless you are absolutely sure who the sender is.

31.11 Summary

Many cybercrimes involve tricking users into abetting crime. Phishing scams send email messages that ask users to pay money or reveal personal information, usually with the promise of large financial gain. Software update scams display a fake message that asks the user to enter a password to install a software update. A similar scam sends a message that asks a user to follow a link to change their password. The user must enter their old password for verification, revealing it to the attacker. In an SSID scam, an attacker creates a Wi-Fi network with an SSID that appears to be legitimate; if a user selects the SSID, the user's device will connect to the attacker's network instead of the real network. An attacker can use an email address or web site URL that differs slightly from the original. To avoid such attacks, users must exercise extreme caution.

Chapter Contents

32 Secure Access From A Distance (VPNs)

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

32

Secure Access From A Distance (VPNs)

32.1 Introduction

The previous chapters discuss technologies and mechanisms that make a network secure, and describe ways in which users can be tricked by scams. This chapter continues the topic of secure communication by explaining a security technology that allows an individual located at a remote location to access an organization's network without risk, as if the computer were physically located inside the organization. The chapter explains how such facilities work and how they are used.

32.2 An Employee At A Remote Location

Many companies have employees who work from remote locations. For example, some employees *telecommute* from home. A telecommuter must obtain Internet access from a local ISP, and then use the Internet to connect to their employer. Other employees hold sales jobs that require them to meet with customers at the customers' locations. A sales employee may need to access company documents or facilities. The question arises, how can an employee have secure, safe access to all company facilities from a remote location?

To understand the question, it is important to know that in most organizations, computers attached directly to the corporate network are granted more privilege than computers that access the organization from the outside. The motivation arises from the premise that only employees' computers are attached directly; suppliers, customers, and

others who access the network from the outside are not part of the organization. Thus, the company adopts policies that distinguish between "insiders" and "outsiders." The company creates policies that specify what insiders and outsiders are permitted to access and what each group is prohibited from accessing. The company then deploys technologies that enforce the policies.

It is important to understand that in most cases, an outsider is prohibited from accessing company facilities and services, even if the outsider possesses login information. For example, an outsider may not be able to access the company's employee database to look up employee email addresses and phone numbers, even if the outsider knows a valid login ID and password for the database server. One way to block such access consists of configuring the organization's firewall† to block all packets that arrive from the Internet destined to the organization's employee database server.

32.3 Secure Remote Desktop

Chapter 28 explains remote desktop technology that allows a user at one location to access a computer at another. Chapter 30 explains how encryption keeps Internet communication confidential. The question arises: can the two technologies be combined to provide secure access between an employee at a remote location and a computer inside the organization? The answer is yes. The organization's firewall can be configured to permit such access, and the use of encryption will ensure the communication will remain confidential. Thus, an employee will be able to access a server inside the organization from any location.

Unfortunately, a secure remote desktop system does not solve the problem completely. To see why, consider a salesperson meeting with a customer at the customer's location. Suppose the salesperson uses remote desktop software on their laptop to access a server inside the company and display documents. Now suppose the salesperson decides to place a copy of the document onto a USB thumb drive, which will be given to the customer. Because the salesperson is using remote desktop software, the document is not on the local laptop. More important, if a user plugs a USB device into the laptop, the device does not appear on the remote computer, meaning that apps running on the remote computer cannot place a copy of data on the USB device. The point is:

> *Although remote desktop technology can be used to provide safe access to a server inside the company, apps running on the remote server cannot store data to the user's local device.*

†Chapter 30 explains firewalls.

32.4 Using A Leased Circuit For Secure Telecommuting

Telephone companies provide one way to solve the telecommuting problem: a leased circuit. Recall that a phone company leases digital circuits and a customer can specify two arbitrary geographic locations when leasing a circuit. More important, the phone company guarantees that the circuit will remain *private* (i.e., only the two designated locations will be able to access the data). Networking professionals say that a circuit provides a *private network connection*.

Recall that a digital circuit must be leased from a phone company, which installs the circuit between an employee's residence and the employee's company. The circuit may run in wires along utility poles. A modem attaches to each end of the circuit. At the employee's house, the modem connects to a computer; at the company, the modem connects to the corporate network. In essence, the circuit "extends" the corporate network to the employee's home — the employee's computer has the same privileges as a computer inside the company. Figure 32.1 illustrates the connections:

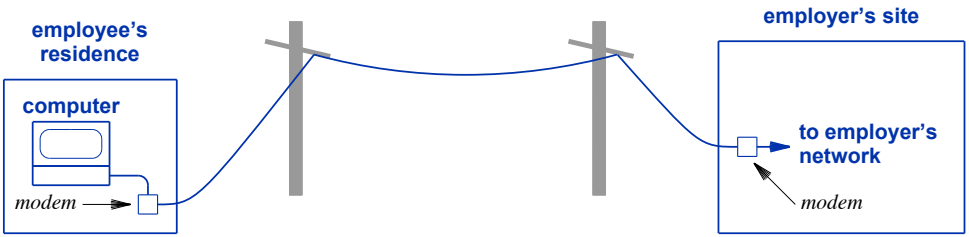


Figure 32.1 Illustration of a leased circuit connecting an employee's residence and the employer's site. The connection is *private* because the phone company guarantees that no outsiders can access the circuit.

Although a leased circuit provides secure telecommuting, the approach has a significant disadvantage: extremely high cost. Even a low-capacity circuit costs much more than a typical Internet connection. In addition, leased circuits do not solve the problem of providing access for employees who travel.

32.5 VPN Technology: Secure, Low-Cost Remote Access

Engineers have created a clever and interesting technology that combines low cost and secure access to allow inexpensive, safe telecommuting. When using the technology, a remote computer is granted full privileges, just as if the computer were present on the company's network. Furthermore, the technology is safe, and can even be used over

wireless networks in a public area such as a hotel room. That is, even if others can eavesdrop, they will not be able to understand the transmissions or learn how to gain access.

The networking industry has given the technology a descriptive name: *Virtual Private Network* (*VPN*). All communication remains *private* because VPN technology ensures that outsiders cannot understand messages. VPN technology is *virtual* because it does not require the installation of physical wires or leased circuits. Instead, all communication travels over the Internet. Thus, to use VPN technology, an employee only needs standard Internet access. To summarize:

> *VPN technology solves the problem of providing safe, low-cost telecommuting by allowing a computer to obtain insider privileges over a conventional Internet connection.*

32.6 VPN From An Employee's Perspective

Interestingly, no special hardware is needed to use a VPN — an employee only needs to load VPN software on their device. In many cases, a company requires the use of VPN technology, and installs VPN software on devices they issue to employees. As an example, we will consider VPN software on a company-issued laptop.

When an employee boots their laptop or changes network connections, VPN software takes over the screen, and asks the employee to enter their company login ID and password. Some companies require the use of two-factor authentication, such as an extra code sent to the employee's phone.† Once the employee has logged in, the employee's desktop appears, and the employee can begin to work.

The laptop has all the privilege of a device connected directly to the employer's network, which means the employee can access corporate servers and other facilities that are only available "inside" the company. Of course, the employee can access web sites, send email, and perform Internet search, just like employees working in their offices.

32.7 How A VPN Works

As one might expect, VPN software communicates with a special *VPN server* at the employer's site. VPN technology uses encryption to ensure that communication remains confidential — all data is encrypted before being sent from the employee's device to the server at the employer's site and all data is encrypted before being sent from the server at the employer's site to the employee's device. Thus, communication between the employee's device and the employer's site remains secure.

Once it is installed, VPN software runs whenever the employee boots the device. More important, device startup cannot be completed until the employee has entered a

†Chapter 30 discusses two-factor authentication.

valid ID and password. Thus, if the employee loses their device, an outsider will not be able to access data on the device.

VPN software works with any Internet connection. If an employee is home, the employee can use a wired connection to their DSL or cable modem, or can use Wi-Fi to connect to a wireless router. When traveling, the employee can connect to any network, including public Wi-Fi networks, such as those in a hotel or coffee shop. Figure 32.2 illustrates an employee using a VPN over a Wi-Fi connection.

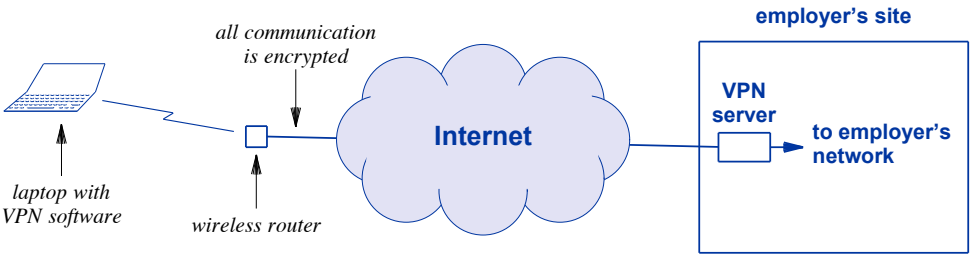


Figure 32.2 Illustration of a laptop with VPN technology that allows an employee to telecommute.

32.8 The Illusion Of A Direct Connection

Although encryption handles the "privacy" aspect of VPN communication, a second technique is used to handle the "virtual" aspect. In essence, VPN software running on an employee's device must create the illusion that the device connects directly to the corporate network. All datagrams sent from and to the computer must have an IP address that is on the corporate network. To achieve the illusion, VPN software controls all communication completely by inserting itself between the Internet software on the device and the network hardware.

Insertion occurs when the device is booted. That is, VPN software runs before the operating system displays a desktop. The VPN software:

1. takes control of the network hardware,
2. obtains a temporary IP address from the local network as usual,
3. prompts the user for a login ID and password,
4. contacts a server to verify that the credentials are valid, and
5. arranges to intercept all outgoing packets.

VPN software sends each outgoing packet to the VPN server, regardless of the packet's destination. The VPN server places the packet on the corporate network as if it were sent by a device inside the company. Routers on the corporate network then forward the packet to its destination. When a packet comes back from the destination, routers on the corporate network forward the packet to the VPN server, which sends the

packet back to the employee's device. VPN software on the employee's device blocks all packets except those that arrive from the VPN server. As a consequence:

> *Packets leaving the employee's device only go the corporate network, and packets entering the employee's device only come from the corporate network, exactly as if the employee's device connects directly to the corporate network.*

Figure 32.3 illustrates the arrangement of Internet software on a device that does not run VPN software and the arrangement when a VPN is used.

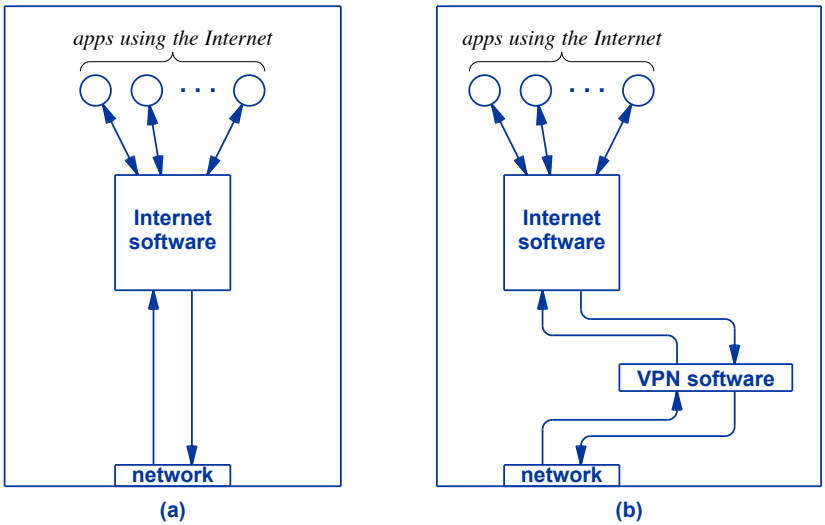


Figure 32.3 (a) Internet software on a device that does not run VPN software, and (b) the arrangement when VPN software is used.

32.9 Obtaining A Corporate IP Address

If a device attaches directly to a corporate network, the device's IP address will belong to the corporate network. To complete the illusion of being directly attached to the corporate network, VPN software on the employee's device must provide a way to obtain a corporate IP address. Therefore, once the VPN software has taken control of the network hardware and arranged to intercept all outgoing packets, it allows the operating system to continue booting. Internet software on the device follows the usual steps to obtain an IP address, completely unaware that the VPN software is capturing each outgoing packet and sending the packet to the VPN server. Therefore, the Internet software:

1. sends a request to obtain an IP address,
2. receives a reply with a temporary IP address, and
3. installs the address and allows apps to use the address.

What IP address does the Internet software on the computer receive? Remember that the VPN software will intercept the address request, and send the packet through the VPN server to the corporate network. When a system on the corporate network assigns an address and replies, the VPN server forwards the reply to the employee's device, and the VPN software passes the reply to the Internet software. Therefore, the Internet software on the user's device will be assigned an IP address on the corporate network!

Only the VPN software can use the local Internet address; all other software on the computer uses an IP address that was obtained from the corporate network. As far as an application program is concerned, the computer appears to connect directly to the corporate network. Outgoing and incoming packets contain the corporate address.

32.10 Exchanging Packets With The VPN Server

Exactly how does a packet travel from VPN software on the employee's device to the VPN server? One method places the packet *inside another packet*, and transfers the result to the VPN server. To understand what a VPN is doing, imagine that you want to exchange letters with a friend, but want to fool your friend into believing you are in Philadelphia when, in fact, you are in Chicago. To succeed in the ruse, your letters must be postmarked from Philadelphia, and the return address must specify Philadelphia. If you have a relative living in Philadelphia, creating the illusion is straightforward. To send a letter, you create a stamped envelope addressed to your friend with a return address that gives the Philadelphia address of your relative. Once an outgoing letter has been created, place the letter inside a larger envelope and address the outer envelope to your relative in Philadelphia. When the outer envelope arrives in Philadelphia, your relative extracts the inner letter and drops it in the mail. Similarly, when your relative receives a reply addressed to you from your friend, your relative places the reply in a larger envelope and sends it to you.

If your relative agrees, you can send letters to many recipients from Philadelphia. Your relative does not need to know in advance the people to whom you will send letters, nor does your relative need to know who will send you letters. Basically, whenever an envelope arrives from you, your relative opens the envelope and mails all the letters found inside. Similarly, whenever a letter arrives addressed to you, your relative places the letter inside an envelope addressed to you in Chicago and forwards it.

The VPN server at the corporation plays the role of your relative in Philadelphia. When it receives a packet sent by VPN software on the employee's device, the VPN server extracts the packet that is inside, and forwards the packet over the corporate network. When a reply comes back, the VPN server places the reply inside a larger packet, and sends the outer packet back to the employee's device. The major difference between letters sent through your relative and packets sent through a VPN server is that

an inner packet is encrypted when it travels between the employee's device and the VPN server. Figure 32.4 lists the steps taken when an employee uses a browser to visit Google.

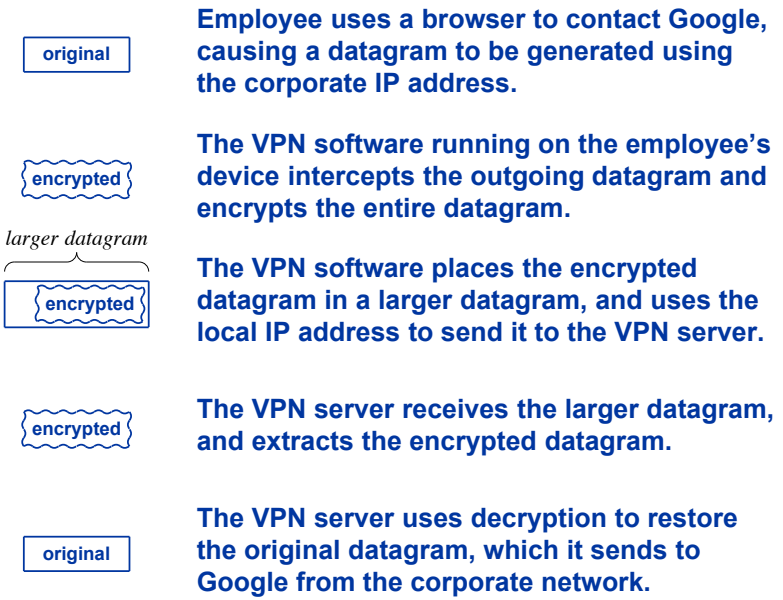


Figure 32.4 The steps taken when an employee who is using a VPN visits Google. The datagram contains a corporate IP address, and is sent from the corporate network, exactly as if the employee's device attached directly to the corporate network.

32.11 The Significance Of VPNs

For many Internet users, VPN technology has revolutionized the way they use the Internet. Business travelers can connect to the corporate network and access all services as if they were local. More important, a traveler does not need to worry about whether access to a network is secure because a VPN provides privacy and prevents unwanted packets from being processed. Thus, a business traveler can use their device in any environment, including customer sites or even a competitor's site, without compromising information or losing privilege.

The point is:

> *An employee can use VPN technology to connect to the corporate office and obtain full privileges from any location. Because it provides secure access, VPN technology changes the way business travelers use the Internet.*

32.12 Summary

Virtual Private Network (VPN) technology provides a way for an employee to obtain privileges on a corporate network as if their device was inside the company and directly attached to the network. No special hardware is needed for a VPN — a conventional device can become a VPN device merely by adding software.

VPN software obtains a local IP address, which is only used to contact the VPN server at the company. Once it has been set up, VPN software intercepts each outgoing packet and sends the packet to the VPN server, and only accepts incoming packets from the VPN server. Only the VPN software knows the local IP address. When Internet software on the employee's device requests an IP address, the request is sent to the corporate network, which means the device will be assigned a corporate IP address. VPNs use encryption to ensure that no outsiders can eavesdrop on communication between the employee's device and the VPN server.

EXERCISES

32.1 If you know someone who travels on business, ask if they use a VPN.

32.2 Suppose an employee from a U.S. company visits China and uses a VPN. If the employee browses a web site in China, where do the packets go?

32.3 If an employee is using a VPN and saves a file on the desktop, will the file be stored on the local device or on a server at the company?

32.4 Answer the previous question for the case where an employee uses remote desktop access instead of a VPN.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter Contents

33 Internet Economics And Electronic Commerce

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

33

Internet Economics And Electronic Commerce

33.1 Introduction

This chapter discusses economic aspects of the Internet. It describes basic terminology, reviews the ISP hierarchy presented in Chapter 10, and considers possible billing schemes. Finally, it explains the relationship between the ISP hierarchy and fees.

33.2 The ISP Hierarchy

Earlier chapters describe the Internet as a set of networks interconnected by routers. In fact, the Internet is not merely a random collection of equipment. Instead, networks and routers that constitute the path between users are owned and operated by Internet Service Providers (*ISPs*).

As Chapter 11 explains, each customer obtains Internet service from one of the ISPs. An ISP uses an access technology to provide wired or wireless communication between the customer's location and the ISP's facilities. For residential customers, wired access technologies include DSL and cable modem technologies, and wireless access includes 4G and 5G cellular technologies. The smallest businesses use the same access technologies as residential customers. Larger businesses often lease dedicated circuits that provide higher capacity than other access technologies.

How does a customer's ISP connect to the rest of the Internet? ISPs are arranged in a hierarchy. Recall from Chapter 10 that ISPs are arranged in a hierarchy with large

Tier 1 ISPs at the core, intermediate *Tier 2 ISPs* at the next level, and *Tier 3* ISPs at the lowest level. Some networking professionals add a fourth tier to the ISP hierarchy to refer to extremely small ISPs that only serve a few customers in a neighborhood. However, we will focus on the three main tiers. Figure 33.1 illustrates how the three tiers form a conceptual hierarchy.

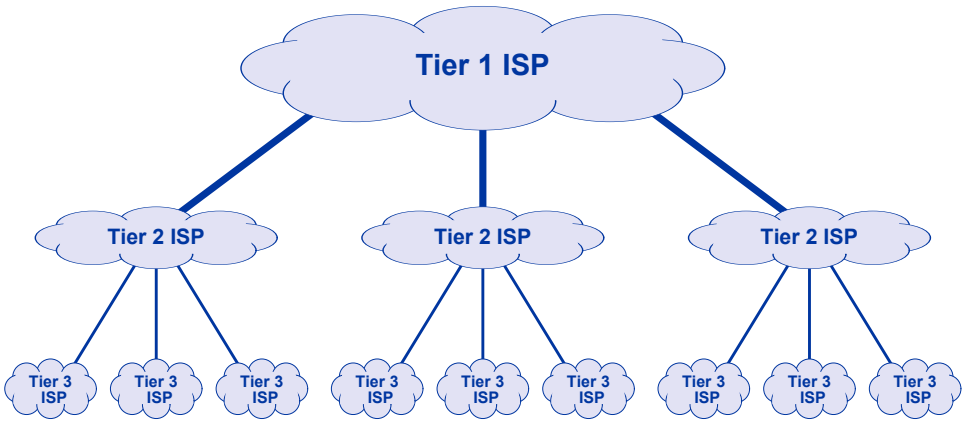


Figure 33.1 Illustrations of the ISP hierarchy with Tier 1 (major) ISPs at the top, Tier 2 (intermediate) ISPs in the middle, and Tier 3 (small) ISPs at the bottom.

Internet traffic follows the hierarchy. When a customer of a Tier 3 ISP communicates with another customer of the same ISP, datagrams stay within the ISP. When a customer of a Tier 3 ISP communicates with a customer of another Tier 3 ISP, datagrams must travel up the hierarchy until they reach a common level. For example, if the sender's Tier 3 ISP and the receiver's Tier 3 ISP both connect to the same Tier 2 ISP, datagrams will travel from the sender's Tier 3 ISPs up to the Tier 2 ISP, down to the receiver's Tier 3 ISP, and then to the receiver.

When a customer connected to a Tier 3 ISP communicates with a customer of the Tier 3 ISP that is farther away, datagrams must travel through a Tier 1 ISP. The datagrams travel from the sender's Tier 3 ISP up to a Tier 2 ISP near the sender, and then up to a Tier 1 ISP. The datagrams then travel down to a Tier 2 ISP near the receiver, down to the receiver's Tier 3 ISP, and finally to the receiver.

In the most extreme cases, the sender and receiver are not reachable from the same Tier 1 ISP. Therefore, before they can be delivered, datagrams must travel up from the sender's Tier 3 ISP, up to a Tier 2 ISP near the sender, up to a Tier 1 ISP, across the connection to another Tier 1 ISP, then down to a Tier 2 ISP near the receiver, down to the receiver's Tier 3 ISP, and finally to the receiver. Now that we understand how ISPs are organized and how datagrams travel through the Internet, we can consider the fees that ISPs charge.

33.3 Network Capacity And Router Hardware

An important difference among the ISP tiers arises from the amount of traffic that each tier handles. A Tier 3 ISP only handles traffic coming from or going to its customers. A Tier 2 ISP handles traffic coming from or going to all the customers of the Tier 3 ISPs that connect directly.

Differences in the amount of traffic require differences in equipment. For example, the routers and network hardware used by a Tier 2 ISP can handle many more packets per second than the routers and network hardware used by a Tier 3 ISP. Tier 1 stands out because the traffic is significantly higher than the traffic in other tiers, and therefore requires extremely high-speed networks and extremely powerful, special-purpose routers.

Interestingly, some large ISPs operate across multiple tiers. For example, a major phone company operates a Tier 1 service that spans continents, a set of Tier 2 services that each span a smaller geographic area, and a set of Tier 3 services that each span a city or part of a city. Similarly, a large cable ISP operates a Tier 1 backbone across the United States, a set of second-level networks in various regions, and Tier 3 services in cities. If the equipment used in the tiers varies, how can a single company offer services across multiple tiers? The answer is that although a single company offers services at all levels, the company divides its internal network into separate tiers, using the highest capacity networks and most powerful routers for the backbone, medium capacity networks and routers for the second tier, and lower capacity networks and routers for the third tier. The point is:

> *Because the networks and routers used in each tier differ from those used in other tiers, a single company that offers ISP services for multiple tiers divides their internal network into separate tiers, each with its own type of hardware.*

33.4 Service Provider Fee Structures

Like any commercial entity, an ISP charges for its services. ISPs have used three types of fee structures to charge subscribers:

- Flat Rate Billing
- Usage Based Billing
- Tiered Flat Rate Billing

Flat rate billing. In the United States, many ISPs offer service that is billed at a *flat rate*. That is, the ISP charges the customer the same fixed rate each month, independent of the number of minutes a customer uses the Internet, the destinations with which the customer communicates, or the amount of data transferred. In return for the

charges, the ISP agrees to forward packets from the customer's computer to destinations on the Internet and from computers on the Internet back to the customer's computer. Flat rate billing means a customer never needs to worry about receiving an unexpectedly high bill because the monthly payment remains constant no matter what the customer accesses. ISPs that offer flat rate billing often advertise "unlimited data" to emphasize that the fee remains the same no matter how much (or how little) data a customer transfers in a given month.

Usage based billing. Early ISPs, especially those that offered dial-up Internet access, based the customer's monthly bill on the number of minutes the customer was connected. Some ISPs counted the number of bytes a customer transferred each month, and then used the count to compute the monthly fee. From a customer's point of view, basing a fee on usage means the monthly amount varies. More important, a customer may not be able to guess how much data a given action will involve (e.g., how much data will be downloaded if a web page contains extensive graphics and animation, or how much data will be transferred when watching a streaming video).

Tiered flat rate billing. As a compromise between flat rate and usage based billing, many ISPs offer a *tiered flat rate service*. The ISP sets a limit on the number of bytes that can be transferred per month, and charges a flat rate for any month in which the customer is under the limit. Usually an ISP has multiple plans, where each plan specifies a limit and a fee. To avoid paying for more than they need, a customer chooses the least expensive plan that satisfies the customer's expected needs. If a customer exceeds their limit during a month, the ISP adds a penalty to the monthly fee. By making the penalty high, an ISP can encourage customers who occasionally exceed their limit to upgrade to a more expensive plan.

33.5 Receiver Pays

Traditional Internet services focus on supplying a user with information. That is, users tend to receive much more data than they send. For example, consider browsing the Web. When a user clicks on a URL, the user's browser sends a short request (a few hundred bytes of data) to a web server. The web server responds by sending the contents of the web page which, with embedded graphics, animations, and ads, can total millions of bytes. The point is that the amount of data a user downloads can be many times larger than the amount of data the user uploads. The imbalance between upload and download has persisted, and even applies to social networking sites where users upload photos and other information. Basically, for every photo a user uploads, the user is likely to view many photos on their friends' pages. Because they are aware of the difference in download and upload sizes, many ISPs have designed their services to focus on high-speed download.

The fee structure for ISPs has followed the premise that users download more data than they upload because more value lies in consuming data than producing it. Therefore, ISPs employ the following rule:

Receiver pays.

The most fundamental question about Internet economics involves the source of revenue for the core of the Internet. Who pays for the expensive infrastructure that Tier 1 and Tier 2 providers need to make global Internet communication possible? Interestingly, the rule that receivers must pay applies up the hierarchy. A Tier 3 must pay the Tier 2 directly up the hierarchy because customers of the Tier 3 will download more data from the Tier 2 than they upload. Similarly, a Tier 2 ISP must pay the Tier 1 directly up the hierarchy because customers below Tier 2 will download more data from the Tier 1 than they generate. We can summarize:

Throughout the Internet an entity that consumes data must pay the ISP that supplied the data. An individual customer pays a Tier 3 provider for service, the Tier 3 ISP pays the Tier 2 provider, and the Tier 2 ISP pays the Tier 1 provider.

33.6 ISP Revenue

Each customer signs a legal contract with an ISP that is known as a *Service Level Agreement* (*SLA*). In addition to stating the fee that a customer must pay, an SLA specifies details about the service that will be provided, such as the capacity of the physical connection between the ISP and customer, and the limit (if any) on the amount of data that can be transferred each month. An SLA may contain separate specifications for each direction. For example, the amount of data that a customer can download (i.e., transfer from Internet sites) each month may differ from the amount the customer can upload (i.e., transfer to Internet sites).

The SLA for a large business is more complex than the SLA for a residential customer, and may include a sliding scale of charges with increasing costs as traffic exceeds prestated thresholds. The SLA for a business may also guarantee a response time for repairs when problems occur, or penalties the ISP must pay the customer if service is down for an extended period. The largest business customers transfer so much data that they often pay for a leased connection directly to a Tier 2 network, bypassing local Tier 3 ISPs. In such cases, the charges include the cost of the leased circuit.

Because it downloads more than it uploads, a Tier 3 ISP must pay the Tier 2 provider above it in the hierarchy. Therefore a Tier 3 provider charges its customers enough to cover the cost. Thus, part of the fee each customer pays to their Tier 3 ISP is used to pay a Tier 2 ISP up the hierarchy. In turn, the Tier 2 ISP must use part of its revenue to pay a Tier 1 ISP for service. The point is:

Although a residential or business subscriber only pays a fee directly to a Tier 3 ISP, the Tier 3 ISP will use part of the fee to pay a Tier 2 for service; the Tier 2 ISP uses part of its revenue to pay a Tier 1 ISP.

33.7 Peering Arrangements Among Tier 1 ISPs

We said that at the core of the Internet, Tier 1 ISPs interconnect. What about fees? Does one of the Tier 1 ISPs become a customer of another? In general, no. Instead, Tier 1 ISPs treat each other as *peers* (i.e., as equals). Consequently, the contract between them is known as a *peering agreement*.

Peering arrangements usually require two Tier 1 ISPs to split the cost of shared facilities (i.e., the equipment and leased circuits used for the interconnection). A peering arrangement also specifies conditional fees for network traffic. If the same amount of traffic passes in each direction between two peers, neither pays a fee for traffic. However, if during a given month one ISP sends much more traffic than the other, the peering agreement follows the principle of requiring a receiver to pay, and specifies that the ISP receiving more traffic must pay a fee. We can summarize:

> *At the center of the Internet, Tier-1 ISPs treat each other as peers, and traffic fees are only assessed if traffic is unequal.*

33.8 Security Technology And E-commerce

Retail purchases by individual consumers account for a significant percentage of e-commerce. The most common form of e-commerce transaction consists of a retail purchase from a catalog. An individual begins by using a web browser to search for an item of interest. To enable e-commerce, the company advertising an item for sale provides a way for the user to purchase the item. When a user is ready to purchase items, the user must enter credit card information or use an alternative payment method.

As Chapter 30 explains, a browser can use encryption to keep payment information confidential. In particular, a browser encrypts communication when the URL begins with the prefix https:// instead of http://. In fact, the security technology built into a browser does more than encrypt messages. Surprisingly, public key encryption is quite general. A browser uses public key encryption (the same technology that it uses to keep messages confidential) to verify authenticity.

Before encryption occurs, the browser must obtain the appropriate encryption key (i.e., the public key of the server), and must verify the authenticity of both the server and the key. All steps are automated — no dialog boxes appear, and the user does not need to enter any of the information. Instead, the browser obtains all information automatically over the Internet.

33.9 Digital Signatures

One mechanism for authentication is known as a *digital signature* because it allows a receiver to know who sent a given electronic document in the same way that a conventional signature allows a receiver to know who sent a written document. The digital signature is created by encrypting information about the document using the sender's private key. Unfortunately, many people are confused by the terminology. A digital signature is an encrypted message; it is *not* merely a scanned version of a conventional signature. The latter would be a *digitized signature*. The chief difference between a digitized signature and a digital signature is that a digitized signature (i.e., a scanned image) can be duplicated, but a digital signature cannot. We will not go into detail about how a digital signature works, but will merely summarize:

> *A digital signature consists of a special form of encrypted message; the encryption technology ensures that a digital signature cannot be forged.*

33.10 Certificates Contain Public Keys

We said that anyone can publish their public key. However, before a browser can verify a digital signature and know which web sites can be trusted, the browser must be absolutely certain it knows the sender's public key. Interestingly, a browser does not ask the user to enter the sender's public key. Instead, trusted companies exist that register public keys and provide them as needed. When a browser needs to obtain or verify a public key, it contacts one of the trusted companies. After the company provides the public key, the browser uses the key to verify authenticity of the original message. Of course, the messages sent between a browser and a trusted company must be secure; once again encryption is used to guarantee security.

Although we said that a browser obtains a public key, the technical term for the information that a browser receives from a trusted source is a *digital certificate*. Each certificate contains a public key plus a digital signature from the trusted company to verify that the message is authentic.

A browser usually obtains digital certificates automatically when they are needed without informing the user. In fact, the only way a user can tell that a browser is obtaining a certificate is to watch the area on the screen where a browser displays its current action. Along with items such as *Contacting host*, a user may see the browser display *Obtaining certificate*.

In a few cases, a browser may not be able to obtain a certificate from a trusted source. In such cases, the browser asks the user whether to proceed with a certificate obtained from another, less trusted source. For situations involving financial records (e.g., submitting credit card information), it is not wise to accept certificates from alternative sources.

33.11 Digital Money

Although there are many details we did not cover, it may appear that encryption technology provides everything needed for consumers to conduct e-commerce: privacy to keep messages confidential, authentication of web sites, and the secure communication a browser requires to obtain digital certificates. According to some users, however, another facility is desirable: digital money.

Credit cards are convenient for making large purchases, but they incur overhead because they require a user to enter the number and expiration date. Furthermore, the user must remember the details of the purchase in case there is a question when the bill arrives at the end of the month. Many groups have proposed an alternative known as *digital money* or *digital cash*. The premise is simple: digital cash should be the electronic equivalent of the cash that people carry to make small purchases and should have less overhead than credit cards.

Several schemes have been proposed; the easiest to understand views digital cash as a form of *debit card*. A user electronically visits their bank to authorize a withdrawal from their account, and places the money in a *digital wallet* (the equivalent of a debit card). The bank returns an ID for the wallet, which the user stores on their computer. To make a purchase, the user specifies their digital wallet ID, and the amount of the purchase is deducted. A user can return to the bank to replenish the amount in their wallet, as needed.

33.12 How Digital Cash Works

Behind the scenes, several steps are required to make digital cash operate. Because a bank cannot transfer real money to an electronic wallet, the bank creates an encrypted message that specifies the bank, an account, and an amount. When a business wants to deduct money, the business must obtain authorization from the user who owns the wallet, send the authorization to the bank, and request a transfer of funds. Such transfers are called *micropayments* to reflect the small amount. All the communication involved in setting up a micropayment must be encrypted to keep it confidential, and encryption must be used to ensure that the store, the wallet, and the purchase are authentic.

Because digital money requires extensive use of encryption, engineering the necessary software is difficult. More important, because a viable system requires merchants, banks, and users to agree on software before they can start using the system, building a new digital cash system is costly. Finally, because digital cash is best suited to small purchases, the profit margin is small. As a result, most companies that have tried to create digital cash services have failed. Nevertheless, a few, such as *Pay Pal*, have succeeded.

33.13 Business And E-commerce

So far, we have described e-commerce from a consumer's point of view — shopping from a retail catalog and buying individual items. How does e-commerce affect businesses? When one business sells to another, e-commerce can replace conventional money transfer mechanisms (e.g., printed checks and bank transfers). We refer to such transactions as *business to business*, abbreviated *B2B*. As with consumer purchases, B2B transactions use encryption technologies to ensure that transactions remain confidential and to guarantee authenticity.

To avoid making many small payments, some businesses use a cumulative payment scheme. In essence, a system is arranged where one business can accumulate daily microcharges without actually transferring money to pay for each charge. At the end of the month, the microcharges are totaled, and a single charge is made (e.g., either by charging a credit card or using an electronic transfer).

33.14 The Controversy Over Net Neutrality

Whenever money is involved, controversy seems to arise, and the Internet is no exception. Governments are considering the issue of taxation and private service providers are considering how to maximize profits. Taxation and regulation are extremely difficult in the electronic domain because communication can cross geopolitical boundaries. For example, when a buyer in one country uses the Internet to contact a seller in another country, which country's regulations and taxes apply?

In the economic realm, a heated controversy has arisen over the method of charging for Internet services. To understand the issue, observe that when the Internet began, phone companies focused on providing voice service using analog equipment. Consequently, phone companies viewed the leases of phone wires as a secondary source of revenue, and established a system in which the cost of a lease depended on the capacity of the circuit — higher capacity costs more. As government allowed more competition in the telephone industry and the cost of digital technology declined, revenues from voice telephone service declined.

By 2005, common carriers faced a difficult challenge: companies like Skype and Vonage began using the Internet to provide voice telephone service.† To recapture revenues, phone companies and major ISPs proposed a new pricing scheme in which the charge to send or receive a given amount of traffic would depend on the company that was sending or receiving as well as the type of traffic being sent. For example, a company such as Google, Vonnage, or Skype would pay more for a given amount of traffic than other companies.

The proposal and surrounding discussions worried consumer advocacy groups for three reasons. First, incremental pricing might mean that only large, profitable companies could afford reasonable Internet service. Startups and small, speciality web sites might become unusable. Second, an ISP might provide better service to its own business customers, meaning that the service a consumer receives would depend on the ISP

†Chapter 26 explains Voice over IP (VoIP).

to which the consumer connects. Third, an ISP might use cost to drive customers away from some content and toward other content (e.g., an ISP might make the content from one source much more expensive than the content from another). Fourth, a business could merely pass along the increased costs to consumers, while carriers and ISPs would receive more money for exactly the same service they already provided.

To prevent ISPs from following what they consider to be unfair practices, consumer groups and others lobbied the government to impose *net neutrality*, a pricing scheme in which costs depend on the volume of traffic rather than the type of traffic, the traffic source, or the traffic destination. We can summarize:

> *The term* net neutrality *refers to a pricing scheme in which the charge for service depends only on the volume of traffic and not the traffic type, source, or destination. Consumers find neutrality desirable; many ISPs prefer an alternative that would increase their revenues.*

33.15 Summary

ISPs are arranged in a hierarchy, with Tier 1 ISPs forming the core of the Internet. A Tier 2 ISP covers a smaller geographic region, and a Tier 3 ISP serves a small area, such as a city. Each tier requires its own type of networks and routers.

ISPs can use flat rate, usage based, or tiered flat rate billing. Independent of the rate structure, ISPs follow a system where the receiver pays. Some of the fee a customer pays to a Tier 3 ISP passes up to the Tier 2 ISP directly up the hierarchy, and then to Tier 1.

Encryption, especially public key encryption, enables secure Internet commerce. A digital signature allows an entity to sign a document in such a way that the signature can be authenticated and not forged. Certificates allow a browser to obtain the public key of a company automatically and safely.

An ongoing controversy has arisen over ISP billing. ISPs would like to charge some customers more than others, and would like to charge based on the content being downloaded. Consumer groups and governments have lobbied for *net neutrality*, a set of policies and rules that form a pricing scheme in which costs depend only on the volume of data rather than the type of traffic, its source, or its destination. The U.S. government enacted, and then withdrew, regulations related to net neutrality.

Chapter Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

34

A Global Digital Library

34.1 Introduction

Previous chapters examine services available on the Internet and show how each can be useful. More important, each chapter explains a basic concept that underlies an Internet service.

This chapter concludes the discussion by considering what the Internet is. The Internet has been called a digital library, and the chapter compares a library to the Internet.

34.2 What Is A Library?

In the ancient world, when illiteracy was widespread, knowledge could only pass from one generation to another through oral communication and mimicry. Because mistakes and omissions were common, folklore was unreliable. The invention of writing changed the situation by making it possible to pass information to succeeding generations accurately and unchanged. Once writing had been invented, libraries were created as a way to collect and preserve documents.

The term library is often associated with a building or part of a building that houses and protects physical copies of books, documents, and maps. However, physical artifacts do not form the basis of a library — they are merely the means to an end. A library's mission lies in the preservation of the information that the documents contain, and finding ways to share that information broadly. We can summarize:

A library is a repository that accumulates and protects human knowledge, and makes the knowledge accessible to successive generations.

34.3 Is The Internet A Digital Library?

The Internet has been characterized as a giant *digital library*. Is the characterization valid? Does the Internet fulfill the mission of a library? We learned that in the strictest sense, the Internet only provides a basic communication facility that allows a computer to send a packet to another computer; all other services must be provided by computers attached to the Internet. Consequently, the Internet itself is not a library. However, we can rephrase the question and ask whether the services available via the Internet constitute a library. To help answer the question, the next sections consider the properties of Internet services and the requirements for a library.

Consider some of the Internet services that previous chapters describe. Services include personal communication provided by social media, electronic mail, and interactive chat; hypermedia browsing and the World Wide Web; automated Internet search provided by search engines; instant messaging; streaming audio and video; file transfer; remote desktop services; and IoT services that allow a user to contact and control devices, such as their appliances or vehicles, from a remote location. Although the example services seem diverse, the list does not include all available Internet facilities or services. For example, *online reference sources* provide online versions of dictionaries and encyclopedias; *language translation services* offer automated translation of text from one language to another; *grocery delivery* services allow a user to enter a list of food items and have them delivered to their residence from a local grocery store; *online dating services* allow individuals to enter profiles and contact potential others who are interested in meeting in person; and *roadmap services* allow a user to enter a destination and receive detailed driving instructions. In fact, so many services exist that an individual would need extensive study just to list all the categories. The point is:

The Internet offers a cornucopia of services; the set is incredibly diverse, and includes services that range from personal communication to global search and control of IoT devices.

34.4 New Services Replace Old Services

Although many services exist, the Internet continues to evolve. As users conceive of new ways to use the Internet, engineers devise new implementations. At any time, some of the most popular services have existed for less than a decade. One cannot appreciate the Internet without understanding that:

As researchers and entrepreneurs discover new ways to store, communicate, reference, access, and use information, new Internet services appear. As users move on to new services, old services eventually disappear.

Does the paradigm of continual change fit the definition of a library? Hardly. A library is designed with a goal of preserving information. For example, consider how a library treats books. Once book sales drop to zero, a bookstore will not bother to restock the book. In contrast, once a library adds a book to its collection, the library works to retain and preserve the book. Internet services resemble a bookstore much more than they resemble a library — most of the information and services used to access the information persist only as long as they remain popular.

An aphorism suggests that once something appears on the Internet, it is never actually forgotten. The aphorism certainly seems accurate for embarrassing mistakes. For example, a video clip from high school might mysteriously reappear years later when the individual applies for a job or runs for public office. It is tempting to imagine a secret storage area somewhere in the Internet that keeps a copy of all data. Of course, no such secret storage area exists. Individuals or groups may choose to post copies of items they find interesting or useful, but some information disappears completely once the owner deletes their copy. Thus, information on the Internet is ephemeral — many discussions last a few minutes, some data lasts for days, and only a few items persist for many years.

34.5 Digital Formats, Standards, And Archival Storage

Another interesting difference between a traditional library and a digital library arises from the way in which information can be represented. In a traditional library, physical artifacts contain symbols from a natural language. A given human may not understand the language being used, but can recognize each of the symbols. In the digital world, all data is represented as a sequence of binary digits, zeros and ones. One cannot know how to group them into items or interpret the meaning unless one knows how the bits were produced.

We use the term *format* to refer to the way information is encoded into binary. A format can be straightforward or complicated. For example, the format of a basic text file can be specified easily by saying that the file is divided into 8-bit *bytes*, and each byte contains a character represented in ASCII.† However, the format used for video is much more complex.

Many formats exist. For example, Chapter 21 mentions the *jpeg* format used for digital images. We use the term *standard data format* to characterize a format such as jpeg for which all the details have been carefully documented and published. Standardized formats foster interoperability. If two computer programs each follow a standard, the data produced by one can be read and processed correctly by the other, and vice versa.

†A table of ASCII values can be found in Figure 6.2 on page 49.

The alternative to a standard data format consists of a *proprietary data format*, known only by a single vendor. The vendor does not reveal the details of the format, and does not allow other vendors to create software that uses the format. Instead, the vendor creates and sells apps that store data using the proprietary format. Consequently, anyone who wants to create or access data that uses the proprietary format must purchase apps from the vendor. For example, a software company that sells word processing apps may choose to store each document in a proprietary format.

Consider the problem of archiving data that has been created using a proprietary format. For example, suppose a library keeps a digital copy of a document that has been created by a word processor that uses a proprietary format. As the years pass, the company that produced the word processing software may cease to exist without ever revealing the details of the format. Therefore, it might not be possible to write new software that can understand the format. Thus, only the original word processing app can read the document. Even if copies of the original app still exist, it may not be possible to run them on a modern computer or a modern operating system. We can summarize.

> *Archival storage of digital documents is surprisingly difficult because a digital document cannot be read unless the format used to create the document is known. Even if an app used to create a proprietary format has been kept along with a document, it may not be possible to run the app many years later.*

34.6 Organizing A Library

Librarians formed an essential component of early libraries — each library needed a librarian who had knowledge of the documents in the library and their location. The librarian chose how to store documents and helped users find documents. Many librarians chose to order documents in the library by the date of acquisition. As libraries grew in size, it became difficult for a librarian to remember all the documents. Finding documents pertinent to a topic became more difficult, and several questions arose. How should the documents in a large library be organized? Does it make sense to place documents in order by date, title, author, or subject matter? Should works of fiction be separated from works of nonfiction? More important, can we replace the librarians who keep knowledge about documents with a system that allows a user to find information on their own? In essence:

> *How can information in a library be organized to make searching easy and fast?*

Recall from Chapter 25 that Melvil Dewey proposed a solution to the problem of organizing a library: order documents by topic. Dewey published a *classification* sys-

tem that assigns a three-digit number to each major category and fractional numbers to subcategories. Dewey's initial proposal consisted of a four-page document with fewer than one thousand categories. The resulting *Dewey Decimal System* was widely adopted, and helped ensure that all libraries ordered their collections the same way.

34.7 The Disadvantage Of Imposing Structure

Classification imposes a *structure* on information to make searching easier. Researchers have investigated the question of whether a better classification scheme can be devised to cover all the information available on the Internet. Unfortunately, a classification scheme that makes some searches easier makes others more difficult. To understand the relationship between classification and searching, consider a trivial example: classifying light bulbs. One possible classification divides light bulbs by technology, and has categories *incandescent*, *fluorescent*, *LED*, *halogen*, and so on. Another possible classification divides light bulbs by the environment in which they can be used, and has categories *indoor*, *outdoor*, and *indoor/outdoor*. If we choose to classify by technology, finding information about halogen light bulbs is trivial because all items related to halogen bulbs will be grouped together. However, finding all information about outdoor light bulbs will be difficult because it will require someone to search documents in the incandescent category, then search documents in the fluorescent category, and so on. The point is:

> *No classification scheme is perfect because each classification scheme makes some searches easy and others more difficult.*

34.8 Searching An Unstructured Collection

In some ways, the Internet completely reverses the approach taken by traditional libraries. To make it easier to search by topic, a traditional library uses a classification scheme to impose a structure, and then orders items in the library according to the classification scheme. The Internet avoids classification and structure, and instead analyzes unstructured documents, generates a list of keywords that identify each document, and then allows users to search by keyword.

Which approach works better? As we have seen, whatever classification a library chooses makes some searches easy, but makes others difficult. The Internet approach of searching unstructured documents offers a much richer search mechanism than a conventional library, and becomes necessary when the collection of documents is extremely large. As Chapter 25 points out, Internet search has a slight drawback because automatic indexing does not understand semantics. For example, consider a document that contains the sentence:

This document has nothing to do with Gothic architecture.

Such a document may be selected in response to a query for Gothic architecture because keyword matching does not take into account the meaning of the phrase *nothing to do with*. We say that an Internet search may report *false positives*.

Despite a few false positives, Internet search has turned out to be extremely accurate and handles arbitrary, unstructured documents. The approach also permits arbitrary queries to be answered quickly, which makes it superior to a traditional classification scheme that makes some searches easy and others difficult.

34.9 What Is The Internet?

We began with the question, "What is the Internet?" If it does not fulfill the entire mission of a library, we cannot conclude that the Internet is simply a global digital library. One answer might be that the Internet is a new form that combines some features of a digital library, some aspects of a newsstand, and other features and services.

> *The Internet is a wildly successful, rapidly growing, global, digital information system built on a remarkably flexible communication technology. The Internet includes a variety of services used to create, browse, access, search, view, and share information on a diverse set of topics. In addition, information that is accessible over the Internet includes audio and video that can be gathered, communicated, and delivered live, without being stored.*

34.10 A Personal Note

The Internet will affect your life in some way every day. When it does, think of what you learned from this book. When you access a service, imagine a distant server connected to the Internet, with packets flowing between you and the server. When you see a URL that begins with *https:*, feel confident that encryption is being used to keep your communication secure. When you see an ad for services that run *in the cloud*, smile and remember that the services are not really inside the Internet, but instead run in computers connected to the Internet. When an ISP offers a higher speed Internet, remind yourself that they are only offering to increase the capacity of the connection between you and the ISP. When you use your smart phone to connect to an IoT device, remember that it would not be possible without the Internet. In the end, to be really impressed with the Internet, imagine the world before the Internet gave us instantaneous, inexpensive communication and global access to information.

Index

D

E

L

M

N

O

P

R

S

T

U

V

W

Y