Computer Networks and Internets

SIXTH EDITION

Douglas E. Comer

ALWAYS LEARNING

PEARSON

Computer Networks and Internets

Computer Networks and Internets

Sixth Edition

Global Edition

DOUGLAS E. COMER
Department of Computer Sciences
Purdue University
West Lafayette, IN 47907

PEARSON

Boston Columbus Indianapolis New York San Francisco Hoboken
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editorial Director, Engineering
and Computer Science: *Marcia J. Horton*
Acquisitions Editor: *Matt Goldstein*
Editorial Assistant: *Jenah Blitz-Stoehr*
Marketing Manager: *Yez Alaya*n
Marketing Assistant: *Jon Bryant*
Senior Managing Editor: *Scott Disanno*
Operations Specialist: *Linda Sager*

Media Editor: *Renata Butera*
Head of Learning Asset Acquisition,
Global Edition: *Laura Dent*
Assistant Acquisitions Editor, Global Edition: *Aditee Agarwal*
Senior Manufacturing Controller, Global Edition:
Trudy Kimber
Project Editor, Global Edition: *Aaditya Bugga*

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England
and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsonglobaleditions.com

Authorized adaptation from the United States edition, entitled Computer Networks and Internets,6th edition, ISBN 978-0-13-358793-7, by Douglas E. Comer, published by Pearson Education © 2015.

ISBN 10: 1-292-06117-0
ISBN 13: 978-1-292-06117-7 (Print)
ISBN 13: 978-1-292-06182-5 (PDF)

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4 3 2 1
19 18 17 16 15

Printed and bound by Courier Westford in the United States of America.

PART II Data Communication Basics

Chapter 7 Transmission Media 147

Chapter 8 Reliability And Channel Coding 169

Chapter 9 Transmission Modes 187

Chapter 10 Modulation And Modems 199

Chapter 11 Multiplexing And Demultiplexing (Channelization) 215

PART II

Data Communications

The basics of media, encoding, transmission, modulation, multiplexing, connections, and remote access

Chapters

Chapter Contents

5

Overview Of Data Communications

5.1 Introduction

The first part of the text discusses network programming and reviews Internet applications. The chapter on socket programming explains the API that operating systems provide to application software, and shows that a programmer can create applications that use the Internet without understanding the underlying mechanisms. In the remainder of the text, we will learn about the complex protocols and technologies that support communication, and see that understanding the complexity can help programmers write better code.

This part of the text explores the transmission of information across physical media, such as wires, optical fibers, and radio waves. We will see that although the details vary, basic ideas about information and communication apply to all forms of transmission. We will understand that data communications provides conceptual and analytical tools that offer a unified explanation of how communications systems operate. More important, data communications tells us what transfers are theoretically possible as well as how the reality of the physical world limits practical transmission systems.

This chapter provides an overview of data communications and explains how the conceptual pieces form a complete communications system. Successive chapters each explain one concept in detail.

5.2 The Essence Of Data Communications

What does data communications entail? As Figure 5.1 illustrates, the subject involves a combination of ideas and approaches from three disciplines.

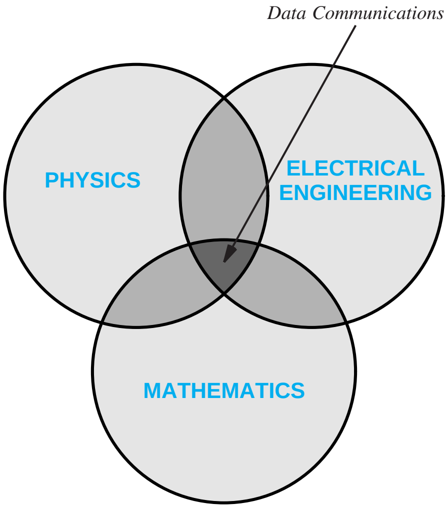


Figure 5.1 The subject of data communications lies at the intersection of Physics, Mathematics, and Electrical Engineering.

Because it involves the transmission of information over physical media, data communications touches on physics. The subject draws on ideas about electric current, light, radio waves, and other forms of electromagnetic radiation. Because information is digitized and digital data is transmitted, data communications uses mathematics and includes mathematical theories and various forms of analysis. Finally, because the ultimate goal is to develop practical ways to design and build transmission systems, data communications focuses on developing techniques that electrical engineers can use. The point is:

> *Although it includes concepts from physics and mathematics, data communications does not merely offer abstract theories. Instead, data communications provides a foundation that is used to construct practical communications systems.*

5.3 Motivation And Scope Of The Subject

Three main ideas provide much of the motivation for data communications and help define the scope.

- The sources of information can be of arbitrary types
- Transmission uses a physical system
- Multiple sources of information can share the underlying medium

The first point is especially relevant considering the popularity of multimedia applications: information is not restricted to bits that have been stored in a computer. Instead, information can also be derived from the physical world, including audio from a microphone and video from a camera. Thus, it is important to understand the possible sources and forms of information and the ways that one form can be transformed into another.

The second point suggests that we must use natural phenomena, such as electricity and electromagnetic radiation, to transmit information. Thus, it is important to understand the types of media that are available and the properties of each. Furthermore, we must understand how physical phenomena can be used to transmit information over each medium, and the relationship between data communications and the underlying transmission. Finally, we must understand the limits of physical systems, the problems that can arise during transmission, and techniques that can be used to detect or solve the problems.

The third point suggests that sharing is fundamental. Indeed, we will see that sharing plays a fundamental role in computer networking. That is, a computer network usually permits multiple pairs of communicating entities to communicate over a given physical medium. Thus, it is important to understand the possible ways underlying facilities can be shared, the advantages and disadvantages of each, and the resulting modes of communication.

5.4 The Conceptual Pieces Of A Communications System

To understand data communications, imagine a working communications system that accommodates multiple sources of information, and allows each source to send to a separate destination. It may seem that communication in such a system is straightforward. Each source needs a mechanism to gather the information, prepare the information for transmission, and transmit the information across the shared physical medium. Similarly, a mechanism is needed that extracts the information for the destination and delivers the information. Figure 5.2 illustrates the simplistic view.

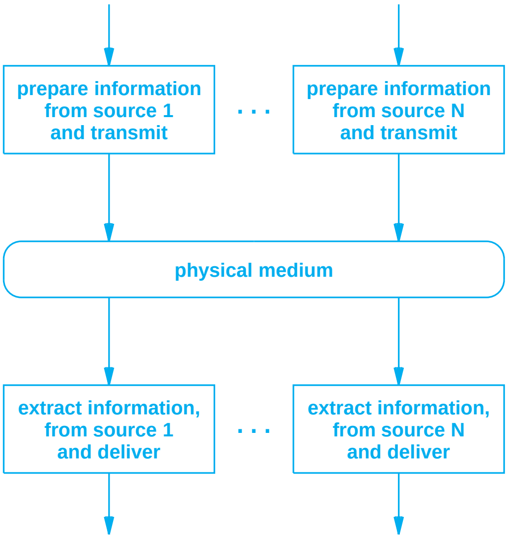

prepare information from source 1 and transmit
. . .
prepare information from source N and transmit
physical medium
extract information, from source 1 and deliver
. . .
extract information, from source N and deliver

Figure 5.2 A simplistic view of data communications with a set of sources sending to a set of destinations across a shared medium.

In practice, data communications is much more complex than the simplistic diagram in Figure 5.2 suggests. Because information can arrive from many types of sources, the techniques used to handle sources vary. Before it can be sent, information must be digitized, and extra data must be added to protect against errors. If privacy is a concern, the information may need to be encrypted. To send multiple streams of information across a shared communication mechanism, the information from each source must be identified, and data from all the sources must be intermixed for transmission. Thus, a mechanism is needed to identify each source, and guarantee that the information from one source is not inadvertently confused with information from another source.

To explain the major aspects of data communications, engineers have derived a conceptual framework that shows how each subtopic fits into a communications system. The idea is that each item in the framework can be studied independently, and once all pieces have been examined, the entire subject will be understood. Figure 5.3 illustrates the framework, and shows how the conceptual aspects fit into the overall organization of a data communications system.

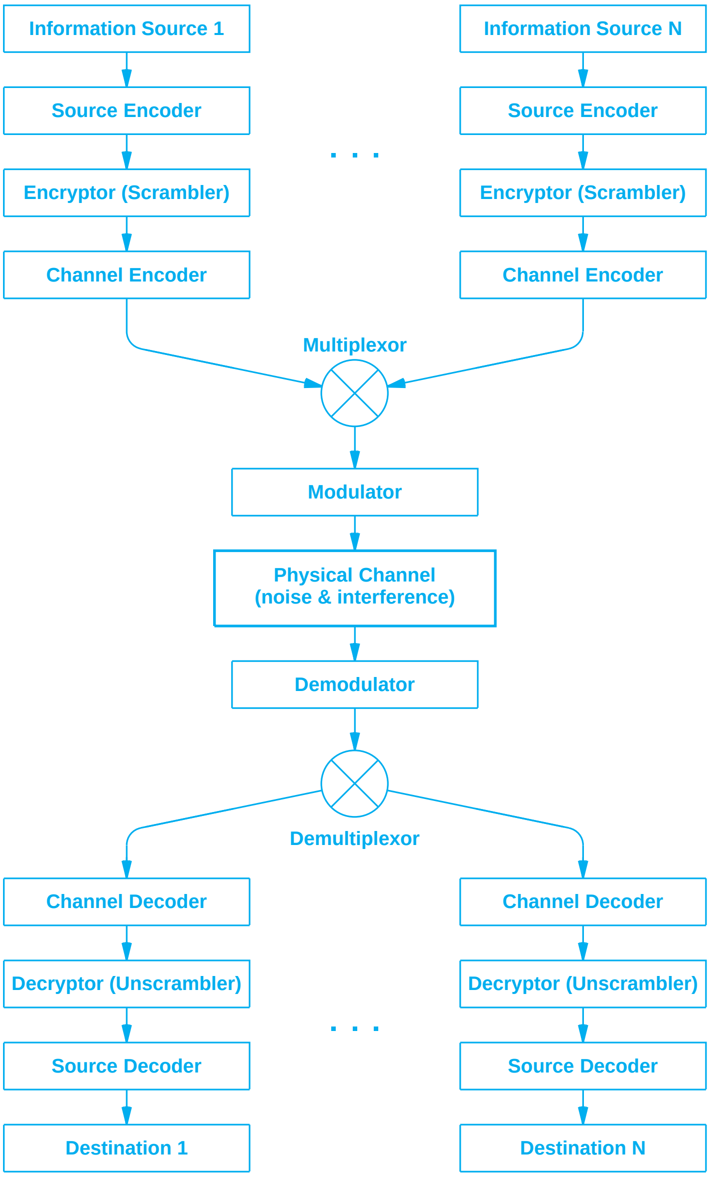


Figure 5.3 A conceptual framework for a data communications system. Multiple sources send to multiple destinations through an underlying physical channel.

5.5 The Subtopics Of Data Communications

Each of the boxes in Figure 5.3 corresponds to one subtopic of data communications. The following paragraphs explain the terminology. Successive chapters each examine one of the conceptual subtopics.

- *Information Sources*. A source of information can be either analog or digital. Important concepts include characteristics of signals, such as amplitude, frequency, and phase. Classification is either periodic (occurring regularly) or aperiodic (occurring irregularly). In addition, the subtopic focuses on the conversion between analog and digital representations of information.

- *Source Encoder and Decoder*. Once information has been digitized, digital representations can be transformed and converted. Important concepts include data compression and its consequences for communications.

- *Encryptor and Decryptor*. To protect information and keep it confidential, the information can be encrypted (i.e., scrambled) before transmission and decrypted upon reception. Important concepts include cryptographic techniques and algorithms.

- *Channel Encoder and Decoder*. Channel coding is used to detect and correct transmission errors. Important topics include methods to detect and limit errors, and practical techniques like parity checking, checksums, and cyclic redundancy codes that are employed in computer networks.

- *Multiplexor and Demultiplexor*. Multiplexing refers to the way information from multiple sources is combined for transmission across a shared medium. Important concepts include techniques for simultaneous sharing as well techniques that allow sources to take turns when using the medium.

- *Modulator and Demodulator*. Modulation refers to the way electromagnetic radiation is used to send information. Concepts include both analog and digital modulation schemes, and devices known as modems that perform the modulation and demodulation.

- *Physical Channel and Transmission*. The subtopic includes transmission media and transmission modes. Important concepts include bandwidth, electrical noise and interference, and channel capacity, as well as transmission modes, such as serial and parallel.

5.6 Summary

Because it deals with transmission across physical media and digital information, data communications draws on physics and mathematics. The focus is on techniques that allow Electrical Engineers to design practical communication mechanisms.

To simplify understanding, engineers have devised a conceptual framework for data communications systems. The framework divides the entire subject into a set of subtopics. Each of the successive chapters in this part of the text discusses one of the subtopics.

EXERCISES

5.1 What are the motivations for data communications?

5.2 What three disciplines are involved in data communications?

5.3 Which piece of a data communications system handles analog input?

5.4 Which piece of a data communications system prevents transmission errors from corrupting data?

5.5 What are the conceptual pieces of a data communications system?

Chapter Contents

6

Information Sources And Signals

6.1 Introduction

The previous chapter provides an overview of data communications, the foundation of all networking. The chapter introduces the topic, gives a conceptual framework for data communications, identifies the important aspects, and explains how the aspects fit together. The chapter also gives a brief description of each conceptual piece.

This chapter begins an exploration of data communications in more detail. The chapter examines the topics of information sources and the characteristics of the signals that carry information. Successive chapters continue the exploration of data communications by explaining additional aspects of the subject.

6.2 Information Sources

Recall that a communications system accepts input from one or more *sources* and delivers the information from a given source to a specified *destination*. For a network, such as the global Internet, the source and destination of information are a pair of application programs that generate and consume data. However, data communications theory concentrates on low-level communications systems, and applies to arbitrary sources of information. For example, in addition to conventional computer peripherals such as keyboards and mice, information sources can include microphones, video cameras, sensors, and measuring devices, such as thermometers and scales. Similarly, destinations can include audio output devices such as earphones and loud speakers as well as devices such as radios (e.g., a Wi-Fi radio) or electric motors. The point is:

> *Throughout the study of data communications, it is important to remember that the source of information can be arbitrary and includes devices other than computers.*

6.3 Analog And Digital Signals

Data communications deals with two types of information: analog and digital. An analog signal is characterized by a continuous mathematical function — when the input changes from one value to the next, it does so by moving through all possible intermediate values. In contrast, a digital signal has a fixed set of valid levels, and each change consists of an instantaneous move from one valid level to another. Figure 6.1 illustrates the concept by showing examples of how the signals from an analog source and a digital source vary over time. In the figure, the analog signal might result if one measured the output of a microphone, and the digital signal might result if one measured the output of a computer keyboard.

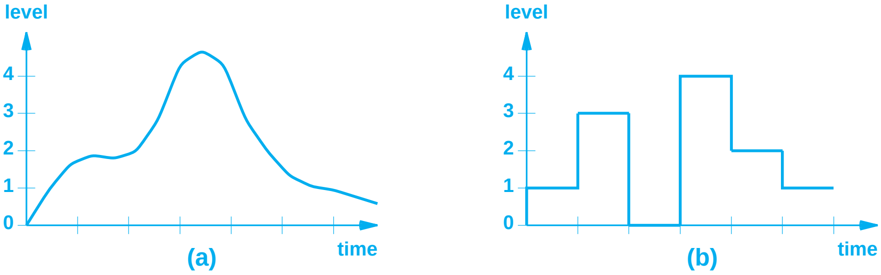


Figure 6.1 Illustration of (a) an analog signal, and (b) a digital signal.

6.4 Periodic And Aperiodic Signals

Signals are broadly classified as *periodic* if they exhibit repetition or *aperiodic* (sometimes called *nonperiodic*), if they do not. For example, the analog signal in Figure 6.1(a) is aperiodic over the time interval shown because the signal does not repeat. Figure 6.2 illustrates a signal that is periodic (i.e., repeating).

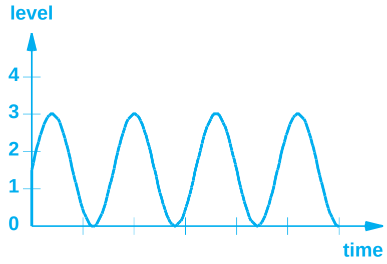


Figure 6.2 A periodic signal repeats.

6.5 Sine Waves And Signal Characteristics

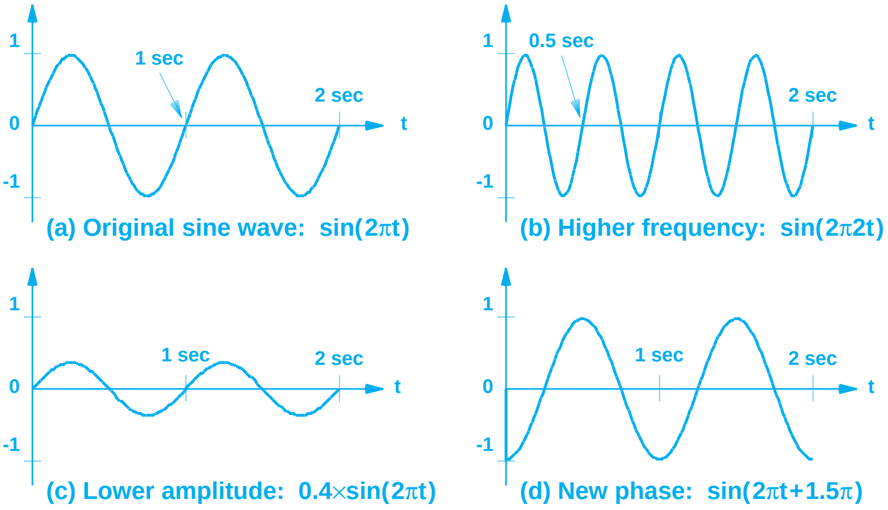
We will see that much of the analysis in data communications involves the use of sinusoidal trigonometric functions, especially *sine*, which is usually abbreviated *sin*. Sine waves are especially important in information sources because natural phenomena produce sine waves. For example, when a microphone picks up an audible tone, the output is a sine wave. Similarly, electromagnetic radiation can be represented as a sine wave. We will specifically be interested in sine waves that correspond to a signal that oscillates in time, such as the wave that Figure 6.2 illustrates. The point is:

> *Sine waves are fundamental to input processing because many natural phenomena produce a signal that corresponds to a sine wave as a function of time.*

There are four important characteristics of signals that relate to sine waves:

- Frequency: the number of oscillations per unit time (usually seconds)
- Amplitude: the difference between the maximum and minimum signal heights
- Phase: how far the start of the sine wave is shifted from a reference time
- Wavelength: the length of a cycle as a signal propagates across a medium

Wavelength is determined by the speed with which a signal propagates (i.e., is a function of the underlying medium). A mathematical expression can be used to specify the other three characteristics. Amplitude is easiest to understand. Recall that $sin(\omega t)$ produces values between -1 to $+1$, and has an amplitude of 1. If the sin function is multiplied by A, the amplitude of the resulting wave is A. Mathematically, the phase is an offset added to t that shifts the sine wave to the right or left along the x-axis. Thus, $sin(\omega t + \phi)$ has a phase of ϕ. The frequency of a signal is measured in the number of sine wave cycles per second, *Hertz*. A complete sine wave requires 2π radians. Therefore, if t is a time in seconds and $\omega = 2\pi$, $sin(\omega t)$ has a frequency of 1 Hertz. Figure 6.3 illustrates the three mathematical characteristics.



(a) Original sine wave: sin($2\pi t$)

(b) Higher frequency: sin($2\pi 2t$)

(c) Lower amplitude: $0.4\times\sin(2\pi t)$

(d) New phase: $\sin(2\pi t+1.5\pi)$

Figure 6.3 Illustration of frequency, amplitude, and phase characteristics.

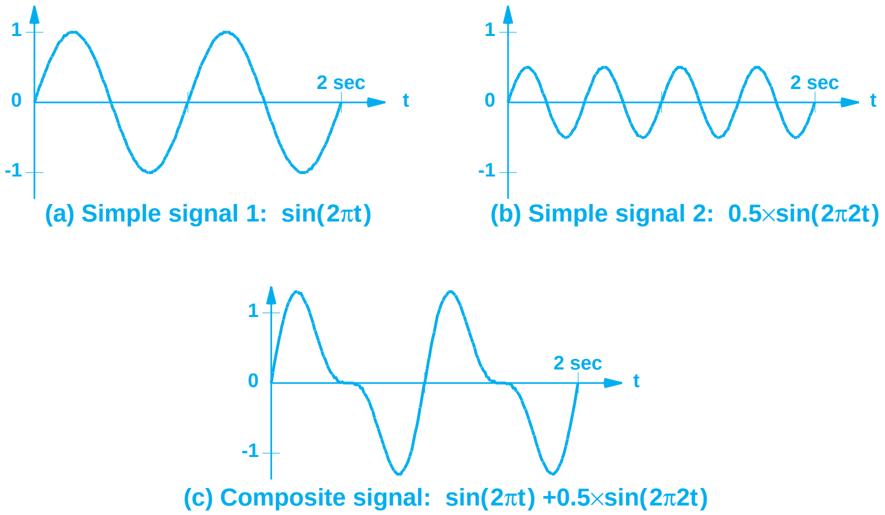
The frequency can be calculated as the inverse of the time required for one cycle, which is known as the *period*. The example sine wave in Figure 6.3(a) has a period of $T=1$ seconds, and a frequency of 1 / T or 1 Hertz. The example in Figure 6.3(b) has a period of $T=0.5$ seconds, so its frequency is 2 Hertz; both are considered extremely *low* frequencies. Typical communication systems use *high* frequencies, often measured in millions of cycles per second. To clarify high frequencies, engineers express time in fractions of a second or express frequency in units such as *megahertz*. Figure 6.4 lists time units and common prefixes used with frequency.

Time Unit	Value	Frequency Unit	Value
Seconds (s)	10^{0} seconds	Hertz (Hz)	10^{0} Hz
Milliseconds (ms)	10^{-3} seconds	Kilohertz (KHz)	10^{3} Hz
Microseconds (μs)	10^{-6} seconds	Megahertz (MHz)	10^{6} Hz
Nanoseconds (ns)	10^{-9} seconds	Gigahertz (GHz)	10^{9} Hz
Picoseconds (ps)	10^{-12} seconds	Terahertz (THz)	10^{12} Hz

Figure 6.4 Prefixes and abbreviations for units of time and frequency.

6.6 Composite Signals

Signals like the ones illustrated in Figure 6.3 are classified as *simple* because they consist of a single sine wave that cannot be decomposed further. In practice, most signals are classified as *composite* because the signal can be decomposed into a set of simple sine waves. For example, Figure 6.5 illustrates a composite signal formed by adding two simple sine waves.



(a) Simple signal 1: sin(2πt)

(b) Simple signal 2: 0.5×sin(2π2t)

(c) Composite signal: sin(2πt) +0.5×sin(2π2t)

Figure 6.5 Illustration of a composite signal formed from two simple signals.

6.7 The Importance Of Composite Signals And Sine Functions

Why is data communications centered on sine functions and composite signals? When we discuss modulation and demodulation, we will understand one of the primary reasons: the signals that result from modulation are usually composite signals. For now, it is only important to understand the motivation:

- Modulation usually forms a composite signal.
- A mathematician named Fourier discovered that it is possible to decompose a composite signal into its constituent parts, a set of sine functions, each with a frequency, amplitude, and phase.

The analysis by Fourier shows that if the composite signal is periodic, the constituent parts will also be periodic. Thus, we will see that most data communications systems use composite signals to carry information: a composite signal is created at the sending end, and the receiver decomposes the signal into the original simple components. The point is:

> *A mathematical method discovered by Fourier allows a receiver to decompose a composite signal into constituent parts.*

6.8 Time And Frequency Domain Representations

Because they are fundamental, composite signals have been studied extensively, and several methods have been invented to represent them. We have already seen one representation in previous figures: a graph of a signal as a function of time. Engineers say that such a graph represents the signal in the *time domain*.

The chief alternative to a time domain representation is known as a *frequency domain* representation. A frequency domain graph shows a set of simple sine waves that constitute a composite function. The y-axis gives the amplitude, and the x-axis gives the frequency. Thus, the function $A\sin(2\pi t)$ is represented by a single line of height A that is positioned at x=t. For example, the frequency domain graph in Figure 6.6 represents a composite from Figure 6.5(c)†.

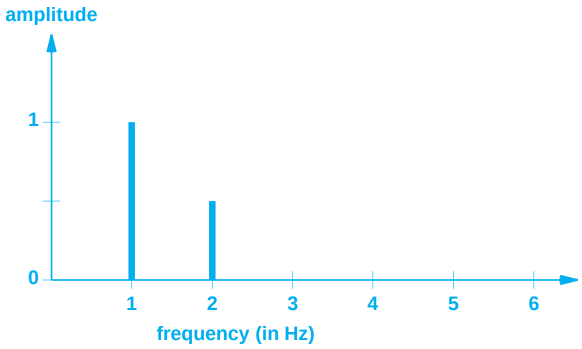


Figure 6.6 Representation of $\sin(2\pi t)$ and $0.5\sin(2\pi 2t)$ in the frequency domain.

The figure shows a set of simple periodic signals. A frequency domain representation can also be used with nonperiodic signals, but aperiodic representation is not essential to an understanding of the subject.

One of the advantages of the frequency domain representation arises from its compactness. Compared to a time domain representation, a frequency domain representation is both small and easy to read because each sine wave occupies a single point along

†Frequency domain diagrams used with real data communications systems have an x-axis that extends to thousands or millions of Hertz.

the x-axis. The advantage becomes clear when a composite signal contains many simple signals.

6.9 Bandwidth Of An Analog Signal

Most users have heard of “network bandwidth”, and understand that a network with high bandwidth is desirable. We will discuss the definition of network bandwidth later. For now, we will explore a related concept, *analog bandwidth*.

We define the bandwidth of an analog signal to be the difference between the highest and lowest frequencies of the constituent parts (i.e., the highest and lowest frequencies obtained by Fourier analysis). In the trivial example of Figure 6.5(c), Fourier analysis produces signals of 1 and 2 Hertz, which means the analog bandwidth is the difference, or 1 Hertz. An advantage of a frequency domain graph becomes clear when one computes analog bandwidth because the highest and lowest frequencies are obvious. For example, the plot in Figure 6.6 makes it clear that the analog bandwidth is 1.

Figure 6.7 shows a frequency domain plot with frequencies measured in Kilohertz (KHz). Such frequencies are in the range audible to a human ear. In the figure, the bandwidth is the difference between the highest and lowest frequency (5 KHz – 1 KHz = 4 KHz).

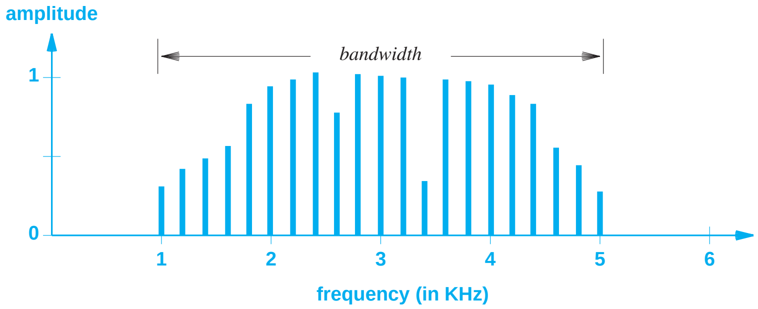


Figure 6.7 A frequency domain plot of an analog signal with a bandwidth of 4 KHz.

To summarize:

> *The* bandwidth *of an analog signal is the difference between the highest and lowest frequency of its components. If the signal is plotted in the frequency domain, the bandwidth is trivial to compute.*

6.10 Digital Signals And Signal Levels

We said in addition to being represented by an analog signal, information can also be represented by a *digital* signal. We further defined a signal to be digital if a fixed set of valid levels has been chosen and at any time, the signal is at one of the valid levels. Some systems use voltage to represent digital values by making a positive voltage correspond to a logical one, and zero voltage correspond to a logical zero. For example, +5 volts can be used for a logical one and 0 volts for a logical zero.

If only two levels of voltage are used, each level corresponds to one data bit (0 or 1). However, some physical transmission mechanisms can support more than two signal levels. When multiple digital levels are available, each level can represent multiple bits. For example, consider a system that uses four levels of voltage: –5 volts, –2 volts, +2 volts, and +5 volts. Each level can correspond to two bits of data as Figure 6.8(b) illustrates.

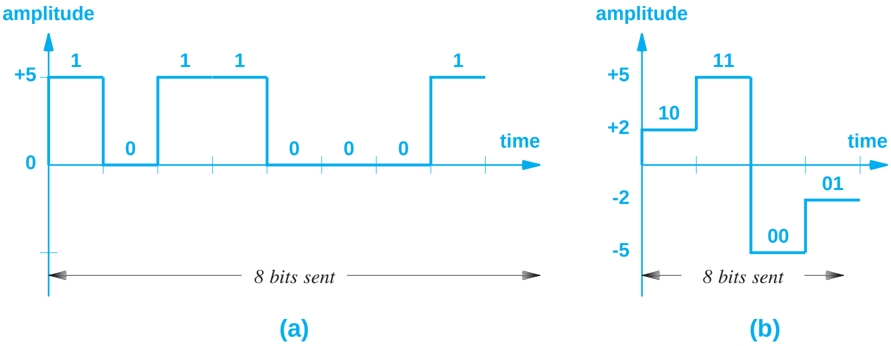


Figure 6.8 (a) A digital signal using two levels, and (b) the same digital signal using four levels.

As the figure illustrates, the chief advantage of using multiple signal levels arises from the ability to represent more than one bit at a time. In Figure 6.8(b), for example, –5 volts represents the two-bit sequence *00*, –2 volts represents *01*, +2 volts represents *10*, and +5 volts represents *11*. Because multiple levels of signal are used, each time slot can transfer two bits, which means that the four-level representation in Figure 6.8(b) takes half as long to transfer the bits as the two-level representation in Figure 6.8(a). Thus, the data rate (bits per second) is doubled.

The relationship between the number of levels required and the number of bits to be sent is straightforward. There must be a signal level for each possible combination of bits. Because 2^n combinations are possible with *n* bits, a communications system must use 2^n levels to represent *n* bits. To summarize:

> *A communications system that uses two signal levels can only send one bit at a given time; a system that supports 2^n signal levels can send n bits at a time.*

It may seem that voltage is an arbitrary quantity, and that one could achieve arbitrary numbers of levels by dividing voltage into arbitrarily small increments. Mathematically, one could create a million levels between 0 and 1 volts merely by using 0.0000001 volts for one level, 0.0000002 for the next level, and so on. Unfortunately, practical electronic systems cannot distinguish between signals that differ by arbitrarily small amounts. Thus, practical systems are restricted to a few signal levels.

6.11 Baud And Bits Per Second

How much data can be sent in a given time? The answer depends on two aspects of the communications system. As we have seen, the rate at which data can be sent depends on the number of signal levels. A second factor is also important: the amount of time the system remains at a given level before moving to the next. For example, the diagram in Figure 6.8(a) shows time along the x-axis, and the time is divided into eight segments, with one bit being sent during each segment. If the communications system is modified to use half as much time for a given bit, twice as many bits will be sent in the same amount of time. The point is:

> *An alternative method of increasing the amount of data that can be transferred in a given time consists of decreasing the amount of time that the system leaves a signal at a given level.*

As with signal levels, the hardware in a practical system places limits on how short the time can be — if the signal does not remain at a given level long enough, the receiving hardware will fail to detect it. Interestingly, the accepted measure of a communications system does not specify a length of time. Instead, engineers measure the inverse: how many times the signal can change per second, which is defined as the *baud*. For example, if a system requires the signal to remain at a given level for .001 seconds, we say that the system operates at 1000 baud.

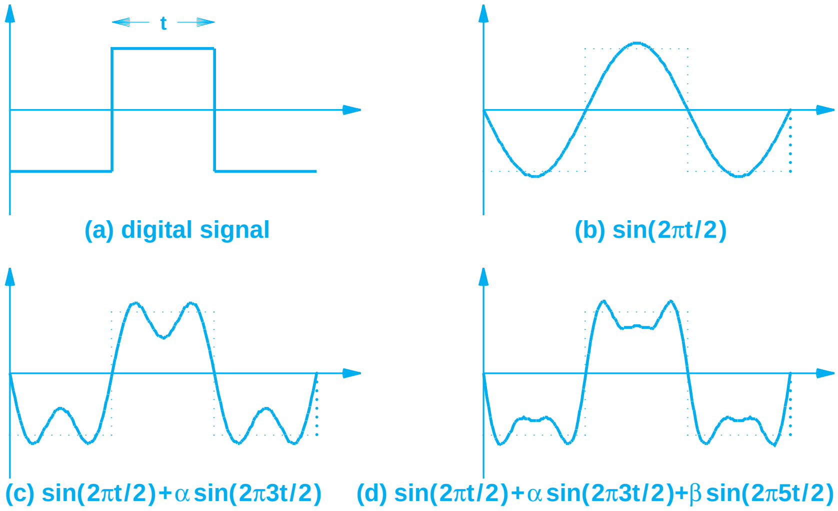
The key idea is that both baud and the number of signal levels control the bit rate. If a system with two signal levels operates at 1000 baud, the system can transfer exactly 1000 bits per second. However, if a system that operates at 1000 baud has four signal levels, the system can transfer 2000 bits per second (because four signal levels can represent two bits). Equation 6.1 expresses the relationship between baud, signal levels, and bit rate.

$$\textit{bits per second} \;=\; \textit{baud} \times \left\lfloor \log_2(\textit{ levels }) \right\rfloor \tag{6.1}$$

6.12 Converting A Digital Signal To Analog

How can a digital signal be converted into an equivalent analog signal? Recall that according to Fourier, an arbitrary curve can be represented as a composite of sine waves, where each sine wave in the set has a specific amplitude, frequency, and phase. Because it applies to any curve, Fourier's theorem also applies to a digital signal. From an engineering perspective, Fourier's result is impractical for digital signals because accurate representation of a digital signal requires an infinite set of sine waves.

Engineers adopt a compromise: conversion of a signal from digital to analog is *approximate*. That is, engineers build equipment to generate analog waves that closely approximate the digital signal. Approximation involves building a composite signal from only a few sine waves. By choosing sine waves that are the correct multiples of the digital signal frequency, as few as three sine waves can be used. The exact details are beyond the scope of this text, but Figure 6.9 illustrates the approximation by showing (a) a digital signal and approximations with (b) a single sine wave, (c) a composite of the original sine wave plus a sine wave of 3 times the frequency, and (d) a composite of the wave in (c) plus one more sine wave at 5 times the original frequency.



(a) digital signal

(b) $\sin(2\pi t/2)$

(c) $\sin(2\pi t/2)+\alpha\sin(2\pi 3t/2)$

(d) $\sin(2\pi t/2)+\alpha\sin(2\pi 3t/2)+\beta\sin(2\pi 5t/2)$

Figure 6.9 Approximation of a digital signal with sine waves.

6.13 The Bandwidth Of A Digital Signal

What is the bandwidth of a digital signal? Recall that the bandwidth of a signal is the difference between the highest and lowest frequency waves that constitute the signal. Thus, one way to calculate the bandwidth consists of applying Fourier analysis to find the constituent sine waves, and then examining the frequencies.

Mathematically, when Fourier analysis is applied to a square wave, such as the digital signal illustrated in Figure 6.9(a), the analysis produces an infinite set of sine waves. Furthermore, frequencies in the set continue to infinity. Thus, when plotted in the frequency domain, the set continues along the x-axis to infinity. The important consequence is:

> *According to the definition of bandwidth, a digital signal has infinite bandwidth because Fourier analysis of a digital signal produces an infinite set of sine waves with frequencies that grow to infinity.*

6.14 Synchronization And Agreement About Signals

Our examples leave out many of the subtle details involved in creating a viable communications system. For example, to guarantee that the sender and receiver agree on the amount of time allocated to each element of a signal, the electronics at both ends of a physical medium must have circuitry to measure time precisely. That is, if one end transmits a signal with 10^9 elements per second, the other end must expect exactly 10^9 elements per second. At slow speeds, making both ends agree is trivial. However, building electronic systems that agree at the high speeds used in modern networks is extremely difficult.

A more fundamental problem arises from the way data is represented in signals. The problem concerns *synchronization* of the sender and receiver. For example, suppose a receiver misses the first bit that arrives, and starts interpreting data starting at the second bit. Or consider what happens if a receiver expects data to arrive at a faster rate than the sender transmits the data. Figure 6.10 illustrates how a mismatch in interpretation can produce errors. In the figure, both the sender and receiver start and end at the same point in the signal, but because the receiver allocates slightly less time per bit, the receiver misinterprets the signal as containing more bits than were sent.

In practice, synchronization errors can be extremely subtle. For example, suppose a receiver's hardware has a timing error of 1 in 10^{-8}. The error might not show up until ten million bits are transmitted in a sequence. Because high-speed communications systems transfer gigabits per second, such small errors can surface quickly and become significant.

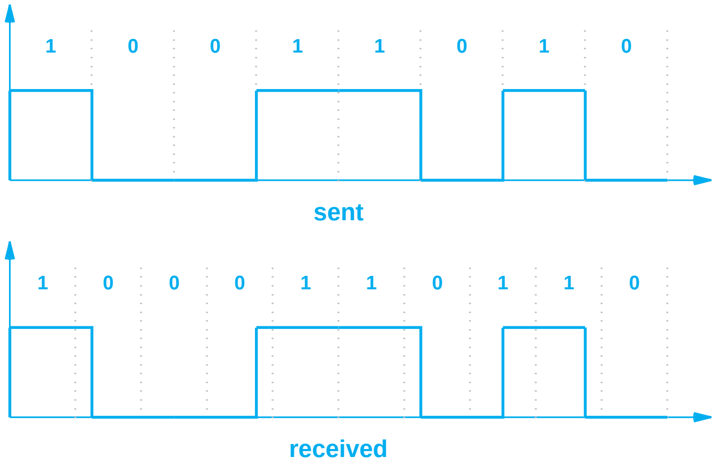


Figure 6.10 Illustration of a synchronization error in which the receiver allows slightly less time per bit than the sender.

6.15 Line Coding

Several techniques have been invented that can help avoid synchronization errors. In general, there are two broad approaches. In one approach, before it transmits data, the sender transmits a known pattern of bits, typically a set of alternating 0s and 1s, that allows the receiver to synchronize. In the other approach, data is represented by the signal in such a way that there can be no confusion about the meaning. We use the term *line coding* to describe the way data is encoded in a signal.

As an example of line coding that eliminates ambiguity, consider how one can use a transmission mechanism that supports three discrete signal levels. To guarantee synchronization, reserve one of the signal levels to start each bit. For example, if the three possible levels correspond to –5, 0, and +5 volts, reserve –5 to start each bit. Logical 0 can be represented by the sequence –5 0, and logical 1 can be represented by the sequence –5 +5. If we specify that no other combinations are valid, the occurrence of –5 volts always starts a bit, and a receiver can use an occurrence of –5 volts to correctly synchronize with the sender. Figure 6.11 illustrates the representation.

Of course, using multiple signal elements to represent a single bit means fewer bits can be transmitted per unit time. Thus, designers prefer schemes that transmit multiple bits per signal element, such as the one that Figure 6.8(b) illustrates†.

†Figure 6.8 can be found on page 134.

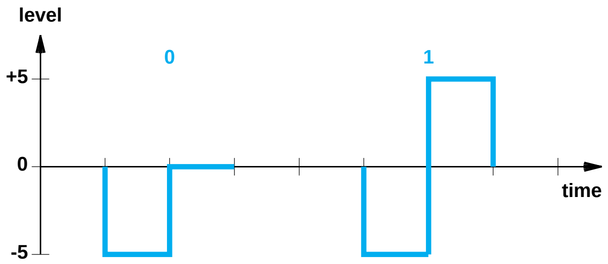


Figure 6.11 Example of two signal elements used to represent each bit.

Figure 6.12 lists the names of line coding techniques in common use, and groups them into related categories. Although the details are beyond the scope of this text, it is sufficient to know that the choice depends on the specific needs of a given communications system.

Category	Scheme	Synchronization
Unipolar	NRZ	No, if many 0s or 1s are repeated
	NRZ-L	No, if many 0s or 1s are repeated
	NRZ-I	No, if many 0s or 1s are repeated
	Biphase	Yes
Bipolar	AMI	No, if many 0s are repeated
Multilevel	2B1Q	No, if many double bits are repeated
	8B6T	Yes
	4D-PAM5	Yes
Multiline	MLT-3	No, if many 0s are repeated

Figure 6.12 Names of line coding techniques in common use.

The point is:

> *A variety of line coding techniques are available that differ in how they handle synchronization as well as other properties such as the bandwidth used.*

6.16 Manchester Encoding Used In Computer Networks

In addition to the list in Figure 6.12, one particular standard for line coding is especially important for computer networks: the *Manchester Encoding* used with Ethernet†.

To understand Manchester Encoding, it is important to know that detecting a transition in signal level is easier than measuring the signal level. The fact, which arises from the way hardware works, explains why the Manchester Encoding uses transitions rather than levels to define bits. That is, instead of specifying that 1 corresponds to a level (e.g., +5 volts), Manchester Encoding specifies that a 1 corresponds to a transition from 0 volts to a positive voltage level. Correspondingly, a 0 corresponds to a transition from a positive voltage level to zero. Furthermore, the transitions occur in the "middle" of the time slot allocated to a bit, which allows the signal to return to the previous level in case the data contains two repeated 0s or two repeated 1s. Figure 6.13(a) illustrates the concept.

A variation known as a *Differential Manchester Encoding* (also called a *Conditional DePhase Encoding*) uses relative transitions rather than absolute. That is, the representation of a bit depends on the previous bit. Each bit time slot contains one or two transitions. A transition *always* occurs in the middle of the bit time. The logical value of the bit is represented by the presence or absence of a transition at the beginning of a bit time: logical 0 is represented by a transition, and logical 1 is represented by no transition. Figure 6.13(b) illustrates Differential Manchester Encoding. Perhaps the most important property of differential encoding arises from a practical consideration: the encoding works correctly even if the two wires carrying the signal are accidentally reversed.

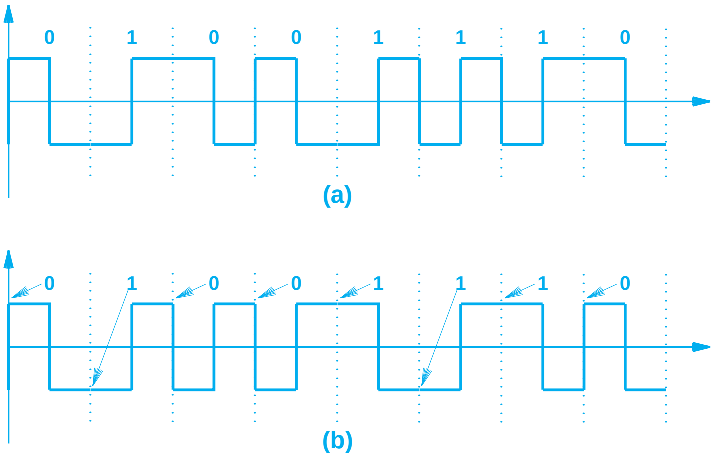


Figure 6.13 (a) Manchester and (b) Differential Manchester Encodings; each assumes the previous bit ended with a low signal level.

†Chapter 15 discusses Ethernet.

6.17 Converting An Analog Signal To Digital

Many sources of information are analog, which means they must be converted to digital form for further processing (e.g., before they can be encrypted). There are two basic approaches:

- Pulse code modulation
- Delta modulation

Pulse code modulation (*PCM*†) refers to a technique where the level of an analog signal is measured repeatedly at fixed time intervals and converted to digital form. Figure 6.14 illustrates the steps.

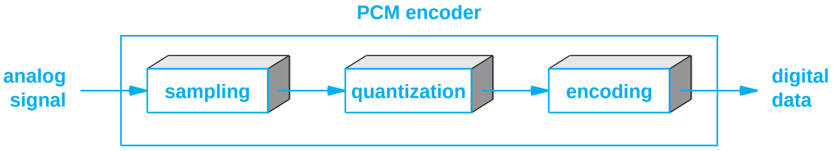


Figure 6.14 The three steps used in pulse code modulation.

Each measurement is known as a *sample*, which explains why the first stage is known as *sampling*. After it has been recorded, a sample is *quantized* by converting it into a small integer value which is then *encoded* into a specific format. The quantized value is not a measure of voltage or any other property of the signal. Instead, the range of the signal from the minimum to maximum levels is divided into a set of slots, typically a power of 2. Figure 6.15 illustrates the concept by showing a signal quantized into eight slots.

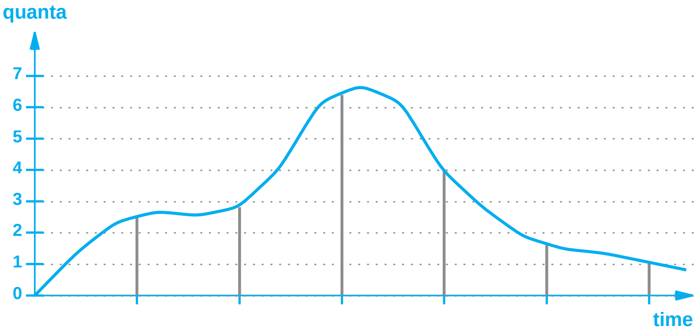


Figure 6.15 An illustration of the sampling and quantization used in pulse code modulation.

†The acronym PCM is ambiguous because it can refer to the general idea or to a specific form of pulse code modulation used by the telephone system. A later section discusses the latter.

In the figure, the six samples are represented by vertical gray lines. Each sample is quantized by choosing the closest quantum interval. For example, the third sample, taken near the peak of the curve is assigned a quantized value of 6.

In practice, slight variations in sampling have been invented. For example, to avoid inaccuracy caused by a brief spike or a dip in the signal, averaging can be used. That is, instead of relying on a single measurement for each sample, three measurements can be taken close together and an arithmetic mean can be computed.

The chief alternative to pulse code modulation is known as *delta modulation*. Delta modulation also takes samples. However, instead of sending a quantization for each sample, delta modulation sends one quantization value followed by a string of values that give the difference between the previous value and the current value. The idea is that transmitting differences requires fewer bits than transmitting full values, especially if the signal does not vary rapidly. The main tradeoff with delta modulation arises from the effect of an error — if any item in the sequence is lost or damaged, all successive values will be misinterpreted. Thus, communications systems that expect data values to be lost or changed during transmission usually use pulse code modulation (PCM).

6.18 The Nyquist Theorem And Sampling Rate

Whether pulse code or delta modulation is used, the analog signal must be sampled. How frequently should an analog signal be sampled? Taking too few samples (known as *undersampling*) means that the digital values only give a crude approximation of the original signal. Taking too many samples (known as *oversampling*) means that more digital data will be generated, which uses extra bandwidth.

A mathematician named Nyquist discovered the answer to the question of how much sampling is required:

$$\textit{sampling rate} \; = \; 2 \times f_{\max} \qquad (6.2)$$

where $f_{\max}$ is the highest frequency in the composite signal. The result, which is known as the *Nyquist Theorem*, provides a practical solution to the problem: sample a signal at least twice as fast as the highest frequency that must be preserved.

6.19 Nyquist Theorem And Telephone System Transmission

As a specific example of the Nyquist Theorem, consider the telephone system that was originally designed to transfer voice. Measurements of human speech have shown that preserving frequencies between 0 and 4000 Hz provides acceptable audio quality. Thus, the Nyquist Theorem specifies that when converting a voice signal from analog to digital, the signal should be sampled at a rate of 8000 samples per second.

To further provide reasonable quality reproduction, the PCM standard used by the phone system quantifies each sample into an 8-bit value. That is, the range of input is divided into 256 possible levels so that each sample has a value between 0 and 255. As a consequence, the rate at which digital data is generated for a single telephone call is:

$$\textit{digitized voice call} \;=\; 8000\,\frac{\textit{samples}}{\textit{second}} \times 8\,\frac{\textit{bits}}{\textit{sample}} \;=\; 64{,}000\,\frac{\textit{bits}}{\textit{second}} \qquad (6.3)$$

As we will see in later chapters, the telephone system uses the rate of 64,000 bits per second (64 Kbps) as the basis for digital communication. We will further see that the Internet uses digital telephone circuits to span long distances.

6.20 Nonlinear Encoding

When each sample only has eight bits, the linear PCM encoding illustrated in Figure 6.15 does not work well for voice. Researchers have devised nonlinear alternatives that can reproduce sounds to which the human ear is most sensitive. Two nonlinear digital telephone standards have been created, and are in wide use:

- *a-law*, a standard used in Europe
- μ-*law*, a standard used in North America and Japan

Both standards use 8-bit samples, and generate 8000 samples per second. The difference between the two arises from a tradeoff between the overall range and sensitivity to noise. The μ-law algorithm has the advantage of covering a wider dynamic range (i.e., the ability to reproduce louder sounds), but has the disadvantage of introducing more distortion of weak signals. The a-law algorithm provides less distortion of weak signals, but has a smaller dynamic range. For international calls, a conversion to a-law encoding must be performed if one side uses a-law and the other uses μ-law.

6.21 Encoding And Data Compression

We use the term *data compression* to refer to a technique that reduces the number of bits required to represent data. Data compression is especially relevant to a communications system, because reducing the number of bits used to represent data reduces the time required for transmission. That is, a communications system can be optimized by compressing data before transmission.

Chapter 28 considers compression in multimedia applications. At this point, we only need to understand the basic definitions of the two types of compression:

- Lossy — some information is lost during compression
- Lossless — all information is retained in the compressed version.

Lossy compression is generally used with data that a human consumes, such as an image, a segment of video, or an audio file. The key idea is that the compression only needs to preserve details to the level of human perception. That is, a change is acceptable if humans cannot detect the change. We will see that well-known compression schemes such as JPEG (used for images) or MPEG-3 (abbreviated MP3 and used for audio recordings) employ lossy compression.

Lossless compression preserves the original data without any change. Thus, lossless compression can be used for documents or in any situation where data must be preserved exactly. When used for communication, a sender compresses the data before transmission, and the receiver decompresses the result. Because the compression is lossless, arbitrary data can be compressed by a sender and decompressed by a receiver to recover an exact copy of the original.

Most lossless compression uses a *dictionary* approach. Compression finds strings that are repeated in the data, and forms a *dictionary* of the strings. To compress the data, each occurrence of a string is replaced by a reference to the dictionary. The sender must transmit the dictionary along with the compressed data. If the data contains strings that are repeated many times, the combination of the dictionary plus the compressed data is smaller than the original data.

6.22 Summary

An information source can deliver analog or digital data. An analog signal has the property of being aperiodic or periodic; a periodic signal has properties of amplitude, frequency, and phase. Fourier discovered that an arbitrary curve can be formed from a sum of sine waves; a single sine wave is classified as simple, and a signal that can be decomposed into multiple sine waves is classified as composite.

Engineers use two main representations of composite signals. A time domain representation shows how the signal varies over time. A frequency domain representation shows the amplitude and frequency of each component in the signal. The bandwidth, which is the difference between the highest and lowest frequencies in a signal is especially clear on a frequency domain graph.

The baud rate of a signal is the number of times the signal can change per second. A digital signal that uses multiple signal levels can represent more than one bit per change, making the effective transmission rate the number of levels times the baud rate. Although it has infinite bandwidth, a digital signal can be approximated with as few as three sine waves.

Various line coding techniques exist. The Manchester Encoding, used with Ethernet networks, is especially important. Rather than using absolute signal levels to represent bits, the Manchester Encoding uses transitions in signal level. The Differential Manchester Encoding uses relative transitions, and has the property that it works even if the two wires are reversed.

Pulse code modulation and delta modulation are used to convert an analog signal to digital. The PCM scheme used by the telephone system employs 8-bit quantization and takes 8000 samples per second, which results in a rate of 64 Kbps.

Compression is lossy or lossless. Lossy compression is most appropriate for images, audio, or video that will be viewed by humans because loss can be controlled to keep changes below the threshold of human perception. Lossless compression is most appropriate for documents or data that must be preserved exactly.

EXERCISES

6.1 Name a common household device that emits an aperiodic signal.

6.2 Give three examples of information sources other than computers.

6.3 State and describe the four fundamental characteristics of a sine wave.

6.4 Why are sine waves fundamental to data communications?

6.5 When is a wave classified as *simple*?

6.6 When shown a graph of a sine wave, what is the quickest way to determine whether the phase is zero?

6.7 On a frequency domain graph, what does the y-axis represent?

6.8 What does Fourier analysis of a composite wave produce?

6.9 Is bandwidth easier to compute from a time domain or frequency domain representation? Why?

6.10 What is the analog bandwidth of a signal?

6.11 What is the definition of *baud*?

6.12 Suppose an engineer increases the number of possible signal levels from two to four. How many more bits can be sent in the same amount of time? Explain.

6.13 What is the bandwidth of a digital signal? Explain.

6.14 Why is an analog signal used to approximate a digital signal?

6.15 Why do some coding techniques use multiple signal elements to represent a single bit?

6.16 What is a synchronization error?

6.17 What is the chief advantage of a Differential Manchester Encoding?

6.18 What aspect of a signal does the Manchester Encoding use to represent a bit?

6.19 If the maximum frequency audible to a human ear is 20,000 Hz, at what rate must the analog signal from a microphone be sampled when converting it to digital?

6.20 When converting an analog signal to digital, what step follows sampling?

6.21 Describe the difference between lossy and lossless compressions, and tell when each might be used.

6.22 What time elapses between each sample for the PCM encoding used in the telephone system?

Chapter Contents

7

Transmission Media

7.1 Introduction

Chapter 5 provides an overview of data communications. The previous chapter considers the topic of information sources. The chapter examines analog and digital information, and explains encodings.

This chapter continues the discussion of data communications by considering transmission media, including wired, wireless, and optical media. The chapter gives a taxonomy of media types, introduces basic concepts of electromagnetic propagation, and explains how shielding can reduce or prevent interference and noise. Finally, the chapter explains the concept of capacity. Successive chapters continue the discussion of data communications.

7.2 Guided And Unguided Transmission

How should transmission media be divided into classes. There are two broad approaches:

- By type of path: communication can follow an exact path such as a wire, or can have no specific path, such as a radio transmission.
- By form of energy: electrical energy is used on wires, radio transmission is used for wireless, and light is used for optical fiber.

We use the terms *guided* and *unguided* transmission to distinguish between physical media such as copper wiring or optical fibers that provide a specific path and a radio transmission that travels in all directions through free space. Informally, engineers use the terms *wired* and *wireless*. Note that the informality can be somewhat confusing because one is likely to hear the term *wired* even when the physical medium is an optical fiber.

7.3 A Taxonomy By Forms Of Energy

Figure 7.1 illustrates how physical media can be classified according to the form of energy used to transmit data. Successive sections describe each of the media types.

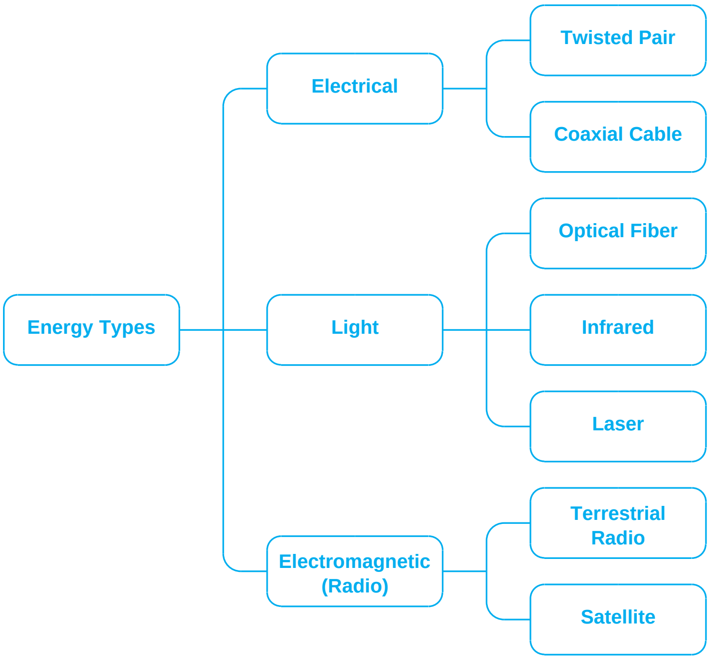


Figure 7.1 A taxonomy of media types according to the form of energy used.

Like most taxonomies, the categories are not perfect, and exceptions exist. For example, a space station in orbit around the earth might employ non-terrestrial communication that does not involve a satellite. Nevertheless, our taxonomy covers most communications.

7.4 Background Radiation And Electrical Noise

Recall from basic physics that electrical current flows along a complete circuit. Thus, all transmissions of electrical energy need two wires to form a circuit — a wire to the receiver and a wire back to the sender. The simplest form of wiring consists of a cable that contains two copper wires. Each wire is wrapped in a plastic coating, which insulates the wires electrically. The outer coating on the cable holds related wires together to make it easier for humans who connect equipment.

Computer networks use an alternative form of wiring. To understand why, one must know three facts.

- Random electromagnetic radiation, called *noise*, permeates the environment. In fact, communications systems generate minor amounts of electrical noise as a side effect of normal operation.
- When it hits metal, electromagnetic radiation induces a small signal, which means that random noise can interfere with signals used for communication.
- Because it absorbs radiation, metal acts as a *shield*. Thus, placing enough metal between a source of noise and a communication medium can prevent noise from interfering with communication.

The first two facts outline a fundamental problem inherent in communication media that use electrical or radio energy. The problem is especially severe near a source that emits random radiation. For example, fluorescent light bulbs and electric motors both emit radiation, especially powerful motors such as those used to operate elevators, air conditioners, and refrigerators. Surprisingly, smaller devices such as paper shredders or electric power tools can also emit enough radiation to interfere with communication. The point is:

> *The random electromagnetic radiation generated by devices such as electric motors can interfere with communication that uses radio transmission or electrical energy sent over wires.*

7.5 Twisted Pair Copper Wiring

The third fact in the previous section explains the wiring used with communications systems. There are three forms of wiring that help reduce interference from electrical noise.

- Unshielded Twisted Pair (UTP)
- Coaxial cable
- Shielded Twisted Pair (STP)

The first form, which is known as *twisted pair* wiring or *unshielded twisted pair* wiring†, is used extensively in communications. As the name implies, twisted pair wiring consists of two wires that are twisted together. Of course, each wire has a plastic coating that insulates the two wires and prevents electrical current from flowing between them.

Surprisingly, twisting two wires makes them less susceptible to electrical noise than leaving them parallel. Figure 7.2 illustrates why.

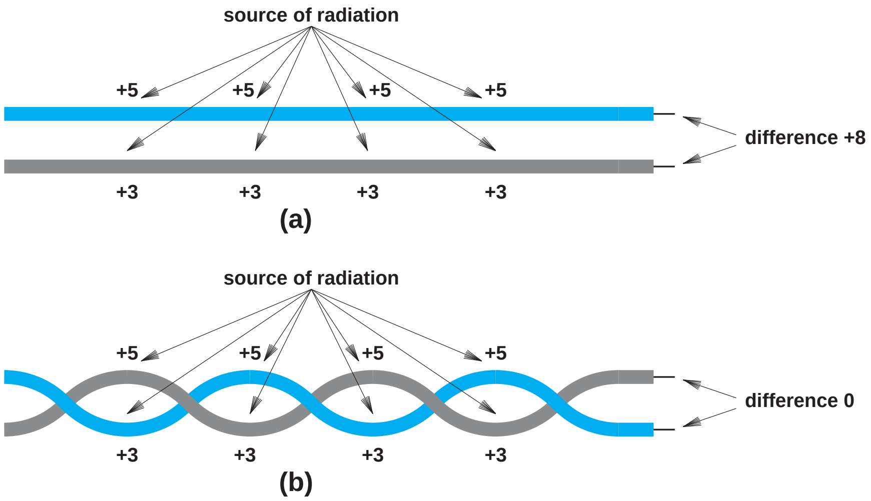


Figure 7.2 Unwanted electromagnetic radiation affecting (a) two parallel wires, and (b) twisted pair wiring.

As the figure shows, when two wires are in parallel, there is a high probability that one of them is closer to the source of electromagnetic radiation than the other. In fact, one wire tends to act as a shield that absorbs some of the electromagnetic radiation. Thus, because it is hidden behind the first wire, the second wire receives less energy. In the figure, a total of 32 units of radiation strikes each of the two cases. In Figure 7.2(a), the top wire absorbs 20 units, and the bottom wire absorbs 12, producing a difference of 8. In Figure 7.2(b), each of the two wires is on top one-half of the time, which means each wire absorbs the same amount of radiation.

Why does equal absorption matter? The answer is that if interference induces exactly the same amount of electrical energy in each wire, no extra current will flow. Thus, the original signal will not be disturbed. The point is:

†A later section explains the term *shielded*.

To reduce the interference caused by random electromagnetic radiation, communications systems use twisted pair wiring rather than parallel wires.

7.6 Shielding: Coaxial Cable And Shielded Twisted Pair

Although it is immune to most background radiation, twisted pair wiring does not solve all problems. Twisted pair wiring tends to have problems with:

- Especially strong electrical noise
- Close physical proximity to the source of noise
- High frequencies used for communication

If the intensity is high (e.g., in a factory that uses electric arc welding equipment) or communication cables run close to the source of electrical noise, even twisted pair may not be sufficient. Thus, if a twisted pair runs above the ceiling in an office building on top of a fluorescent light fixture, interference may result. Furthermore, it is difficult to build equipment that can distinguish between valid high frequency signals and noise, which means that even a small amount of noise can cause interference when high frequencies are used.

To handle situations where twisted pair is insufficient, forms of wiring are available that have extra metal shielding. The most familiar form is the wiring used for cable television. Known as *coaxial cable* (*coax*), the wiring has a thick metal shield, formed from braided wires, that completely surrounds a center wire that carries the signal. Figure 7.3 illustrates the concept.

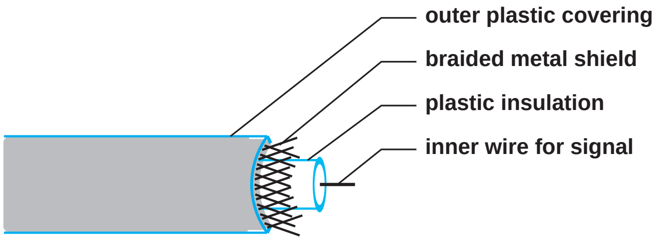


Figure 7.3 Illustration of coaxial cable with a shield surrounding the signal wire.

The shield in a coaxial cable forms a flexible cylinder around the inner wire that provides a barrier to electromagnetic radiation from any direction. The barrier also prevents signals on the inner wire from radiating electromagnetic energy that could affect other wires. Consequently, a coaxial cable can be placed adjacent to sources of electrical noise and other cables, and can be used for high frequencies. The point is:

> *The heavy shielding and symmetry makes coaxial cable immune to noise, capable of carrying high frequencies, and prevents signals on the cable from emitting noise to surrounding cables.*

Using braided wire instead of a solid metal shield keeps coaxial cable flexible, but the heavy shield does make coaxial cable less flexible than twisted pair wiring. Variations of shielding have been invented that provide a compromise: the cable is more flexible, but has slightly less immunity to electrical noise. One popular variation is known as *shielded twisted pair* (*STP*). An STP cable has a thinner, more flexible metal shield surrounding one or more twisted pairs of wires. In most versions of STP cable, the shield consists of metal foil, similar to the aluminum foil used in a kitchen. STP cable has the advantages of being more flexible than a coaxial cable and less susceptible to electrical interference than *unshielded twisted pair* (*UTP*).

7.7 Categories Of Twisted Pair Cable

The telephone companies originally specified standards for twisted pair wiring used in the telephone network. More recently, three standards organizations worked together to create standards for twisted pair cables used in computer networks. The *American National Standards Institute* (*ANSI*), the *Telecommunications Industry Association* (*TIA*), and the *Electronic Industries Alliance* (*EIA*) created a list of wiring categories, with strict specifications for each. Figure 7.4 summarizes the main categories.

Category	Description	Data Rate (in Mbps)
CAT 1	Unshielded twisted pair used for telephones	< 0.1
CAT 2	Unshielded twisted pair used for T1 data	2
CAT 3	Improved CAT2 used for computer networks	10
CAT 4	Improved CAT3 used for Token Ring networks	20
CAT 5	Unshielded twisted pair used for networks	100
CAT 5E	Extended CAT5 for more noise immunity	125
CAT 6	Unshielded twisted pair tested for 200 Mbps	200
CAT 7	Shielded twisted pair with a foil shield around the entire cable plus a shield around each twisted pair	600

Figure 7.4 Twisted pair wiring categories and a description of each.

7.8 Media Using Light Energy And Optical Fibers

According to the taxonomy in Figure 7.1, three forms of media use light energy to carry information:

- Optical fibers
- Infrared transmission
- Point-to-point lasers

The most important type of media that uses light is an *optical fiber*. Each fiber consists of a thin strand of glass or transparent plastic encased in a plastic cover. A typical optical fiber is used for communication in a single direction — one end of the fiber connects to a laser or LED used to transmit light, and the other end of the fiber connects to a photosensitive device used to detect incoming light. To provide two-way communication, two fibers are used, one to carry information in each direction. Thus, optical fibers are usually collected into a cable by wrapping a plastic cover around them; a cable has at least two fibers, and a cable used between large sites with multiple network devices may contain many fibers.

Although it cannot be bent at a right angle, an optical fiber is flexible enough to form into a circle with diameter less than two inches without breaking. The question arises: why does light travel around a bend in the fiber? The answer comes from physics: when light encounters the boundary between two substances, its behavior depends on the density of the two substances and the angle at which the light strikes the boundary. For a given pair of substances, there exists a *critical angle*, θ, measured with respect to a line that is perpendicular to the boundary. If the angle of incidence is exactly equal to the critical angle, light travels along the boundary. When the angle is less than θ degrees, light crosses the boundary and is *refracted*, and when the angle is greater than θ degrees, light is reflected as if the boundary were a mirror. Figure 7.5 illustrates the concept.

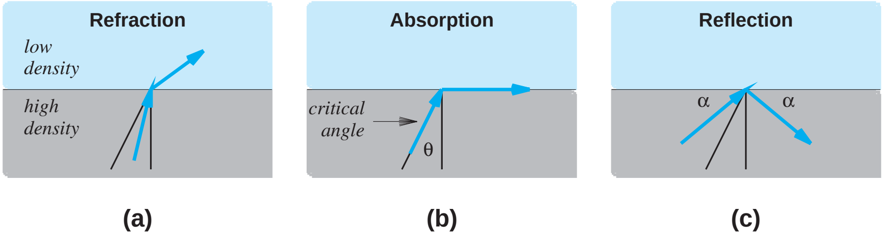


Figure 7.5 Behavior of light at a density boundary when the angle of incidence is (a) less than the critical angle θ, (b) equal to the critical angle, and (c) greater than the critical angle.

Figure 7.5(c) explains why light stays inside an optical fiber — a substance called *cladding* is bonded to the fiber to form a boundary. As it travels along, light is reflected off the boundary.

Unfortunately, reflection in an optical fiber is not perfect. Reflection absorbs a small amount of energy. Furthermore, if a photon takes a zig-zag path that reflects from the walls of the fiber many times, the photon will travel a slightly longer distance than a photon that takes a straight path. The result is that a pulse of light sent at one end of a fiber emerges with less energy and is *dispersed* (i.e., stretched) over time, as Figure 7.6 illustrates.

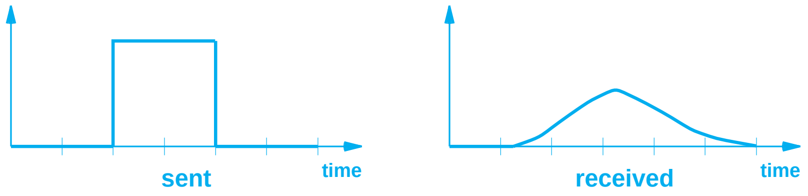


Figure 7.6 A light pulse as sent and received over an optical fiber.

7.9 Types Of Fiber And Light Transmission

Although it is not a problem for optical fibers used to connect a computer to a nearby device, dispersion becomes a serious problem for long optical fibers, such as those used between two cities or under an ocean. Consequently, three forms of optical fibers have been invented that provide a choice between performance and cost:

- *Multimode, step index fiber* is the least expensive, and is used when performance is unimportant. The boundary between the fiber and the cladding is abrupt which causes light to reflect frequently. Therefore, dispersion is high.

- *Multimode, graded index fiber* is slightly more expensive than the multimode, step index fiber. However, it has the advantage of making the density of the fiber increase near the edge, which reduces reflection and lowers dispersion.

- *Single mode fiber* is the most expensive, and provides the least dispersion. The fiber has a smaller diameter and other properties that help reduce reflection. Single mode is used for long distances and higher bit rates.

Single mode fiber and the equipment used at each end are designed to focus light. As a result, a pulse of light can travel thousands of kilometers without becoming dispersed. Minimal dispersion helps increase the rate at which bits can be sent because a pulse corresponding to one bit does not disperse into the pulse that corresponds to a successive bit.

How is light sent and received on a fiber? The key is that the devices used for transmission must match the fiber. The available mechanisms include:

- Transmission: Light Emitting Diode (LED) or Injection Laser Diode (ILD)
- Reception: photo-sensitive cell or photodiode

In general, LEDs and photo-sensitive cells are used for short distances and slower bit rates common with multimode fiber. Single mode fiber, used over long distances with high bit rates, generally requires ILDs and photodiodes.

7.10 Optical Fiber Compared To Copper Wiring

Optical fiber has several properties that make it more desirable than copper wiring. Optical fiber is immune to electrical noise, has higher bandwidth, and light traveling across a fiber does not attenuate as much as electrical signals traveling across copper. However, copper wiring is less expensive. Furthermore, because the ends of an optical fiber must be polished before they can be used, installation of copper wiring does not require as much special equipment or expertise as optical fiber. Finally, because they are stronger, copper wires are less likely to break if accidentally pulled or bent. Figure 7.7 summarizes the advantages of each media type.

Optical Fiber

- Immune to electrical noise
- Less signal attenuation
- Higher bandwidth

Copper Wiring

- Lower overall cost
- Less expertise/equipment needed
- Less easily broken

Figure 7.7 Advantages of optical fiber and copper wiring.

7.11 Infrared Communication Technologies

InfraRed (*IR*) communication technologies use the same type of energy as a typical television remote control: a form of electromagnetic radiation that behaves like visible light but falls outside the range that is visible to a human eye. Like visible light, infrared disperses quickly. Infrared signals can reflect from a smooth, hard surface, and an opaque object as thin as a sheet of paper can block the signal, as does moisture in the atmosphere.

The point is:

> *Infrared communication technologies are best suited for use indoors in situations where the path between sender and receiver is short and free from obstruction.*

The most commonly used infrared technology is intended to connect a computer to a nearby peripheral, such as a printer. An interface on the computer and an interface on the printer each send an infrared signal that covers an arc of approximately 30 degrees. Provided the two devices are aligned, each can receive the other's signal. The wireless aspect of infrared is especially attractive for laptop computers because a user can move around a room and still have access to a printer. Figure 7.8 lists the three commonly used infrared technologies along with the data rate that each supports.

Name	Expansion	Speed
IrDA-SIR	Slow-speed Infrared	0.115 Mbps
IrDA-MIR	Medium-speed Infrared	1.150 Mbps
IrDA-FIR	Fast-speed Infrared	4.000 Mbps

Figure 7.8 Three common infrared technologies and the data rate of each.

7.12 Point-To-Point Laser Communication

Because they connect a pair of devices with a beam that follows the line-of-sight, the infrared technologies described above can be classified as providing *point-to-point* communication. In addition to infrared, other point-to-point communication technologies exist. One form of point-to-point communication uses a beam of coherent light produced by a *laser*.

Like infrared, laser communication follows line-of-sight, and requires a clear, unobstructed path between the communicating sites. Unlike an infrared transmitter, however, a laser beam does not cover a broad area. Instead, the beam is only a few centimeters wide. Consequently, the sending and receiving equipment must be aligned precisely to ensure that the sender's beam hits the sensor in the receiver's equipment. In a typical communications system, two-way communication is needed. Thus, each side must have both a transmitter and receiver, and both transmitters must be aligned carefully. Because alignment is critical, point-to-point laser equipment is usually mounted permanently.

Laser beams have the advantage of being suitable for use outdoors, and can span greater distances than infrared. As a result, laser technology is especially useful in cities to transmit from building to building. For example, imagine a large corporation with offices in two adjacent buildings. A corporation is not permitted to string wires across streets between buildings. However, a corporation can purchase laser communication equipment and permanently mount the equipment, either on the sides of the two buildings or on the roofs. Once the equipment has been purchased and installed, the operating costs are relatively low.

To summarize:

> *Laser technology can be used to create a point-to-point communications system. Because a laser emits a narrow beam of light, the transmitter and receiver must be aligned precisely; typical installations affix the equipment to a permanent structure, such as the roof of a building.*

7.13 Electromagnetic (Radio) Communication

Recall that the term *unguided* is used to characterize communication technologies that can propagate energy without requiring a medium such as a wire or optical fiber. The most common form of unguided communication mechanisms consists of *wireless* networking technologies that use electromagnetic energy in the *Radio Frequency* (*RF*) range. RF transmission has a distinct advantage over light because RF energy can traverse long distances and penetrate objects such as the walls of a building.

The exact properties of electromagnetic energy depend on the frequency. We use the term *spectrum* to refer to the range of possible frequencies; governments around the world allocate frequencies for specific purposes. In the U.S., the *Federal Communications Commission* sets rules for how frequencies are allocated, and sets limits on the amount of power that communication equipment can emit at each frequency. Figure 7.9 shows the overall electromagnetic spectrum and general characteristics of each piece. As the figure shows, one part of the spectrum corresponds to infrared light described above. The spectrum used for RF communications spans frequencies from approximately 3 KHz to 300 GHz, and includes frequencies allocated to radio and television broadcast as well as satellite and microwave communications†.

†Google provides an interesting system that shows spectrum availability at various points in the United States at URL: https://support.google.com/spectrumdatabase/

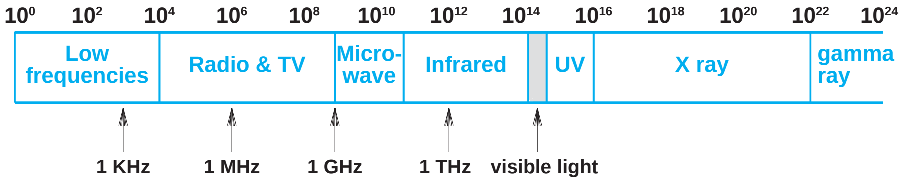


Figure 7.9 Major pieces of the electromagnetic spectrum with frequency in Hz shown on a log scale.

7.14 Signal Propagation

Chapter 6 explains that the amount of information an electromagnetic wave can represent depends on the wave's frequency. The frequency of an electromagnetic wave also determines how the wave *propagates*. Figure 7.10 describes the three broad types of wave propagation.

Classification	Range	Type Of Propagation
Low Frequency	< 2 MHz	Wave follows earth's curvature, but can be blocked by unlevel terrain
Medium Frequency	2 to 30 MHz	Wave can reflect from layers of the atmosphere, especially the ionosphere
High Frequency	> 30 MHz	Wave travels in a direct line, and will be blocked by obstructions

Figure 7.10 Electromagnetic wave propagation at various frequencies.

According to the figure, the lowest frequencies of electromagnetic radiation follow the earth's surface, which means that if the terrain is relatively flat, it will be possible to place a receiver beyond the horizon from a transmitter. With medium frequencies, a transmitter and receiver can be farther apart because the signal can bounce off the ionosphere to travel between them. Finally, the highest frequencies of radio transmission behave like light — the signal propagates in a straight line from the transmitter to the receiver, and the path must be free from obstructions. The point is:

> *The frequencies used for wireless networking technologies cannot be chosen arbitrarily because governments control the use of spectrum and each frequency has characteristics such as wave propagation, power requirements, and susceptibility to noise.*

Wireless technologies are classified into two broad categories as follows:

- *Terrestrial*. Communication uses equipment such as radio or microwave transmitters that is relatively close to the earth's surface. Typical locations for antennas or other equipment include the tops of hills, man-made towers, and tall buildings.
- *Nonterrestrial*. Some of the equipment used in communication is outside the earth's atmosphere (e.g., a satellite in orbit around the earth).

Chapter 16 presents specific wireless technologies, and describes the characteristics of each. For now, it is sufficient to understand that the frequency and amount of power used can affect the speed at which data can be sent, the maximum distance over which communication can occur, and characteristics such as whether the signal can penetrate solid objects.

7.15 Types Of Satellites

The laws of physics (specifically *Kepler's Law*) govern the motion of an object, such as a satellite, that orbits the earth. In particular, the period (i.e., time required for a complete orbit) depends on the distance from the earth. Consequently, communication satellites are classified into three broad categories, depending on their distance from the earth. Figure 7.11 lists the categories, and describes each.

Orbit Type	Description
Low Earth Orbit (LEO)	Has the advantage of low delay, but the disadvantage that from an observer's point of view on the earth, the satellite appears to move across the sky
Medium Earth Orbit (MEO)	An elliptical (rather than circular) orbit used to provide communication at the North and South Poles†
Geostationary Earth Orbit (GEO)	Has the advantage that the satellite remains at a fixed position with respect to a location on the earth's surface, but the disadvantage of being farther away

Figure 7.11 The three basic categories of communication satellites.

†In 2013, a commercial company (O3b) announced that it would create the first MEO satellite cluster with the intention of providing Internet service to the currently unserved population (the other three billion).

7.16 Geostationary Earth Orbit (GEO) Satellites

As Figure 7.11 explains, the main tradeoff in communication satellites is between height and orbital period. The chief advantage of a satellite in *Geostationary Earth Orbit* (*GEO*) arises because the orbital period is exactly the same as the rate at which the earth rotates. If positioned above the equator, a GEO satellite remains in exactly the same location over the earth's surface at all times. A stationary satellite position means that once a *ground station* has been aligned with the satellite, the equipment never needs to move. Figure 7.12 illustrates the concept.

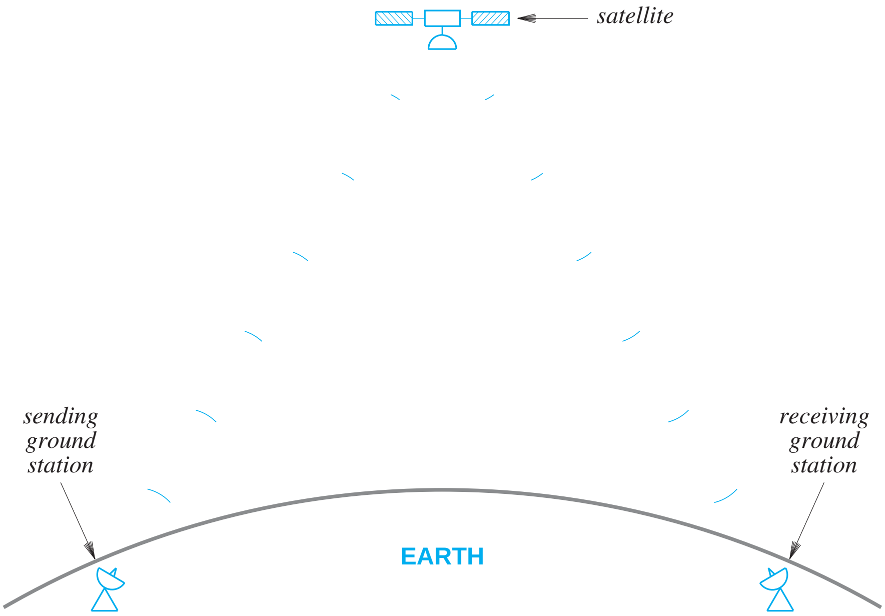


Figure 7.12 A GEO satellite and ground stations permanently aligned.

Unfortunately, the distance required for a geostationary orbit is 35,785 kilometers or 22,236 miles, which is approximately one tenth the distance to the moon. To understand what such a distance means for communication, consider a radio wave traveling to a GEO satellite and back. At the speed of light, $3{\times}10^8$ meters per second, the trip takes:

$$\frac{2 \times 35.8 \times 10^6 \, meters}{3 \times 10^8 \; meters/sec} = 0.238 \; sec \tag{7.1}$$

Although it may seem unimportant, a delay of approximately 0.2 seconds can be significant for some applications. In a telephone call or a video teleconference, a human can notice a 0.2 second delay. For electronic transactions such as a stock exchange offering a limited set of bonds, delaying an offer by 0.2 seconds may mean the difference between a successful and unsuccessful offer. To summarize:

> *Even at the speed of light, a signal takes more than 0.2 seconds to travel from a ground station to a GEO satellite and back to another ground station.*

7.17 GEO Coverage Of The Earth

How many GEO communication satellites are possible? Interestingly, there is a limited amount of “space” available in the geosynchronous orbit above the equator because communication satellites using a given frequency must be separated from one another to avoid interference. The minimum separation depends on the power of the transmitters, but may require an angular separation of between *4* and *8* degrees. Thus, without further refinements, the entire *360*-degree circle above the equator can only hold *45* to *90* satellites.

What is the minimum number of satellites needed to cover the earth? Three. To see why, consider Figure 7.13, which illustrates the earth with three GEO satellites positioned around the equator with 120° separation. The figure illustrates how the signals from the three satellites cover the circumference. In the figure, the size of the earth and the distance of the satellites are drawn to scale.

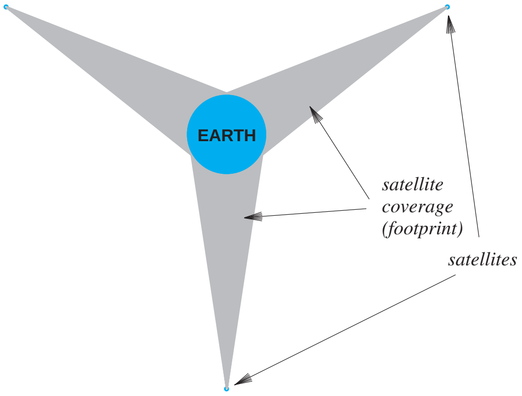


Figure 7.13 The signals from three GEO satellites are sufficient to cover the earth.

7.18 Low Earth Orbit (LEO) Satellites And Clusters

For communication, the primary alternative to GEO is known as *Low Earth Orbit* (*LEO*), which is defined as altitudes up to 2000 Kilometers. As a practical matter, a satellite must be placed above the fringe of the atmosphere to avoid the drag produced by encountering gases. Thus, LEO satellites are typically placed at altitudes of 500 Kilometers or higher. LEO offers the advantage of short delays (typically 1 to 4 milliseconds), but the disadvantage that the orbit of a satellite does not match the rotation of the earth. Thus, from an observer's point of view on the earth, an LEO satellite appears to move across the sky, which means a ground station must have an antenna that can rotate to track the satellite. Tracking is difficult because satellites move rapidly. The lowest altitude LEO satellites orbit the earth in approximately 90 minutes; higher LEO satellites require several hours.

The general technique used with LEO satellites is known as *clustering* or *array deployment*. A large group of LEO satellites are designed to work together. In addition to communicating with ground stations, a satellite in the group can also communicate with other satellites in the group. Members of the group stay in communication, and agree to forward messages, as needed. For example, consider what happens when a user in Europe sends a message to a user in North America. A ground station in Europe transmits the message to the satellite currently overhead. The cluster of satellites communicate to forward the message to the satellite in the cluster that is currently over a ground station in North America. Finally, the satellite currently over North America transmits the message to a ground station. To summarize:

> *A cluster of LEO satellites work together to forward messages. Members of the cluster must know which satellite is currently over a given area of the earth, and forward messages to the appropriate member for transmission to a ground station.*

7.19 Tradeoffs Among Media Types

The choice of medium is complex, and involves the evaluation of multiple factors. Items that must be considered include:

- Cost: materials, installation, operation, and maintenance
- Data rate: number of bits per second that can be sent
- Delay: time required for signal propagation or processing
- Affect on signal: attenuation and distortion
- Environment: susceptibility to interference and electrical noise
- Security: susceptibility to eavesdropping

7.20 Measuring Transmission Media

We have already mentioned the two most important measures of performance used to assess a transmission medium:

- *Propagation delay*: the time required for a signal to traverse the medium
- *Channel capacity*: the maximum data rate that the medium can support

Chapter 6 explains that in the 1920s, a researcher named Nyquist discovered a fundamental relationship between the bandwidth of a transmission system and its capacity to transfer data. Known as the *Nyquist Theorem*, the relationship provides a theoretical bound on the maximum rate at which data can be sent without considering the effect of noise. If a transmission system uses K possible signal levels and has an analog bandwidth B, the Nyquist Theorem states that the maximum data rate in bits per second, D, is:

$$D = 2B \log_2 K \tag{7.2}$$

7.21 The Effect Of Noise On Communication

The Nyquist Theorem provides an absolute maximum that cannot be achieved in practice. In particular, engineers have observed that a real communications system is subject to small amounts of electrical *noise* and that such noise makes it impossible to achieve the theoretical maximum transmission rate. In 1948, Claude Shannon extended Nyquist's work to specify the maximum data rate that could be achieved over a transmission system that experiences noise. The result, called *Shannon's Theorem*†, can be stated as:

$$C = B \log_2(1 + S/N) \tag{7.3}$$

where C is the effective limit on the channel capacity in bits per second, B is the hardware bandwidth, and S/N is the *signal-to-noise ratio*, the ratio of the average signal power divided by the average noise power.

As an example of Shannon's Theorem, consider a transmission medium that has a bandwidth of 1 KHz, an average signal power of 70 units, and an average noise power of 10 units. The channel capacity is:

$$C = 10^3 \times \log_2(1 + 7) = 10^3 \times 3 = 3{,}000 \textit{ bits per second}$$

†The result is also called the *Shannon–Hartley Law*.

The signal-to-noise ratio is often given in *decibels* (abbreviated *dB*), where a decibel is defined as a measure of the difference between two power levels. Figure 7.14 illustrates the measurement.

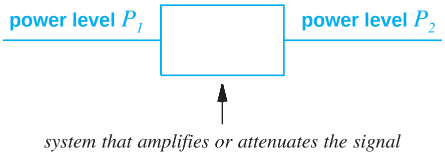


Figure 7.14 Power levels measured on either side of a system.

Once two power levels have been measured, the difference is expressed in decibels, defined as follows:

$$dB \ = \ 10 \log_{10} \left(\frac{P_2}{P_1} \right) \tag{7.4}$$

Using dB as a measure may seem usual, but has two interesting advantages. First, a negative dB value means that the signal has been *attenuated* (i.e., reduced), and a positive dB value means the signal has been *amplified*. Second, if a communications system has multiple parts arranged in a sequence, the decibel measures of the parts can be summed to produce a measure of the overall system.

The voice telephone system has a signal-to-noise ratio of approximately 30 dB and an analog bandwidth of approximately 3000 Hz. To convert signal-to-noise ratio dB into a simple fraction, divide by 10 and use the result as a power of 10 (i.e., $30/10 = 3$ and $10^3 = 1000$, so the signal-to-noise ratio is 1000). Shannon's Theorem can be applied to determine the maximum number of bits per second that can be transmitted across the telephone network:

$$C \ = \ 3000 \times \log_2(\ 1 \ + \ 1000 \ )$$

or approximately 30,000 bps. Engineers recognize this as a fundamental limit — faster transmission speeds will only be possible if the signal-to-noise ratio can be improved.

7.22 The Significance Of Channel Capacity

The theorems of Nyquist and Shannon, described above, have consequences for engineers who design data communication networks. Nyquist's work has provided an incentive to explore complex ways to encode bits on signals:

The Nyquist Theorem encourages engineers to explore ways to encode bits on a signal because a clever encoding allows more bits to be transmitted per unit time.

In some sense, Shannon's Theorem is more fundamental because it represents an absolute limit derived from the laws of physics. Much of the noise on a transmission line, for example, can be attributed to background radiation in the universe left over from the Big Bang. Thus,

Shannon's Theorem informs engineers that no amount of clever encoding can overcome the laws of physics that place a fundamental limit on the number of bits per second that can be transmitted in a real communications system.

7.23 Summary

A variety of transmission media exists that can be classified as guided / unguided or divided according to the form of energy used (electrical, light, or radio transmission). Electrical energy is used over wires. To protect against electrical interference, copper wiring can consist of twisted pairs or can be wrapped in a shield.

Light energy can be used over optical fiber or for point-to-point communication using infrared or lasers. Because it reflects from the boundary between the fiber and cladding, light stays in an optical fiber provided the angle of incidence is greater than the critical angle. As it passes along a fiber, a pulse of light disperses; dispersion is greatest in multimode fiber and least in single mode fiber. Single mode fiber is more expensive.

Wireless communication uses electromagnetic energy. The frequency used determines both the bandwidth and the propagation behavior; low frequencies follow the earth's surface, higher frequencies reflect from the ionosphere, and the highest frequencies behave like visible light by requiring a direct, unobstructed path from the transmitter to the receiver.

The chief nonterrestrial communication technology relies on satellites. The orbit of a GEO satellite matches the earth's rotation, but the high altitude incurs a delay measured in tenths of seconds. LEO satellites have low delay, and move across the sky quickly; clusters are used to relay messages.

The Nyquist Theorem gives a theoretical limit on the channel capacity of transmission media when no noise is present; Shannon's Theorem specifies the channel capacity in realistic situations where noise is present. The signal-to-noise ratio, a term in Shannon's Theorem, is often measured in decibels.

EXERCISES

7.1 What are the three energy types used when classifying physical media according to energy used?

7.2 What is the difference between guided and unguided transmission?

7.3 What three types of wiring are used to reduce interference from noise?

7.4 What happens when noise encounters a metal object?

7.5 Draw a diagram that illustrates the cross section of a coaxial cable.

7.6 Explain how twisted pair cable reduces the effect of noise.

7.7 Explain why light does not leave an optical fiber when the fiber is bent into an arc.

7.8 If you are installing computer network wiring in a new house, what category of twisted pair cable would you choose? Why?

7.9 List the three forms of optical fiber, and give the general properties of each.

7.10 What is dispersion?

7.11 What is the chief disadvantage of optical fiber as opposed to copper wiring?

7.12 What light sources and sensors are used with optical fibers?

7.13 Can laser communication be used from a moving vehicle? Explain.

7.14 What is the approximate conical angle that can be used with infrared technology?

7.15 What are the two broad categories of wireless communications?

7.16 Why might low-frequency electromagnetic radiation be used for communications? Explain.

7.17 If messages are sent from Europe to the United States using a GEO satellite, how long will it take for a message to be sent and a reply to be received?

7.18 List the three types of communications satellites, and give the characteristics of each.

7.19 What is propagation delay?

7.20 How many GEO satellites are needed to reach all populated areas on the earth?

7.21 If two signal levels are used, what is the data rate that can be sent over a coaxial cable that has an analog bandwidth of 6.2 MHz?

7.22 What is the relationship between bandwidth, signal levels, and data rate?

7.23 If a telephone system can be created with a signal-to-noise ratio of 40 dB and an analog bandwidth of 3000 Hz, how many bits per second could be transmitted?

7.24 If a system has an average power level of 100, an average noise level of 33.33, and a bandwidth of 100 MHz, what is the effective limit on channel capacity?

7.25 If a system has an input power level of 9000, and an output power level of 3000, what is the difference when expressed in dB?

This page is intentionally left blank.

Chapter Contents

8

Reliability And Channel Coding

8.1 Introduction

Chapters in this part of the text each present one aspect of data communications, the foundation for all computer networking. The previous chapter discusses transmission media, and points out the problem of electromagnetic noise. This chapter continues the discussion by examining errors that can occur during transmission and techniques that can be used to control errors.

The concepts presented here are fundamental to computer networking, and are used in communication protocols at many layers of the stack. In particular, the approaches to error control and techniques appear throughout the Internet protocols discussed in the fourth part of the text.

8.2 The Three Main Sources Of Transmission Errors

All data communications systems are susceptible to errors. Some of the problems are inherent in the physics of the universe, and some result either from devices that fail or from equipment that does not meet the engineering standards. Extensive testing can eliminate many of the problems that arise from poor engineering, and careful monitoring can identify equipment that fails. However, small errors that occur during transmission are more difficult to detect than complete failures, and much of computer networking focuses on ways to control and recover from such errors. There are three main categories of transmission errors:

- *Interference.* As Chapter 7 explains, electromagnetic radiation emitted from devices such as electric motors and background cosmic radiation cause noise that can disturb radio transmissions and signals traveling across wires.
- *Distortion.* All physical systems distort signals. As a pulse travels along an optical fiber, the pulse disperses. Wires have properties of capacitance and inductance that block signals at some frequencies while admitting signals at other frequencies. Simply placing a wire near a large metal object can change the set of frequencies that can pass through the wire. Similarly, metal objects can block some frequencies of radio waves, while passing others.
- *Attenuation.* As a signal passes across a medium, the signal becomes weaker. Engineers say that the signal has been *attenuated*. Thus, signals on wires or optical fibers become weaker over long distances, just as a radio signal becomes weaker with distance.

Shannon's Theorem suggests one way to reduce errors: increase the signal-to-noise ratio (either by increasing the signal or lowering noise). Even though mechanisms like shielded wiring can help lower noise, a physical transmission system is always susceptible to errors, and it may not be possible to increase the signal-to-noise ratio.

Although errors cannot be eliminated completely, many transmission errors can be detected. In some cases, errors can be corrected automatically. We will see that error detection adds overhead. Thus, all error handling is a tradeoff in which a system designer must decide whether a given error is likely to occur, and if so, what the consequences will be (e.g., a single bit error in a bank transfer can make a difference of over a million dollars, but a one bit error in an image is less important). The point is:

> *Although transmission errors are inevitable, error detection mechanisms add overhead. Therefore, a designer must choose exactly which error detection and compensation mechanisms will be used.*

8.3 Effect Of Transmission Errors On Data

Instead of examining physics and the exact cause of transmission errors, data communications focuses on the effect of errors on data. Figure 8.1 lists the three principal ways transmission errors affect data.

Although any transmission error can cause each of the possible data errors, the figure points out that an underlying transmission error often manifests itself as a specific data error. For example, extremely short duration interference, called a *spike*, is often the cause of a single bit error. Longer duration interference or distortion can produce burst errors. Sometimes a signal is neither clearly 1 nor clearly 0, but falls in an ambiguous region, which is known as an *erasure*.

Type Of Error	Description
Single Bit Error	A single bit in a block of bits is changed and all other bits in the block are unchanged (often results from very short-duration interference)
Burst Error	Multiple bits in a block of bits are changed (often results from longer-duration interference)
Erasure (Ambiguity)	The signal that arrives at a receiver is ambiguous and does not clearly correspond to either a logical 1 or a logical 0 (can result from distortion or interference)

Figure 8.1 The three types of data errors in a data communications system.

For a burst error, the *burst size*, or *length*, is defined as the number of bits from the start of the corruption to the end of the corruption. Figure 8.2 illustrates the definition.

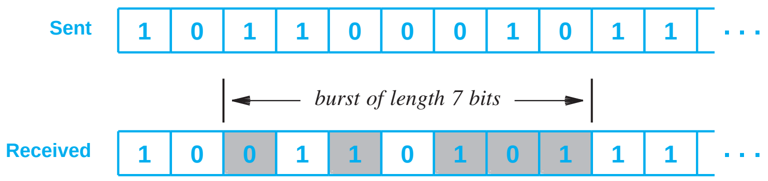


Figure 8.2 Illustration of a burst error with changed bits marked in gray.

8.4 Two Strategies For Handling Channel Errors

A variety of mathematical techniques have been developed that overcome data errors and increase reliability. Known collectively as *channel coding*, the techniques can be divided into two broad categories:

- Forward Error Correction (FEC) mechanisms
- Automatic Repeat reQuest (ARQ) mechanisms

The basic idea of forward error correction is straightforward: add additional information to data that allows a receiver to verify that data arrives correctly and to correct errors, if possible. Figure 8.3 illustrates the conceptual organization of a forward error correction mechanism.

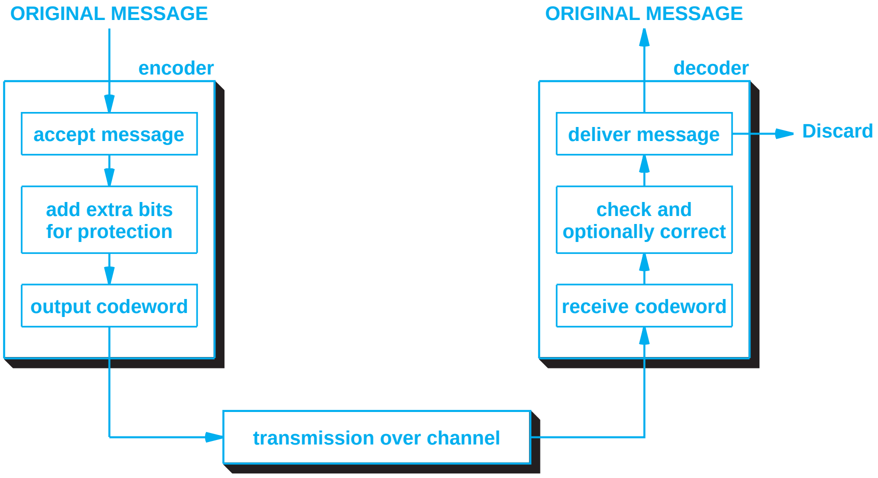


Figure 8.3 The conceptual organization of a forward error correction mechanism.

Basic *error detection mechanisms* allow a receiver to detect when an error has occurred; forward error correction mechanisms allow a receiver to determine exactly which bits have been changed and to compute correct values. The second approach to channel coding, known as an ARQ†, requires the cooperation of a sender — a sender and receiver exchange messages to ensure that all data arrives correctly.

8.5 Block And Convolutional Error Codes

The two types of forward error correction techniques are:

- *Block Error Codes*. A block code divides the data to be sent into a set of blocks, and attaches extra information known as *redundancy* to each block. The encoding for a given block of bits depends only on the bits themselves, not on bits that were sent earlier. Block error codes are *memoryless* in the sense that the encoding mechanism does not carry state information from one block of data to the next.

- *Convolutional Error Codes*. A convolutional code treats data as a series of bits, and computes a code over a continuous series. Thus, the code computed for a set of bits depends on the current input and some of the previous bits in the stream. Convolutional codes are said to be codes with *memory*.

†Section 8.15 introduces ARQ.

When implemented in software, convolutional error codes usually require more computation than block error codes. However, convolutional codes often have a higher probability of detecting problems.

8.6 An Example Block Error Code: Single Parity Checking

To understand how additional information can be used to detect errors, consider a *single parity checking* (*SPC*) mechanism. One form of SPC defines a block to be an 8-bit unit of data (i.e., a single *byte*). On the sending side, an encoder adds an extra bit, called a *parity bit* to each byte before transmission; a receiver removes the parity bit and uses it to check whether bits in the byte are correct.

Before parity can be used, the sender and receiver must be configured for either *even parity* or *odd parity*. When using even parity, the sender chooses a parity bit of 0 if the byte has an even number of 1 bits, and 1 if the byte has an odd number of 1 bits. The way to remember the definition is: even or odd parity specifies whether the 9 bits sent across a channel have an even or odd number of 1 bits. Figure 8.4 lists examples of data bytes and the value of the parity bit that is sent when using even or odd parity.

To summarize:

> *Single parity checking (SPC) is a basic form of channel coding in which a sender adds an extra bit to each byte to make an even (or odd) number of 1 bits and a receiver verifies that the incoming data has the correct number of 1 bits.*

Original Data	Even Parity	Odd Parity
0 0 0 0 0 0 0 0	0	1
0 1 0 1 1 0 1 1	1	0
0 1 0 1 0 1 0 1	0	1
1 1 1 1 1 1 1 1	0	1
1 0 0 0 0 0 0 0	1	0
0 1 0 0 1 0 0 1	1	0

Figure 8.4 Data bytes and the corresponding value of a single parity bit when using even parity or odd parity.

Single parity checking is a weak form of channel coding that can detect errors, but cannot correct them. Furthermore, parity mechanisms can only handle errors where an odd number of bits are changed. If one of the nine bits (including the parity bit) is changed during transmission, the receiver will declare that the incoming byte is invalid.

However, if a burst error occurs in which two, four, six, or eight bits change value, the receiver will incorrectly classify the incoming byte as valid.

8.7 The Mathematics Of Block Error Codes And (n,k) Notation

Observe that forward error correction takes as input a set of messages and inserts additional bits to produce an encoded version. Mathematically, we define the set of all possible messages to be a set of *datawords*, and define the set of all possible encoded versions to be a set of *codewords*. If a dataword contains k bits and r additional bits are added to form a codeword, we say that the result is an

$$(n,\, k)\ \ \textit{encoding scheme}$$

where $n = k + r$. The key to successful error detection lies in choosing a subset of the 2^n possible combinations that are valid codewords. The valid subset is known as a *codebook*.

As an example, consider single parity checking. The set of datawords consists of any possible combination of eight bits. Thus, $k = 8$ and there are 2^8 or 256 possible data words. The data sent consists of $n = 9$ bits, so there are 2^9 or 512 possibilities. However, only half of the 512 values form valid codewords.

Think of the set of all possible n-bit values and the valid subset that forms the codebook. If an error occurs during transmission, one or more of the bits in a codeword will be changed, which will either produce another valid codeword or an invalid combination. For example, in the single parity scheme discussed above, a change to a single bit of a valid codeword produces an invalid combination, but changing two bits produces another valid codeword. Obviously, we desire an encoding where an error produces an invalid combination. To generalize:

> *An ideal channel coding scheme is one where any change to bits in a valid codeword produces an invalid combination.*

8.8 Hamming Distance: A Measure Of A Code's Strength

No channel coding scheme is ideal — changing enough bits will always transform to a valid codeword. Thus, for a practical scheme, the question becomes: what is the minimum number of bits of a valid codeword that must be changed to produce another valid codeword?

To answer the question, engineers use a measure known as the *Hamming distance*, named after a theorist at Bell Laboratories who was a pioneer in the field of information theory and channel coding. Given two strings of n bits each, the Hamming distance is defined as the number of differences (i.e., the number of bits that must be changed to transform one bit string to the other). Figure 8.5 illustrates the definition.

d(000,001) = 1	d(000,101) = 2
d(101,100) = 1	d(001,010) = 2
d(110,001) = 3	d(111,000) = 3

Figure 8.5 Examples of Hamming distance for various pairs of 3-bit strings.

One way to compute the Hamming distance consists of taking the *exclusive or* (*xor*) between two strings and counting the number of 1 bits in the answer. For example, consider the Hamming distance between strings 110 and 011. The *xor* of the two strings is:

$$1\,1\,0 \quad \oplus \quad 0\,1\,1 \quad = \quad 1\,0\,1$$

which contains two 1 bits. Therefore, the Hamming distance between 011 and 101 is 2.

8.9 The Hamming Distance Among Strings In A Codebook

Recall that we are interested in whether errors can transform a valid codeword into another valid codeword. To measure such transformations, we compute the Hamming distance between all pairs of codewords in a given codebook. As a trivial example, consider odd parity applied to 2-bit data words. Figure 8.6 lists the four possible datawords, the four possible codewords that result from appending a parity bit, and the Hamming distances for pairs of codewords.

Dataword	Codeword
0 0	0 0 1
0 1	0 1 0
1 0	1 0 0
1 1	1 1 1

(a)

d(001,010) = 2	d(010,100) = 2
d(001,100) = 2	d(010,111) = 2
d(001,111) = 2	d(100,111) = 2

(b)

Figure 8.6 (a) The datawords and codewords for a single parity encoding of 2-bit data strings using odd parity, and (b) the Hamming distance for all pairs of codewords.

An entire set of codewords is known as a *codebook*. We use $d_{\min}$ to denote the *minimum Hamming distance* among pairs in a codebook. The concept gives a precise answer to the question of how many bit errors can cause a transformation from one valid codeword into another valid code word. In the single parity example of Figure 8.6, the set consists of the Hamming distance between each pair of codewords, and $d_{\min} = 2$. The definition means that there is at least one valid codeword that can be transformed into another valid codeword if two bit errors occur during transmission. The point is:

> *To find the minimum number of bit changes that can transform a valid codeword into another valid codeword, compute the minimum Hamming distance between all pairs in the codebook.*

8.10 The Tradeoff Between Error Detection And Overhead

For a set of codewords, a large value of $d_{\min}$ is desirable because the code is immune to more bit errors — if fewer than $d_{\min}$ bits are changed, the code can detect that error(s) occurred. Equation (8.1) specifies the relationship between $d_{\min}$ and e, the maximum number of bit errors that can be detected:

$$e = d_{\min} - 1 \tag{8.1}$$

The choice of error code is a tradeoff — although it detects more errors, a code with a higher value of $d_{\min}$ sends more redundant information than an error code with a lower value of $d_{\min}$. To measure the amount of overhead, engineers define a *code rate* that gives the ratio of a dataword size to the codeword size. Equation (8.2) defines the code rate, R, for an (n, k) error coding scheme:

$$R = \frac{k}{n} \tag{8.2}$$

8.11 Error Correction With Row And Column (RAC) Parity

We have seen how a channel coding scheme can detect errors. To understand how a code can be used to correct errors, consider an example. Assume a dataword consists of $k=12$ bits. Instead of thinking of the bits as a single string, imagine arranging them into an array of three rows and four columns, with a parity bit added for each row and for each column. Figure 8.7 illustrates the arrangement, which is known as a *Row And Column* (*RAC*) code. The example RAC encoding has $n=20$, which means that it is a $(20, 12)$ encoding.

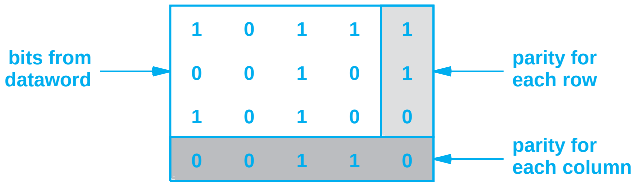


Figure 8.7 An example of row and column encoding with data bits arranged in a 3 × 4 array and an even parity bit added for each row and each column.

To see how error correction works, assume that when data bits in Figure 8.7 are transmitted, one bit is corrupted. The receiver arranges the bits that arrived into an array, recomputes the parity for each row and column, and compares the result to the value received. The changed bit causes two of the parity checks to fail, as Figure 8.8 illustrates.

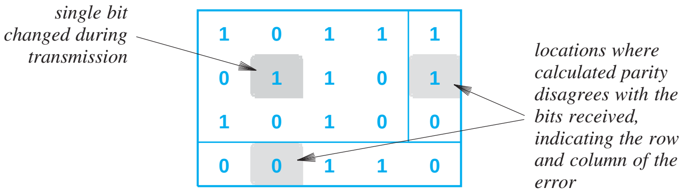

single bit changed during transmission
locations where calculated parity disagrees with the bits received, indicating the row and column of the error
1 0 1 1 1
0 1 1 0 1
1 0 1 0 0
0 0 1 1 0

Figure 8.8 Illustration of how a single-bit error can be corrected using a row and column encoding.

As the figure illustrates, a single bit error will cause two calculated parity bits to disagree with the parity bit received. The two disagreements correspond to the row and column of the error. A receiver uses the calculated parity bits to determine exactly which data bit is in error, and then corrects the data bit. Thus, a RAC encoding can correct any error that changes a single data bit.

What happens to a RAC code if an error changes more than one bit in a given block? RAC can only correct single-bit errors. In cases of multi-bit errors where an odd number of bits are changed, a RAC encoding will be able to detect, but not correct, the problem.

To summarize:

> *A Row And Column (RAC) encoding allows a receiver to correct any single bit error and to detect errors in which an odd number of bits are changed.*

8.12 The 16-Bit Checksum Used In The Internet

A particular channel coding scheme plays a key role in the Internet. Known as the *Internet checksum*, the code consists of a 16-bit 1s complement checksum. The Internet checksum does not impose a fixed size on a dataword. Instead, the algorithm allows a message to be arbitrarily long, and computes a checksum over the entire message. In essence, the Internet checksum treats data in a message as a series of 16-bit integers, as Figure 8.9 illustrates.

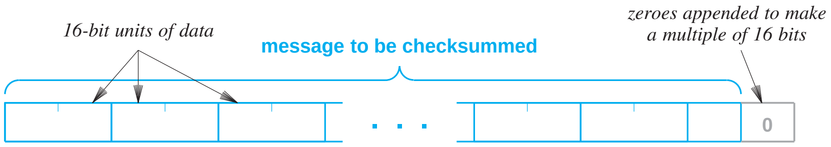

16-bit units of data
message to be checksummed
zeroes appended to make a multiple of 16 bits
. . .
0

Figure 8.9 The Internet checksum divides data into 16-bit units, appending zeroes if the data is not an exact multiple of 16 bits.

To compute a checksum, a sender adds the numeric values of the 16-bit integers, and transmits the result. To validate the message, a receiver performs the same computation. Algorithm 8.1 gives the details of the computation.

Algorithm 8.1

Given:

A message, M, of arbitrary length

Compute:

A 16-bit 1s complement checksum, C, using 32-bit arithmetic

Method:

```
Pad M with zero bits to make an exact multiple of 16 bits
Set a 32-bit checksum integer, C, to 0;
for ( each 16-bit group in M ) {
    Treat the 16 bits as an integer and add to C;
}
Extract the high-order 16 bits of C and add them to C;
The inverse of the low-order 16 bits of C is the checksum;
If the checksum is zero, substitute the all 1s form of zero.
```

Algorithm 8.1 The 16-bit checksum algorithm used in the Internet protocols.

The key to understanding the algorithm is to realize that the checksum is computed in 1s complement arithmetic instead of the 2s complement arithmetic found on most computers, and uses 16 bit integers instead of 32 or 64 bit integers. Thus, the algorithm is written to use 32-bit 2s complement arithmetic to perform a 1s complement computation. During the *for* loop, the addition may overflow. Thus, following the loop, the algorithm adds the overflow (the high-order bits) back into the sum. Figure 8.10 illustrates the computation.

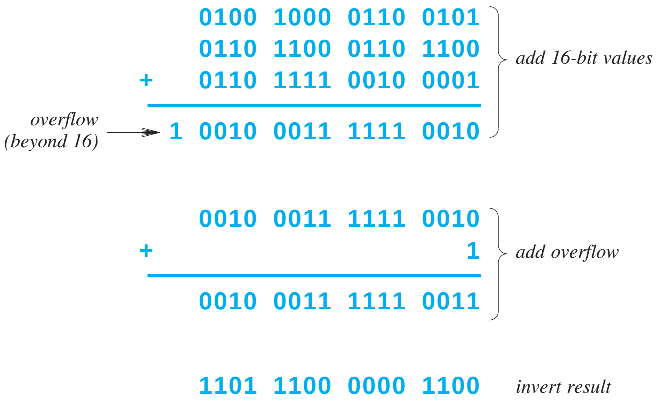


Figure 8.10 An example of Algorithm 8.1 applied to six octets of data.

Why is a checksum computed as the arithmetic inverse of the sum instead of the sum? The answer is efficiency: a receiver can apply the same checksum algorithm as the sender, but can include the checksum itself. Because it contains the arithmetic inverse of the total, adding the checksum to the total will produce zero. Thus, a receiver includes the checksum in the computation, and then tests to see if the resulting sum is zero.

A final detail of 1s complement arithmetic arises in the last step of the algorithm. Ones complement arithmetic has two forms of zero: all zeroes and all ones. The Internet checksum uses the all-ones form to indicate that a checksum was computed and the value of the checksum is zero; the Internet protocols use the all-zeroes form to indicate that no checksum was computed.

8.13 Cyclic Redundancy Codes (CRCs)

A form of channel coding known as a *Cyclic Redundancy Code* (*CRC*) is used in high-speed data networks. CRC codes have three key properties that make them important, as Figure 8.11 summarizes.

Arbitrary Length Message	As with a checksum, the size of a dataword is not fixed, which means a CRC can be applied to an arbitrary length message
Excellent Error Detection	Because the value computed depends on the sequence of bits in a message, a CRC provides excellent error detection capability
Fast Hardware Implementation	Despite its sophisticated mathematical basis, a CRC computation can be carried out extremely fast by hardware

Figure 8.11 The three key aspects of a CRC that make it important in data networking.

The term *cyclic* is derived from a property of the codewords: a circular shift of the bits of any codeword produces another codeword. Figure 8.12 illustrates a $(7, 4)$ cyclic redundancy code that was introduced by Hamming.

Dataword	Codeword
0000	0000 000
0001	0001 011
0010	0010 110
0011	0011 101
0100	0100 111
0101	0101 100
0110	0110 001
0111	0111 010
1000	1000 101
1001	1001 110
1010	1010 011
1011	1011 000
1100	1100 010
1101	1101 001
1110	1110 100
1111	1111 111

Figure 8.12 An example $(7, 4)$ cyclic redundancy code.

CRC codes have been studied extensively, and a variety of mathematical explanations and computational techniques have been produced. The descriptions seem so disparate that it is difficult to understand how they can all refer to the same concept. Principal views include:

- *Mathematicians* explain a CRC computation as the remainder from a division of two polynomials with binary coefficients, one representing the message and another representing a fixed divisor.
- *Theoretical computer scientists* explain a CRC computation as the remainder from a division of two binary numbers, one representing the message and the other representing a fixed divisor.
- *Cryptographers* explain a CRC computation as a mathematical operation in a Galois field of order 2, written GF(2).
- *Computer programmers* explain a CRC computation as an algorithm that iterates through a message and uses table lookup to obtain an additive value for each step.
- *Hardware architects* explain a CRC computation as a small hardware pipeline unit that takes as input a sequence of bits from a message and produces a CRC without using division or iteration.

As an example of the views above, consider the division of binary numbers under the assumption of no carries. Because no carries are performed, subtraction is performed modulo two, and we can think of subtraction as being replaced by *exclusive or*. Figure 8.13 illustrates the computation by showing the division of 1010, which represents a message, by a constant chosen for a specific CRC, 1011.

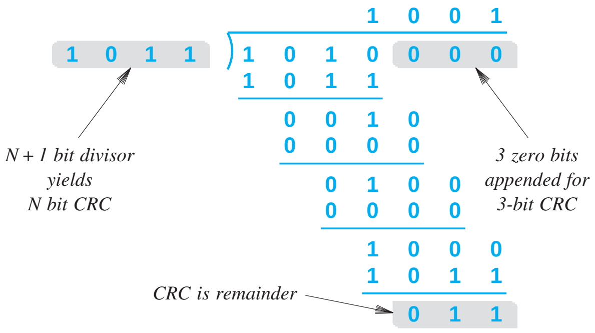


Figure 8.13 Illustration of a CRC computation viewed as the remainder of a binary division with no carries (i.e., where subtraction becomes exclusive or).

To understand how mathematicians can view the above as a polynomial division, think of each bit in a binary number as the coefficient of a term in a polynomial. For example, we can think of the divisor in Figure 8.13, *1011*, as coefficients in the following polynomial:

$$1 \times x^3 + 0 \times x^2 + 1 \times x^1 + 1 \times x^0 \quad = \quad x^3 + x + 1$$

Similarly, the dividend in Figure 8.13, *1010000*, represents the polynomial:

$$x^6 + x^4$$

We use the term *generator polynomial* to describe a polynomial that corresponds to a divisor. The selection of a generator polynomial is key to creating a CRC with good error detection properties. Therefore, much mathematical analysis has been conducted on generator polynomials. We know, for example, that an ideal polynomial is irreducible (i.e., can only be divided evenly by itself and 1) and that a polynomial with more than one non-zero coefficient can detect all single-bit errors.

8.14 An Efficient Hardware Implementation Of CRC

The hardware needed to compute a CRC is surprisingly straightforward. CRC hardware is arranged as a shift register with *exclusive or* (*xor*) gates between some of the bits. When computing a CRC, the hardware is initialized so that all bits in the shift register are zero. Then data bits are shifted in, one at a time. Once the last data bit has been shifted in, the value in the shift register is the CRC.

The shift register operates once per input bit, and all parts operate at the same time, like the production line in a factory. During a cycle, each stage of the register either accepts the bit directly from the previous stage, or accepts the output from an *xor* operation. The *xor* always involves the bit from the previous stage and a feedback bit from a later stage.

Figure 8.14 illustrates the hardware needed for the 3-bit CRC computation from Figure 8.13. Because an *xor* operation and *shift* can each be performed at high speed, the arrangement can be used for high-speed computer networks.

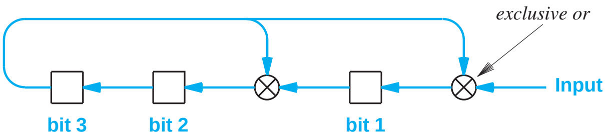


Figure 8.14 A hardware unit to compute a 3-bit CRC for $x^3 + x^1 + 1$.

8.15 Automatic Repeat Request (ARQ) Mechanisms

An Automatic Repeat reQuest (ARQ) approach to error correction requires a sender and receiver to communicate metainformation. That is, whenever one side sends a message to another, the receiving side sends a short *acknowledgement* message back. For example, if *A* sends a message to *B*, *B* sends an acknowledgement back to *A*. Once it receives an acknowledgement, *A* knows that the message arrived correctly. If no acknowledgement is received after T time units, A assumes the message was lost and *retransmits* a copy.

ARQ is especially useful in cases where the underlying system provides error detection, but not error correction. For example, many computer networks use a CRC to detect transmission errors. In such cases, an ARQ scheme can be added to guarantee delivery — if a transmission error occurs, the receiver discards the message and the sender retransmits another copy.

Chapter 25 will discuss the details of an Internet protocol that uses the ARQ approach. In addition to showing how the timeout-and-retransmission paradigm works in practice, the chapter explains how the sender and receiver identify the data being acknowledged, and discusses how long a sender waits before retransmitting.

8.16 Summary

Physical transmission systems are susceptible to interference, distortion, and attenuation, all of which can cause errors. Transmission errors can result in single-bit errors or burst errors, and erasures can occur whenever a received signal is ambiguous (i.e., neither clearly 1 nor clearly 0). To control errors, data communications systems employ a forward error correction mechanism or use an automatic repeat request (ARQ) technique.

Forward error correction arranges for a sender to add redundant bits to the data and encode the result before transmission across a channel, and arranges for a receiver to decode and check incoming data. A coding scheme is (n, k) if a dataword contains k bits and a codeword contains n bits.

One measure of an encoding assesses the chance that an error will change a valid codeword into another valid codeword. The minimum Hamming distance provides a precise measure.

Simplistic block codes, such as a single parity bit added to each byte, can detect an odd number of bit errors, but cannot detect an even number of bit changes. A Row And Column (RAC) code can correct single-bit errors, and can detect any multi-bit error in which an odd number of bits are changed in a block.

The 16-bit checksum used in the Internet can be used with an arbitrary size message. The checksum algorithm divides a message into 16-bit blocks, and computes the arithmetic inverse of the 1s-complement sum of the blocks; the overflow is added back into the checksum.

Cyclic redundancy codes (CRCs) are used in high-speed data networks because a CRC accepts a message of arbitrary length, provides extremely good error detection, and has an efficient hardware implementation. CRC techniques have a mathematical basis, and have been studied extensively. A CRC computation can be viewed as computing the remainder of a binary division, computing the remainder of a polynomial division, or an operation using Galois field theory. The hardware to perform a CRC computation uses a shift register and *exclusive or* operations.

EXERCISES

8.1 How do transmission errors affect data?

8.2 List and explain the three main sources of transmission errors.

8.3 What is a codeword, and how is it used in forward error correction?

8.4 In a burst error, how is burst length measured?

8.5 What does an ideal channel coding scheme achieve?

8.6 Give an example of a block error code used with character data.

8.7 Compute the Hamming distance for the following pairs: $(0000, 0001)$, $(0101, 0001)$, $(1111, 1001)$, and $(0001, 1110)$.

8.8 Define the concept of *Hamming distance*.

8.9 Explain the concept of *code rate*. Is a high code rate or low code rate desirable?

8.10 How does one compute the minimum number of bit changes that can transform a valid codeword into another valid codeword?

8.11 What can a RAC scheme achieve that a single parity bit scheme cannot?

8.12 Generate a RAC parity matrix for a $(20, 12)$ coding of the dataword *100011011111*.

8.13 What are the characteristics of a CRC?

8.14 Write a computer program that computes a 16-bit Internet checksum.

8.15 List and explain the function of each of the two hardware building blocks used to implement CRC computation.

8.16 Show the division of 10010101010 by 10101.

8.17 Express the two values in the previous exercise as polynomials.

8.18 Write a computer program that implements the (7,4) cyclic redundancy code in Figure 8.12.

Chapter Contents

9

Transmission Modes

9.1 Introduction

Chapters in this part of the text cover fundamental concepts that underlie data communications. This chapter continues the discussion by focusing on the ways data is transmitted. The chapter introduces common terminology, explains the advantages and disadvantages of parallelism, and discusses the important concepts of synchronous and asynchronous communication. Later chapters show how the ideas presented here are used in networks throughout the Internet.

9.2 A Taxonomy Of Transmission Modes

We use the term *transmission mode* to refer to the manner in which data is sent over the underlying medium. Transmission modes can be divided into two fundamental categories:

- Serial — one bit is sent at a time
- Parallel — multiple bits are sent at the same time

As we will see, serial transmission is further categorized according to timing of transmissions. Figure 9.1 gives an overall taxonomy of the transmission modes discussed in the chapter.

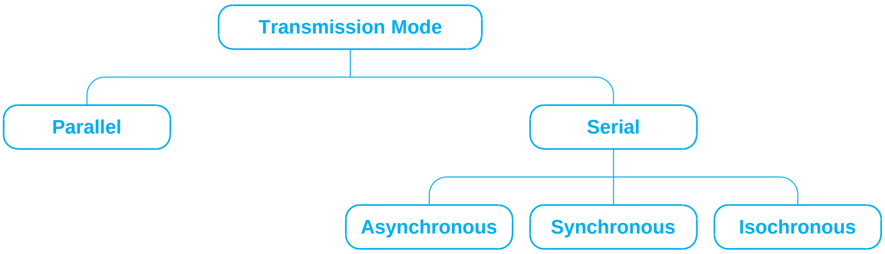


Figure 9.1 A taxonomy of transmission modes.

9.3 Parallel Transmission

The term *parallel transmission* refers to a transmission mechanism that transfers multiple data bits at the same time over separate media. In general, parallel transmission is used with a wired medium that uses multiple, independent wires. Furthermore, the signals on all wires are synchronized so that a bit travels across each of the wires at precisely the same time. Figure 9.2 illustrates the concept, and shows why engineers use the term *parallel* to characterize the wiring.

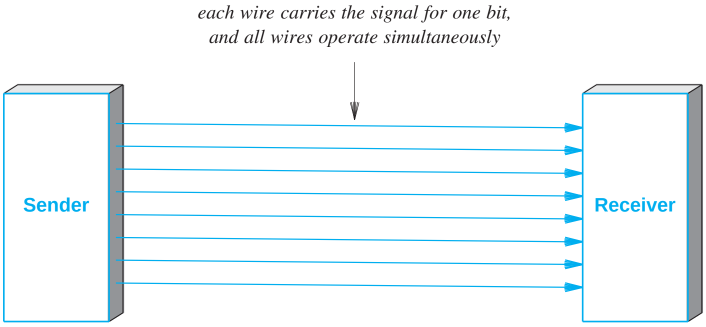

each wire carries the signal for one bit,
and all wires operate simultaneously
Sender
Receiver

Figure 9.2 Illustration of parallel transmission that uses 8 wires to send 8 bits at the same time.

The figure omits two important details. First, in addition to the parallel wires that each carry data, a parallel interface usually contains other wires that allow the sender and receiver to coordinate. Second, to make installation and troubleshooting easy, the wires for a parallel transmission system are placed in a single physical cable. Thus, one expects to see a single, large cable connecting a sender and receiver rather than a set of independent physical wires.

A parallel mode of transmission has two chief advantages:

- *High Throughput*. Because it can send N bits at the same time, a parallel interface can send N bits in the same time it takes a serial interface to send one bit.
- *Match To Underlying Hardware*. Internally, computer and communication hardware uses parallel circuitry. Thus, a parallel interface matches the internal hardware well.

9.4 Serial Transmission

The alternative to parallel transmission, known as *serial transmission*, sends one bit at a time. With the emphasis on speed, it may seem that anyone designing a data communications system would choose parallel transmission. However, most communications systems use serial mode. There are three main reasons. First, a serial transmission system costs less because fewer physical wires are needed and intermediate electronic components are less expensive. Second, parallel systems require each wire to be exactly the same length (even a difference of millimeters can cause problems). Third, at extremely high data rates, signals on parallel wires can cause electromagnetic noise that interferes with signals on other wires.

To use serial transmission, the sender and receiver must contain a small amount of hardware that converts data from the parallel form used in the device to the serial form used on the wire. Figure 9.3 illustrates the configuration.

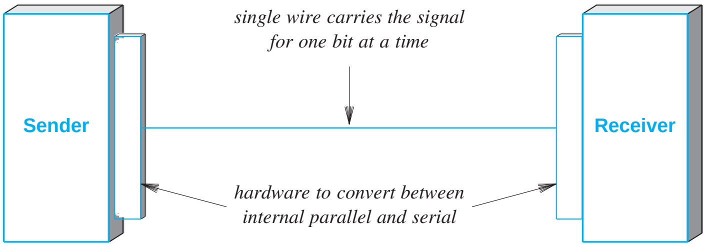

single wire carries the signal
for one bit at a time
Sender
Receiver
hardware to convert between
internal parallel and serial

Figure 9.3 Illustration of a serial transmission mode.

The hardware needed to convert data between an internal parallel form and a serial form can be straightforward or complex, depending on the type of serial communication mechanism. In the simplest case, a single chip that is known as a *Universal Asynchronous Receiver and Transmitter* (*UART*) performs the conversion. A related chip, *Universal Synchronous-Asynchronous Receiver and Transmitter* (*USART*) handles conversion for synchronous networks.

9.5 Transmission Order: Bits And Bytes

Serial transmission mode introduces an interesting question: when sending bits, which bit should be sent across the medium first? For example, consider an integer. Should a sender transmit the *Most Significant Bit* (*MSB*) or the *Least Significant Bit* (*LSB*) first?

Engineers use the term *little-endian* to describe a system that sends the LSB first, and the term *big-endian* to describe a system that sends the MSB first. Either form can be used, but the sender and receiver must agree.

Interestingly, the order in which bits are transmitted does not settle the entire question of transmission order. Data in a computer is divided into bytes, and each byte is further divided into bits (typically 8 bits per byte). Thus, it is possible to choose a byte order and a bit order independently. For example, Ethernet technology specifies that data is sent byte big-endian and bit little-endian. Figure 9.4 illustrates the order in which Ethernet sends bits from a 32-bit quantity.

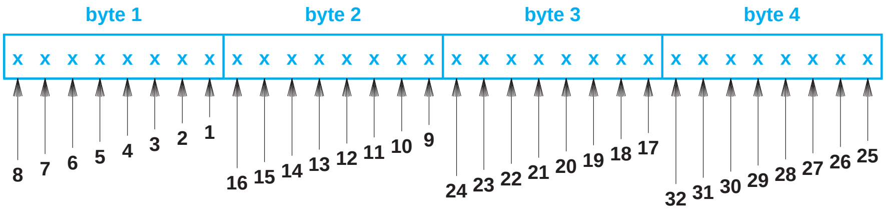


Figure 9.4 Illustration of byte big-endian, bit little-endian order in which the least-significant bit of the most-significant byte is sent first.

9.6 Timing Of Serial Transmission

Serial transmission mechanisms can be divided into three broad categories, depending on how transmissions are spaced in time:

- *Asynchronous* transmission can occur at any time, with an arbitrary delay between the transmission of two data items.
- *Synchronous* transmission occurs continuously with no gap between the transmission of two data items.
- *Isochronous* transmission occurs at regular intervals with a fixed gap between the transmission of two data items.

9.7 Asynchronous Transmission

A transmission system is classified as *asynchronous* if the system allows the physical medium to be idle for an arbitrary time between two transmissions. The asynchronous style of communication is well-suited to applications that generate data at random (e.g., a user typing on a keyboard, or a user who clicks on a link to obtain a web page, reads for awhile, and then clicks on a link to obtain another page).

The disadvantage of asynchrony arises from the lack of coordination between sender and receiver — while the medium is idle, a receiver cannot know how long the medium will remain idle before more data arrives. Thus, asynchronous technologies usually arrange for a sender to transmit a few extra bits before each data item to inform the receiver that a data transfer is starting. The extra bits allow the receiver's hardware to synchronize with the incoming signal. In some asynchronous systems, the extra bits are known as a *preamble*; in others, the extra bits are known as *start bits*. To summarize:

> *Because it permits a sender to remain idle an arbitrarily long time between transmissions, an asynchronous transmission mechanism sends extra information before each transmission that allows a receiver to synchronize with the signal.*

9.8 RS-232 Asynchronous Character Transmission

As an example of asynchronous communication, consider the transfer of characters across copper wires between a computer and a device such as a keyboard. An asynchronous communication technology standardized by the *Electronic Industries Alliance* (*EIA*) has become the most widely accepted for character communication. Known as *RS-232-C* and commonly abbreviated *RS-232*†, the EIA standard specifies the details of the physical connection (e.g., the connection must be less than 50 feet long), electrical details (e.g., the voltage ranges from –15 volts to +15 volts), and the line coding (e.g., negative voltage corresponds to logical 1 and positive voltage corresponds to logical 0).

Because it is designed for use with devices such as keyboards, the RS-232 standard specifies that each data item represents one character. The hardware can be configured to control the exact number of bits per second and to send seven-bit or eight-bit characters. Although a sender can delay arbitrarily long before sending a character, once transmission begins, a sender transmits all bits of the character one after another with no delay between them. When it finishes transmission, the sender leaves the wire with a negative voltage (corresponding to logical *1*) until another character is ready for transmission.

How does a receiver know where a new character starts? RS-232 specifies that a sender transmit an extra *0* bit (called a *start bit*) before transmitting the bits of a charac-

†Although the later RS-449 standard provides slightly more functionality, most engineers still use the original name.

ter. Furthermore, RS-232 specifies that a sender must leave the line idle between characters for at least the time required to send one bit. Thus, one can think of a phantom *1* bit appended to each character. In RS-232 terminology, the phantom bit is called a *stop bit*. Figure 9.5 illustrates how voltage varies when a start bit, eight bits of a character, and a stop bit are sent.

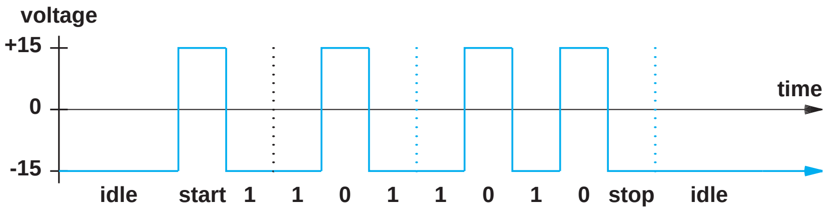


Figure 9.5 Illustration of voltage during transmission of an 8-bit character when using RS-232.

To summarize:

> *The RS-232 standard used for asynchronous, serial communication over short distances precedes each character with a start bit, sends each bit of the character, and follows each character with an idle period at least one bit long (stop bit).*

9.9 Synchronous Transmission

The chief alternative to asynchronous transmission is known as *synchronous transmission*. At the lowest level, a synchronous mechanism transmits bits of data continually, with no idle time between bits. That is, after transmitting the final bit of one data byte, the sender transmits a bit of the next data byte.

The chief advantage of a synchronous mechanism arises because the sender and receiver constantly remain synchronized, which means less synchronization overhead. To understand the overhead, compare the transmission of 8-bit characters on an asynchronous system as illustrated in Figure 9.5 and on a synchronous system as illustrated in Figure 9.6. Each character sent using RS-232 requires an extra start bit and stop bit, meaning that each 8-bit character requires a minimum of 10 bit times, even if no idle time is inserted. On a synchronous system, each character is sent without start or stop bits.

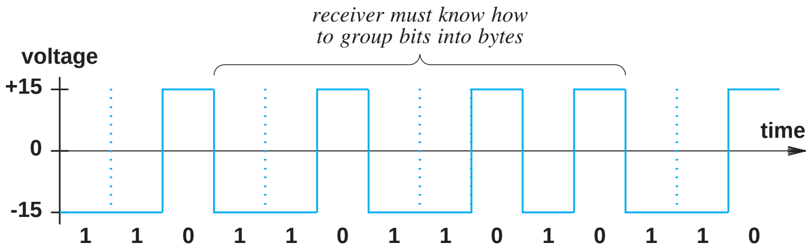


Figure 9.6 Illustration of synchronous transmission where the first bit of a byte immediately follows the last bit of the previous byte.

The point is:

> *When compared to synchronous transmission an asynchronous RS-232 mechanism has 25% more overhead per character.*

9.10 Bytes, Blocks, And Frames

If the underlying synchronous mechanism must send bits continually, what happens if a sender does not have data ready to send at all times? The answer lies in a technique known as *framing*: an interface is added to a synchronous mechanism that accepts and delivers a *block* of bytes known as a *frame*. To ensure that the sender and receiver stay synchronized, a frame starts with a special sequence of bits. Furthermore, most synchronous systems include a special *idle sequence* (or *idle byte*) that is transmitted when the sender has no data to send. Figure 9.7 illustrates the concept.

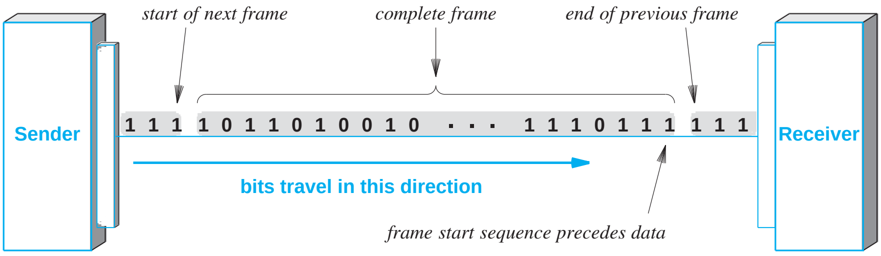


Figure 9.7 Illustration of framing on a synchronous transmission system.

The consequence of framing can be summarized:

> *Although the underlying mechanism transmits bits continuously, the use of an idle sequence and framing permits a synchronous transmission mechanism to provide a byte-oriented interface and to allow idle gaps between blocks of data.*

9.11 Isochronous Transmission

The third type of serial transmission system does not provide a new underlying mechanism. Instead, it can be viewed as an important way to use synchronous transmission. Known as *isochronous transmission*†, the system is designed to provide steady bit flow for multimedia applications that contain voice or video. Delivering such data at a steady rate is essential because variations in delay, which are known as *jitter*, can disrupt reception (i.e., cause pops or clicks in audio or make video freeze for a short time).

Instead of using the presence of data to drive transmission, an isochronous network is designed to accept and send data at a fixed rate, *R*. In fact, the interface to the network is such that data *must* be handed to the network for transmission at exactly *R* bits per second. For example, an isochronous mechanism designed to transfer voice operates at a rate of 64,000 bits per second. A sender must generate digitized audio continuously, and a receiver must be able to accept and play the stream.

An underlying network can use framing and may choose to transmit extra information along with data. However, to be isochronous, a system must be designed so the sender and receiver see a continuous stream of data, with no extra delays at the start of a frame. Thus, an isochronous network that provides a data rate of *R* bits per second usually has an underlying synchronous mechanism that operates at slightly more than *R* bits per second.

9.12 Simplex, Half-Duplex, And Full-Duplex Transmission

A communications channel is classified as one of three types, depending on the direction of transfer:

- Simplex
- Full-Duplex
- Half-Duplex

Simplex. A *simplex* mechanism is the easiest to understand. As the name implies, a simplex mechanism can only transfer data in a single direction. For example, a single optical fiber acts as a simplex transmission mechanism because the fiber has a transmit-

†Isochronous is pronounced *eye-sock'-run-us.*

ting device (i.e., an LED or laser) at one end and a receiving device (i.e., a photosensitive receptor) at the other. Simplex transmission is analogous to broadcast radio or television. Figure 9.8(a) illustrates simplex communication.

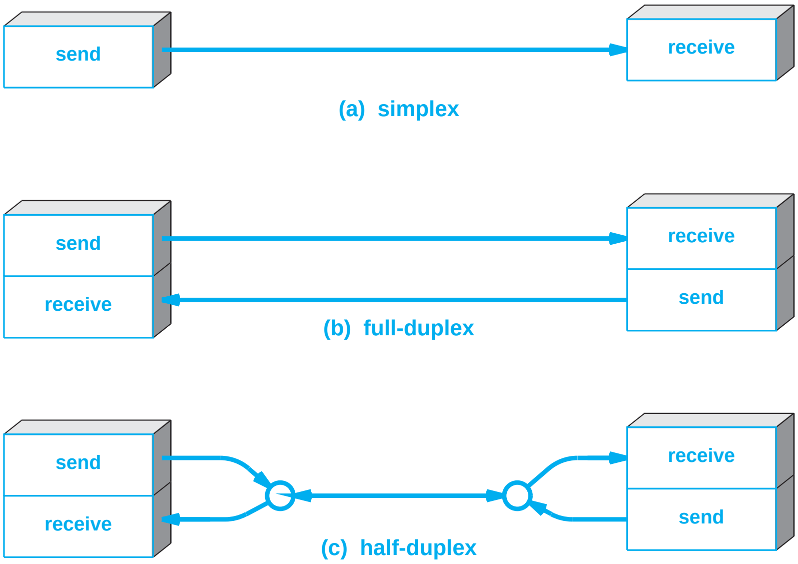


Figure 9.8 Illustration of the three modes of operation.

Full-Duplex. A *full-duplex* mechanism is also straightforward: the underlying system allows transmission in two directions simultaneously. Typically a full-duplex mechanism consists of two *simplex* mechanisms, one carrying information in each direction, as Figure 9.8(b) illustrates. For example, a pair of optical fibers can be used to provide full-duplex communication by running the two in parallel and arranging to send data in opposite directions. Full duplex communication is analogous to a voice telephone conversation in which a participant can speak even if they are able to hear background music at the other end.

Half-Duplex. A *half-duplex* mechanism involves a shared transmission medium. The shared medium can be used for communication in each direction, but the communication cannot proceed simultaneously. Thus, half-duplex communication is analogous to using walkie-talkies where only one side can transmit at a time. An additional mechanism is needed at each end of a half-duplex communication that coordinates transmission to ensure that only one side transmits at a given time. Figure 9.8(c) illustrates half-duplex communication.

9.13 DCE And DTE Equipment

The terms *Data Communications Equipment* (*DCE*) and *Data Terminal Equipment* (*DTE*) were originally created by AT&T to distinguish between the communications equipment owned by the phone company and the *terminal* equipment owned by a subscriber.

The terminology persists: if a business leases a data circuit from a phone company, the phone company installs DCE equipment at the business, and the business purchases DTE equipment that attaches to the phone company's equipment.

From an academic point of view, the important concept behind the DCE-DTE distinction is not ownership of the equipment. Instead, it lies in the ability to define an arbitrary interface for a user. For example, if the underlying network uses synchronous transmission, the DCE equipment can provide either a synchronous or isochronous interface to the user's equipment. Figure 9.9 illustrates the conceptual organization†.

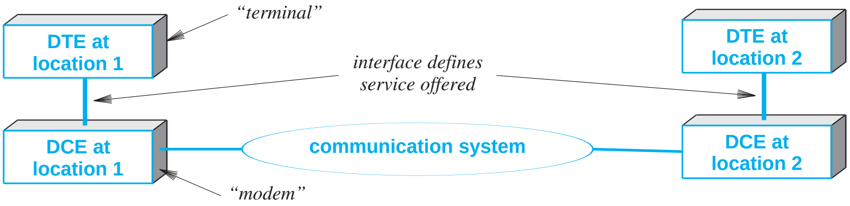


Figure 9.9 Illustration of Data Communications Equipment and Data Terminal Equipment providing a communication service between two locations.

Several standards exist that specify a possible interface between DCE and DTE. For example, the RS-232 standard described in this chapter and the RS-449 standard designed as a replacement can each be used. In addition, a standard known as *X.21* is available.

9.14 Summary

Communications systems use parallel or serial transmission. A parallel system has multiple wires, and at any time, each wire carries the signal for one bit. Thus, a parallel transmission system with K wires can send K bits at the same time. Although parallel communication offers higher speed, most communications systems use lower-cost serial mechanisms that send one bit at a time.

†Note: the terms DCE and DTE are also used to distinguish between two types of connectors, even if the equipment is not owned by a phone company (e.g., the connector on a PC and the connector on an external modem).

Serial communication requires a sender and receiver to agree on timing and the order in which bits are sent. Transmission order refers to whether the most-significant or least-significant bit is sent first and whether the most-significant or least-significant byte is sent first.

The three types of timing are: asynchronous, in which transmission can occur at any time and the communications system can remain idle between transmissions, synchronous, in which bits are transmitted continually and data is grouped into frames, and isochronous, in which transmission occurs at regular intervals with no extra delay at frame boundaries.

A communications system can be simplex, full-duplex, or half-duplex. A simplex mechanism sends data in a single direction. A full-duplex mechanism transfers data in two directions simultaneously, and a half-duplex mechanism allows two-way transfer, but only allows a transfer in one direction at a given time.

The distinction between Data Communications Equipment and Data Terminal Equipment was originally devised to denote whether a provider or a subscriber owned equipment. The key concept arises from the ability to define an interface for a user that offers a different service than the underlying communications system.

EXERCISES

9.1 What are the advantages of parallel transmission? What is the chief disadvantage?

9.2 Describe the difference between serial and parallel transmission.

9.3 What is the difference between synchronous and asynchronous transmission?

9.4 When transmitting a 32-bit 2's complement integer in big-endian order, when is the sign bit transmitted?

9.5 What is a start bit, and with which type of serial transmission is a start bit used?

9.6 Which type (or types) of serial transmission is appropriate for video transmission? For a keyboard connection to a computer?

9.7 When two humans hold a conversation, do they use simplex, half-duplex, or full-duplex transmission?

9.8 When using a synchronous transmission scheme, what happens when a sender does not have data to send?

9.9 Use the Web to find the definition of the DCE and DTE pinouts used on a DB-25 connector. (Hint: pins 2 and 3 are transmit or receive.) On a DCE type connector, does pin 2 transmit or receive?

9.10 Is a modem classified as DTE or DCE?

Chapter Contents

10

Modulation And Modems

10.1 Introduction

Chapters in this part of the text each cover one aspect of data communications. Previous chapters discuss information sources, explain how a signal can represent information, and describe forms of energy used with various transmission media.

This chapter continues the discussion of data communications by focusing on the use of high-frequency signals to carry information. The chapter discusses how information is used to change a high-frequency electromagnetic wave, explains why the technique is important, and describes how analog and digital inputs are used. Later chapters extend the discussion by explaining how the technique can be used to devise a communications system that transfers multiple, independent streams of data over a shared transmission medium simultaneously.

10.2 Carriers, Frequency, And Propagation

Many long distance communications systems use a continuously oscillating electromagnetic wave called a *carrier*. The system makes small changes to the carrier that represent information being sent. To understand why carriers are important, recall from Chapter 7 that the frequency of electromagnetic energy determines how the energy propagates. One motivation for the use of carriers arises from the desire to select a frequency that will propagate well, independent of the rate that data is being sent.

10.3 Analog Modulation Schemes

We use the term *modulation* to refer to changes made in a carrier according to the information being sent. Conceptually, modulation takes two inputs, a carrier and a signal, and generates a modulated carrier as output, as Figure 10.1 illustrates.

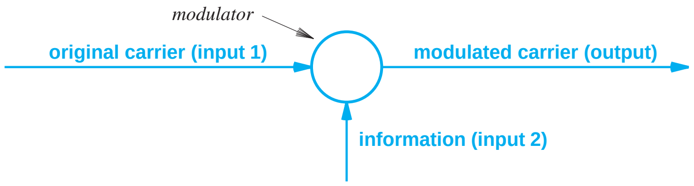


Figure 10.1 The concept of modulation with two inputs.

In essence, a sender must change one of the fundamental characteristics of the wave. Thus, there are three primary techniques that modulate an electromagnetic carrier according to a signal:

- Amplitude modulation
- Frequency modulation
- Phase shift modulation

The first two methods of modulation are the most familiar and have been used extensively. Indeed, they did not originate with computer networks — they were devised and used for broadcast radio, and are also used for broadcast television.

10.4 Amplitude Modulation

A technique known as *amplitude modulation* varies the amplitude of a carrier in proportion to the information being sent (i.e., according to a signal). The carrier continues oscillating at a fixed frequency, but the amplitude of the wave varies. Figure 10.2 illustrates an unmodulated carrier wave, an analog information signal, and the resulting amplitude modulated carrier.

Amplitude modulation is easy to understand because only the amplitude (i.e., magnitude) of the sine wave is modified. Furthermore, a time-domain graph of a modulated carrier has a shape similar to the signal that was used. For example, if one imagines an *envelope* consisting of a curve that connects the peaks of the sine wave in Figure 10.2(c), the resulting curve has the same shape as the information signal in Figure 10.2(b).

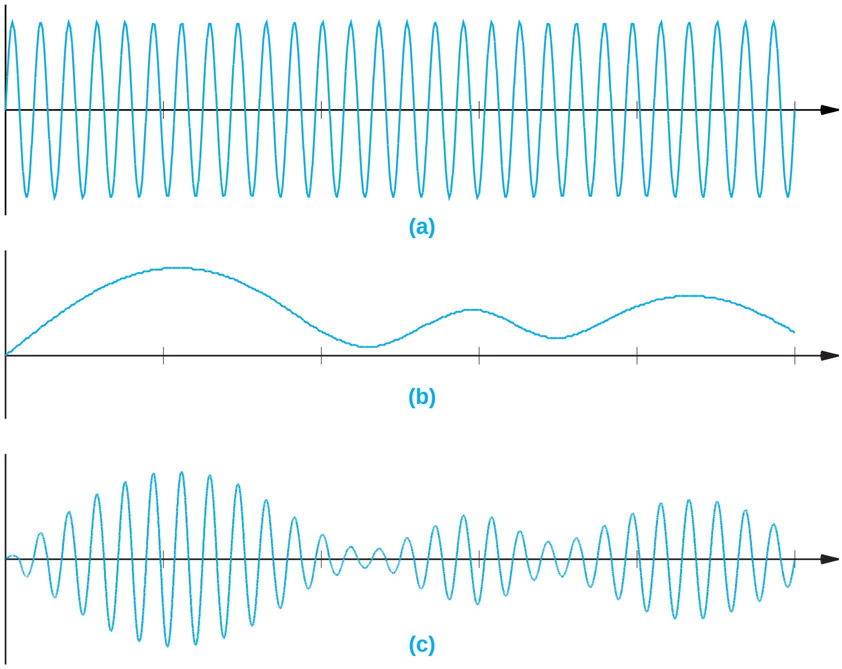


Figure 10.2 Illustration of (a) an unmodulated carrier wave, (b) an analog information signal, and (c) an amplitude modulated carrier.

10.5 Frequency Modulation

An alternative to amplitude modulation is known as *frequency modulation*. When frequency modulation is employed, the amplitude of the carrier remains fixed, but the frequency changes according to the signal: when the signal is stronger, the carrier frequency increases slightly, and when the signal is weaker, the carrier frequency decreases slightly. Figure 10.3 illustrates a carrier wave modulated with frequency modulation according to the signal in Figure 10.2(b).

As the figure shows, frequency modulation is more difficult to visualize because slight changes in frequency are not as clearly visible. However, one can notice that the modulated wave has higher frequencies when the signal used for modulation is stronger.

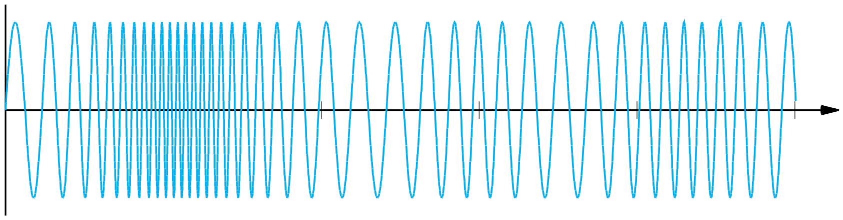


Figure 10.3 Illustration of a carrier wave with frequency modulation according to the signal in Figure 10.2(b).

10.6 Phase Shift Modulation

The third property of a sine wave is its *phase*, the offset from a reference time at which the sine wave begins. It is possible to use changes in phase to represent a signal. We use the term *phase shift* to characterize such changes.

Although modulating phase is possible in theory, the technique is seldom used with an analog signal. To understand why, observe that if phase changes after cycle *k*, the next sine wave will start slightly later than the time at which cycle *k* completes. A slight delay resembles a change in frequency. Thus, for analog input, phase shift modulation can be thought of as a special form of frequency modulation. We will see, however, that phase shifts are important when a digital signal is used to modulate a carrier.

10.7 Amplitude Modulation And Shannon's Theorem

The illustration in Figure 10.2(c) shows the amplitude varying from a maximum to almost zero. Although it is easy for a human to understand, the figure is slightly misleading: in practice, modulation only changes the amplitude of a carrier slightly, depending on a constant known as the *modulation index*.

To understand why practical systems do not allow for a modulated signal to approach zero, consider Shannon's Theorem. Assuming the amount of noise is constant, the signal-to-noise ratio will approach zero as the signal approaches zero. Thus, keeping the carrier wave near maximum ensures that the signal-to-noise ratio remains as large as possible, which permits the transfer of more bits per second.

10.8 Modulation, Digital Input, And Shift Keying

The description of modulation above shows how an analog information signal is used to modulate a carrier. The question arises: how can digital input be used? The answer lies in straightforward modifications of the modulation schemes described above: instead of modulation that is proportional to a continuous signal, digital schemes use discrete values. Furthermore, to distinguish between analog and digital modulation, we use the term *shift keying* rather than modulation.

In essence, shift keying operates similar to analog modulation. Instead of a continuum of possible values, digital shift keying has a fixed set. For example, amplitude modulation allows the amplitude of a carrier to vary by arbitrarily small amounts in response to a change in the signal being used. In contrast, amplitude shift keying uses a fixed set of possible amplitudes. In the simplest case, a full amplitude can correspond to a logical 1 and a significantly smaller amplitude can correspond to a logical 0. Similarly, frequency shift keying uses two basic frequencies. Figure 10.4 illustrates a carrier wave, a digital input signal, and the resulting waveforms for *Amplitude Shift Keying* (*ASK*) and *Frequency Shift Keying* (*FSK*).

10.9 Phase Shift Keying

Although amplitude and frequency changes work well for audio, both require at least one cycle of a carrier wave to send a single bit unless a special encoding scheme is used (e.g., unless positive and negative parts of the signal are changed independently). The Nyquist Theorem described in Chapter 6 suggests that the number of bits sent per unit time can be increased if the encoding scheme permits multiple bits to be encoded in a single cycle of the carrier. Thus, data communications systems often use techniques that can send more bits. In particular, *phase shift keying* changes the phase of the carrier wave abruptly to encode data. Each such change is called a *phase shift*. After a phase shift, the carrier continues to oscillate, but it immediately jumps to a new point in the sine wave cycle. Figure 10.5 illustrates how a phase shift affects a sine wave.

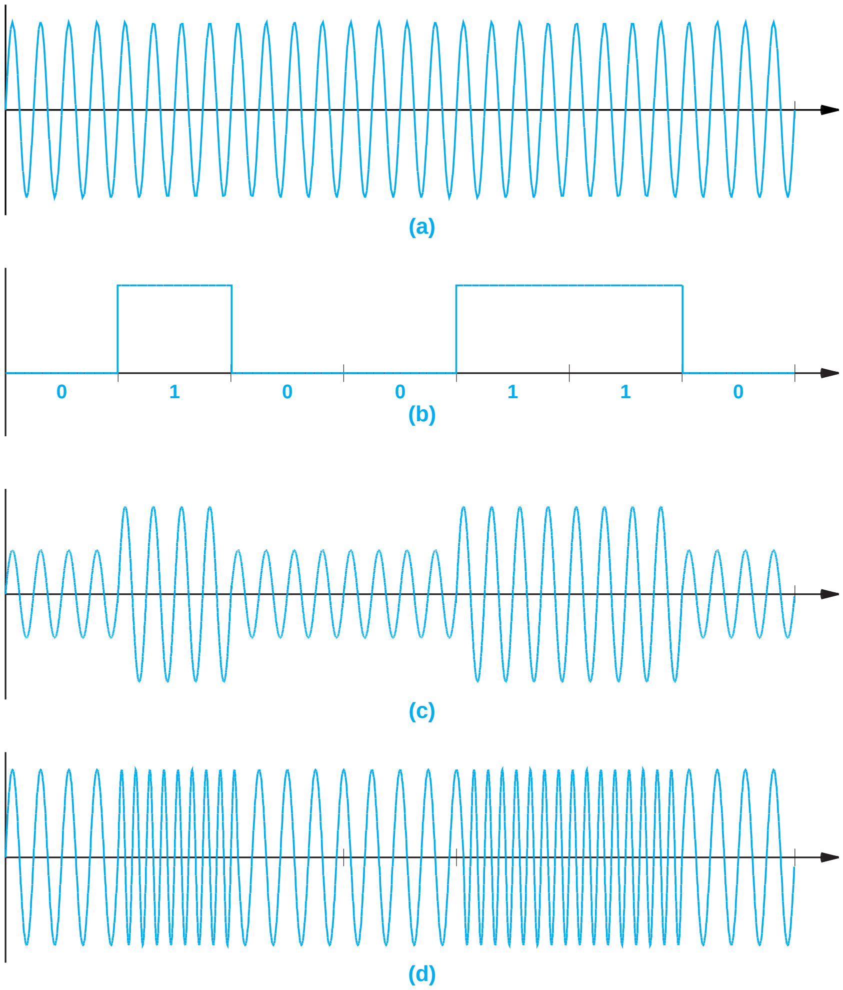


Figure 10.4 Illustration of (a) carrier wave, (b) digital input signal, (c) amplitude shift keying, and (d) frequency shift keying.

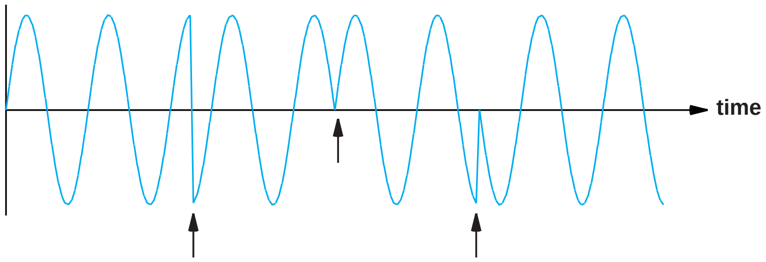


Figure 10.5 An illustration of phase shift modulation with arrows indicating times at which the carrier abruptly jumps to a new point in the sine wave cycle.

A phase shift is measured by the angle of the change. For example, the leftmost shift in Figure 10.5 changes the angle by $\pi/2$ radians or 180°. The second phase change in the figure also corresponds to a 180° shift. The third phase change corresponds to a shift of –90° (which is equivalent to 270°).

10.10 Phase Shift And A Constellation Diagram

How can data be encoded using phase shifts? In the simplest case, a sender and receiver can agree on the number of bits per second, and can use no phase shift to denote a logical 0, and the presence of a phase shift to denote a logical 1. For example, a system might use a 180° phase shift. A *constellation diagram* is used to express the exact assignment of data bits to specific phase changes. Figure 10.6 illustrates the concept.

Hardware can do more than detect the presence of a phase shift — a receiver can measure the amount a carrier shifted during a phase change. Thus, it is possible to devise a communications system that recognizes a set of phase shifts, and use each particular phase shift to represent specific values of data. Usually, systems are designed to use a power of two possible shifts, which means a sender can use bits of data to select among the shifts.

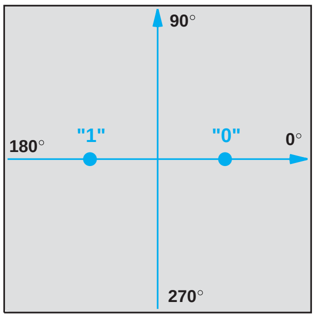


Figure 10.6 A constellation diagram that shows logical 0 as a 0° phase shift and logical 1 as a 180° phase shift.

Figure 10.7 shows the constellation diagram for a system that uses four possible phase shifts (i.e., 2^2). At each stage of transmission, a sender uses two bits of data to select among the four possible shift values.

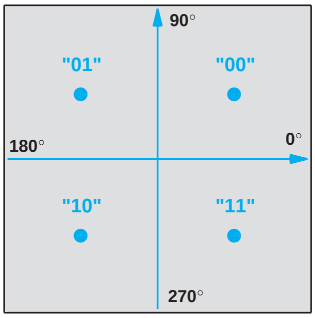


Figure 10.7 A constellation diagram for a system that uses four possible phase shifts that each represent two data bits.

To summarize:

The chief advantage of mechanisms like phase shift keying arises from the ability to represent more than one data bit at a given change. A constellation diagram shows the assignment of data bits to phase changes.

Many variations of phase shift keying exist and are used in practical networks. For example, a phase shift mechanism such as the one illustrated in Figure 10.6 that permits a sender to transfer one bit at a time is classified as a *Binary Phase Shift Keying* (*BPSK*) mechanism. The notation *2-PSK* is used to denote the two possible values. Similarly, the variation illustrated in Figure 10.7 is known as a *4-PSK* mechanism.

In theory, it is possible to increase the data rate by increasing the range of phase shifts. Thus, a 16-PSK mechanism can send twice as many bits per second as a 4-PSK mechanism. In practice, however, noise and distortion limit the ability of hardware to distinguish among minor differences in phase shifts. The point is:

> *Although many variations of phase shift keying exist, noise and distortion limit the ability of practical systems to distinguish among arbitrarily small differences in phase changes.*

10.11 Quadrature Amplitude Modulation

If hardware is incapable of detecting arbitrary phase changes, how can the data rate be increased further? The answer lies in a combination of modulation techniques that change two characteristics of a carrier at the same time. The most sophisticated technology combines amplitude modulation and phase shift keying. Known as *Quadrature Amplitude Modulation*† (*QAM*), the approach uses both change in phase and change in amplitude to represent values.

To represent QAM on a constellation diagram, we use distance from the origin as a measure of amplitude. For example, Figure 10.8 shows the constellation diagram for a variant known as *16QAM* with dark gray areas indicating the amplitudes.

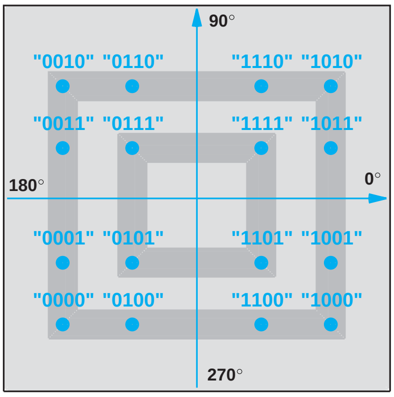


Figure 10.8 A constellation diagram for 16QAM in which distance from the origin reflects amplitude.

†The literature and networking industry usually use the term *quadrature amplitude modulation*, even though the term *quadrature amplitude shift keying* would be more accurate.

10.12 Modem Hardware For Modulation And Demodulation

A hardware mechanism that accepts a sequence of data bits and applies modulation to a carrier wave according to the bits is called a *modulator*; a hardware mechanism that accepts a modulated carrier wave and recreates the sequence of data bits that was used to modulate the carrier is called a *demodulator*. Thus, transmission of data requires a modulator at one end of the transmission medium and a demodulator at the other. In practice, most communications systems are full duplex communication, which means each location needs both a modulator, which is used to send data, and a demodulator, which is used to receive data. To keep cost low and make the pair of devices easy to install and operate, manufacturers combine modulation and demodulation mechanisms into a single device called a *modem* (*mo*dulator and *dem*odulator). Figure 10.9 illustrates how a pair of modems use a 4-wire connection to communicate.

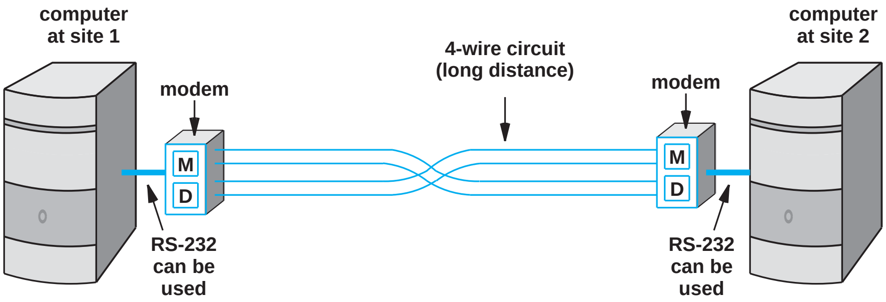


Figure 10.9 Illustration of two modems that use a 4-wire connection.

As the figure indicates, modems are designed to provide communication over long distances. A 4-wire circuit connecting two modems can extend inside a building, across a corporate campus between buildings, or between cities†.

10.13 Optical And Radio Frequency Modems

In addition to dedicated wires, modems are also used with other media, including RF transmission and optical fibers. For example, a pair of *Radio Frequency* (*RF*) modems can be used to send data via radio (e.g., in a Wi-Fi network), and a pair of *optical modems* can be used to send data across a pair of optical fibers. Although such modems use entirely different media than modems that operate over dedicated wires, the principle remains the same: at the sending end, a modem modulates a carrier; at the receiving end, data is extracted from the modulated carrier.

†A circuit that crosses public property must be leased from a service provider, usually a telephone company.

10.14 Dialup Modems

Another interesting application of modems involves the voice telephone system. Instead of using an electrical signal as a carrier, a *dialup modem* uses an audio tone. As with conventional modems, the carrier is modulated at the sending end and demodulated at the receiving end. Thus, besides the ability to place and receive telephone calls, the chief difference between dialup and conventional modems arises from the lower bandwidth of audible tones.

When dialup modems were first designed, the approach made complete sense — a dialup modem converted data into a modulated analog carrier because the telephone system transported analog signals. Ironically, the interior of a modern telephone system is digital. Thus, on the sending side, a dialup modem uses data to modulate an audible carrier, which is transmitted to the phone system. The phone system digitizes the incoming audio, transports a digital form internally, and converts the digitized version back to analog audio for delivery. The receiving modem demodulates the analog carrier, and extracts the original digital data. Figure 10.10 illustrates the ironic use of analog and digital signals by dialup modems.

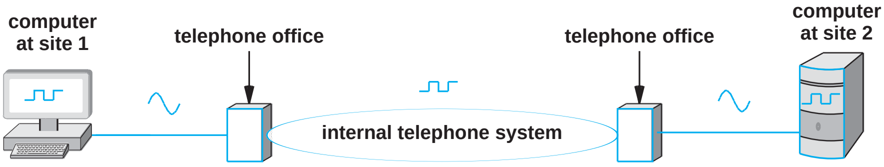


Figure 10.10 Illustration of digital and analog signals (denoted by a square wave and a sine wave) that occur when a dialup modem is used to send data from one computer to another.

As the figure indicates, a dialup modem is usually embedded in a computer. We use the term *internal modem* to denote an embedded device, and the term *external modem* to denote a separate physical device.

10.15 QAM Applied To Dialup

Quadrature Amplitude Modulation is also used with dialup modems as a way to maximize the rate at which data can be sent. To understand why, consider Figure 10.11, which shows the bandwidth available on a dialup connection. As the figure illustrates, most telephone connections transfer frequencies between 300 and 3000 Hz, but a given connection may not handle the extremes well. Thus, to guarantee better reproduction and lower noise, dialup modems use frequencies between 600 and 3000 Hz, which means the available bandwidth is 2400 Hz. A QAM scheme can increase the data rate dramatically.

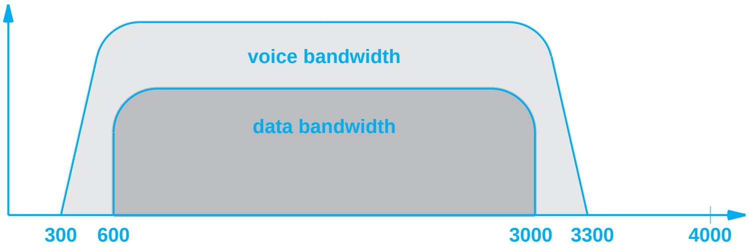


Figure 10.11 Illustration of the voice and data bandwidth on a dialup telephone connection.

10.16 V.32 And V.32bis Dialup Modems

As an example of dialup modems that use QAM, consider the *V.32* and *V.32bis* standards. Figure 10.12 illustrates the QAM constellation for a V.32 modem that uses 32 combinations of amplitude shift and phase shift to achieve a data rate of 9600 bps in each direction.

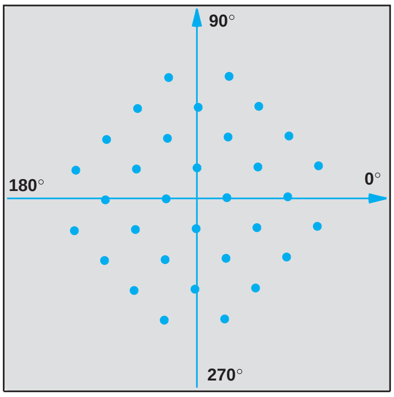


Figure 10.12 Illustration of the QAM constellation for a V.32 dialup modem.

A V.32bis modem uses 128 combinations of phase shift and amplitude shift to achieve a data rate of 14,400 bps in each direction. Figure 10.13 illustrates the constellation. Sophisticated signal analysis is needed to detect the minor change that occurs from a point in the constellation to a neighboring point.

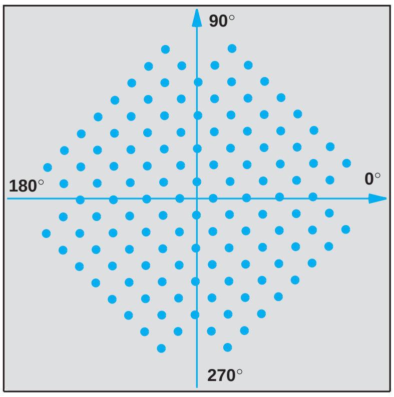


Figure 10.13 Illustration of the QAM constellation for a V.32bis dialup modem.

10.17 Summary

Long distance communications systems use a modulated carrier wave to transfer information. A carrier is modulated by changing the amplitude, frequency, or phase. Amplitude and frequency modulation are the most common forms used with an analog input.

When a digital signal is used as input, modulation is known as shift keying. As with analog modulation, shift keying changes a carrier. However, only a fixed set of possibilities are allowed. A constellation diagram is used to represent the possibilities for phase shift keying. If the system permits a power of two possibilities, multiple input bits can be used to select a possibility at each point in time. Quadrature Amplitude Modulation combines amplitude shift keying and phase shift keying to produce more possibilities.

A modem is a hardware device that includes circuitry to perform both modulation and demodulation; a pair of modems is used for full-duplex communication. Optical, RF, and dialup modems also exist. Because bandwidth is limited, dialup modems use Quadrature Amplitude Modulation schemes. A V.32 modem uses 32 possible combinations of phase shifts and amplitude changes; a V.32bis modem uses 128 possible combinations.

EXERCISES

10.1 When using amplitude modulation, does it make sense for a 1 Hz carrier to be modulated by a 2 Hz sine wave? Why or why not?

10.2 List the three basic types of analog modulation, and describe the differences among them.

10.3 What is the difference between shift keying and modulation?

10.4 Using Shannon’s Theorem, explain why practical amplitude modulation systems keep the carrier near maximum strength.

10.5 Search the Web and find a constellation diagram for 32QAM. How many points are defined in each quadrant?

10.6 In phase shift keying, is it possible to have a phase shift of 90°? Of 270°? Of 360°? Draw an example to explain your answer.

10.7 Assuming a signal-to-noise ration of 30 dB, what is the maximum data rate that can be achieved for the dialup bandwidth illustrated in Figure 10.11?

10.8 Figure 10.9 shows a full-duplex configuration with four wires, two of which are used to transmit in each direction. Argue that it should be possible to use three wires instead.

10.9 In the previous question, why are four wires preferable?

Chapter Contents

11

Multiplexing And Demultiplexing (Channelization)

11.1 Introduction

Chapters in this part of the text cover the fundamentals of data communications. The previous chapter discusses the concept of modulation, and explains how a carrier wave can be modulated to carry analog or digital information.

This chapter continues the discussion of data communications by introducing multiplexing. The chapter describes the motivation, and defines basic types of multiplexing that are used throughout computer networks and the Internet. The chapter also explains how modulated carriers provide the basis for many multiplexing mechanisms.

11.2 The Concept Of Multiplexing

We use the term *multiplexing* to refer to the combination of information streams from multiple sources for transmission over a shared medium, and *multiplexor* to denote a mechanism that implements the combination. Similarly, we use the term *demultiplexing* to refer to the separation of a combination back into separate information streams, and *demultiplexor* to refer to a mechanism that implements separation. Multiplexing and demultiplexing are not restricted to hardware or to individual bit streams — in later chapters, we will see that the idea of combining and separating communication forms a foundation for many aspects of computer networking. Figure 11.1 illustrates the concept.

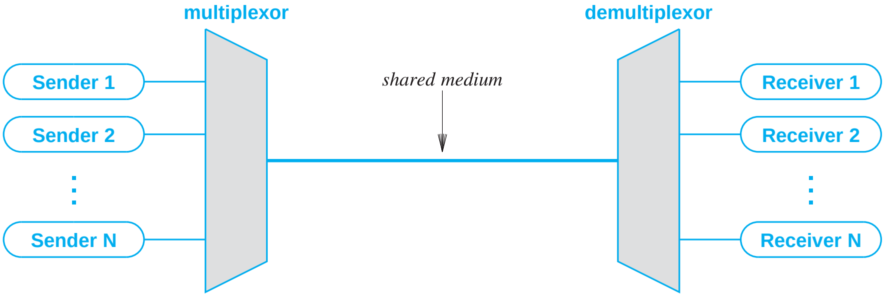


Figure 11.1 The concept of multiplexing in which independent pairs of senders and receivers share a transmission medium.

In the figure, each sender communicates with a single receiver. Although each pair carries on independent communication, all pairs share a single transmission medium. The multiplexor combines information from the senders for transmission in such a way that the demultiplexor can separate the information for receivers. Users experience forms of multiplexing every day. For example, multiple computers at a residence can use a wireless router to communicate with an Internet site. Each computer carries on a separate conversation, and all conversations are multiplexed over the connection between the residence and an ISP.

11.3 The Basic Types Of Multiplexing

At the physical layer, there are four basic approaches to multiplexing that each have a set of variations and implementations.

- Frequency division multiplexing
- Wavelength division multiplexing
- Time division multiplexing
- Code division multiplexing

Frequency and time division multiplexing are widely used. Wavelength division multiplexing is a form of frequency division multiplexing used for optical fiber. Code division multiplexing is a mathematical approach used in some cell phone systems.

11.4 Frequency Division Multiplexing (FDM)

Frequency Division Multiplexing (*FDM*) is easy to understand because it forms the basis for broadcast radio. The underlying principle arises from the physics of transmission: a set of radio stations can transmit electromagnetic signals simultaneously without interference, provided they each use a separate *channel* (i.e., carrier frequency). Data communications systems apply the same principle by simultaneously sending multiple carrier waves over a single copper wire or using wavelength division multiplexing to send multiple frequencies of light over an optical fiber. At the receiving end, a demultiplexor applies a set of filters that each extract a small range of frequencies near one of the carrier frequencies. Figure 11.2 illustrates the organization.

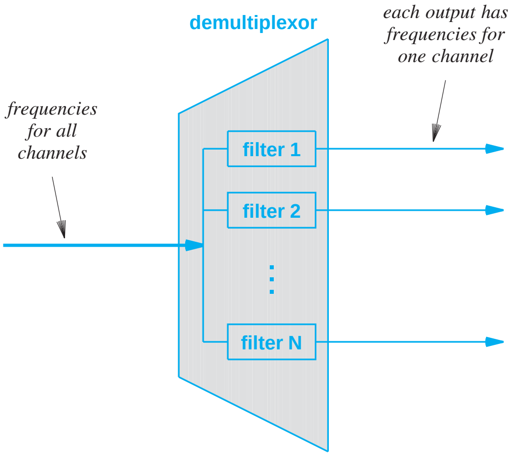

demultiplexor
each output has frequencies for one channel
frequencies for all channels
filter 1
filter 2
filter N

Figure 11.2 Illustration of basic FDM demultiplexing where a set of filters each selects the frequencies for one channel and suppresses other frequencies.

A key idea is that the filters used in FDM only examine frequencies. If a sender and receiver pair are assigned a particular carrier frequency, the FDM mechanism will separate the frequency from others without otherwise modifying the signal. Thus, any of the modulation techniques discussed in Chapter 10 can be used with any carrier.

The point is:

> *Because carrier waves on separate frequencies do not interfere, frequency division multiplexing provides each sender and receiver pair with a private communication channel over which any modulation scheme can be used.*

The most significant advantage of FDM arises from the simultaneous use of a transmission medium by multiple pairs of communicating entities. We imagine FDM as providing each pair with a private transmission path as if the pair had a separate physical transmission medium. Figure 11.3 illustrates the concept.

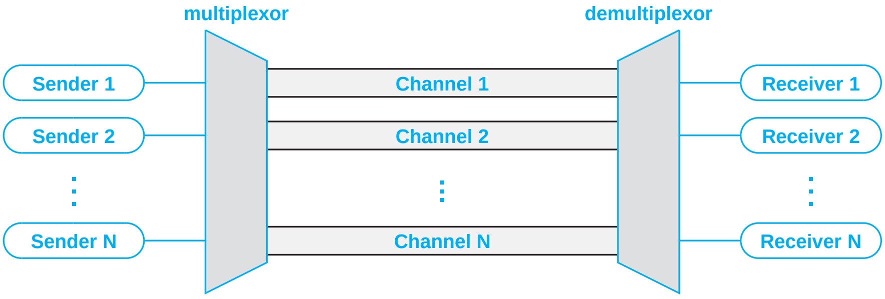


Figure 11.3 The conceptual view of Frequency Division Multiplexing (FDM) providing a set of independent channels.

Of course, any practical FDM system imposes limits on the set of frequencies that can be used for channels. If the frequencies of two channels are arbitrarily close, interference can occur. Furthermore, demultiplexing hardware that receives a combined signal must be able to divide the signal into separate carriers. For broadcast radio in the U.S., the *Federal Communications Commission* (*FCC*) regulates stations to ensure that adequate *spacing* occurs between the carrier frequencies. For data communications systems, designers follow the same approach by choosing a set of carrier frequencies with a gap between them known as a *guard band*.

As an example of channel allocation, consider the assignment in Figure 11.4 that allocates 200 KHz to each of 6 channels with a guard band of 20 KHz between adjacent channels.

Channel	Frequencies Used
1	100 KHz - 300 KHz
2	320 KHz - 520 KHz
3	540 KHz - 740 KHz
4	760 KHz - 960 KHz
5	980 KHz - 1180 KHz
6	1200 KHz - 1400 KHz

Figure 11.4 An example assignment of frequencies to channels with a guard band between adjacent channels.

When plotted in the frequency domain, the guard band is clearly visible. Figure 11.5 contains the plot for the assignment in Figure 11.4.

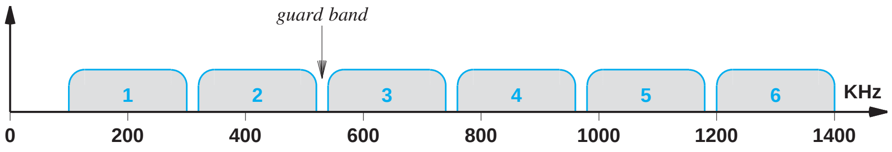


Figure 11.5 A frequency domain plot of the channel allocation from Figure 11.4 with a guard band visible between channels.

11.5 Using A Range Of Frequencies Per Channel

If a carrier uses a single frequency, why does the example allocate blocks of frequencies? To understand the motivation, consider general characteristics of FDM:

- *Long-lived.* FDM predates modern data communications — the idea of dividing the electromagnetic spectrum into channels arose in early experiments with radio.
- *Widely Used.* FDM has been used in broadcast radio and television, cable television, and the AMPS cellular telephone system.
- *Analog.* FDM multiplexing and demultiplexing hardware accepts and delivers analog signals. Even if a carrier has been modulated to contain digital information, FDM hardware treats the carrier as an analog wave.
- *Versatile.* Because it filters on ranges of frequency without examining other aspects of signals, FDM is versatile.

The analog characteristic has the disadvantage of making frequency division multiplexing susceptible to noise and distortion†, but the advantage of providing flexibility. In particular, most FDM systems assign each sender and receiver pair a range of frequencies and the ability to choose how the frequencies can be used. There are two primary ways that systems use a range of frequencies.

- Increase the data rate
- Increase immunity to interference

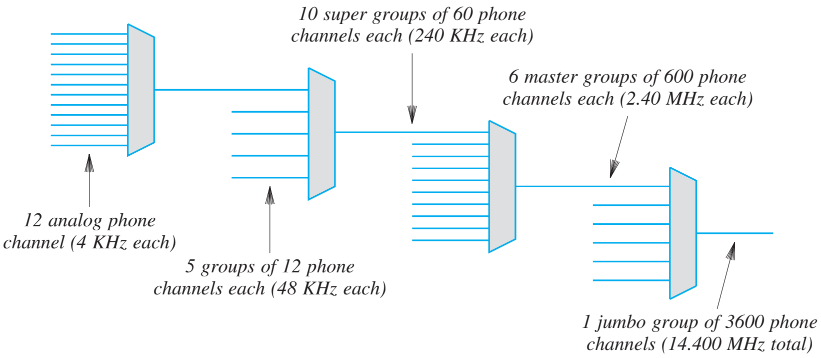
To increase the overall data rate, a sender divides the frequency range of the channel into *K* carriers, and sends *1/K* of the data over each carrier. In essence, a sender performs frequency division multiplexing within the channel that has been allocated. Some systems use the term *subchannel allocation* to refer to the subdivision.

†Data communications systems that use FDM often require coaxial cable to provide more immunity to noise.

To increase immunity to interference, a sender uses a technique known as *spread spectrum*. Various forms of spread spectrum can be used, but the basic idea is to divide the range of the channel into K carriers, transmit the same data over multiple channels, and allow a receiver to use a copy of the data that arrives with fewest errors. The scheme, used extensively in wireless networks, works extremely well in cases where noise is likely to interfere with some frequencies at a given time.

11.6 Hierarchical FDM

Some of the flexibility in FDM arises from the ability of hardware to shift frequencies. If a set of incoming signals all use the frequency range between 0 and 4 KHz, multiplexing hardware can leave the first stage as is, map the second stage onto the range 4 KHz to 8 KHz, map the third stage onto the range 8 KHz to 12 KHz, and so on. The technique forms the basis for a hierarchy of FDM multiplexors that each map their inputs to a larger, continuous band of frequencies. Figure 11.6 illustrates the concept of *hierarchical FDM*†.



10 super groups of 60 phone channels each (240 KHz each)
6 master groups of 600 phone channels each (2.40 MHz each)
12 analog phone channel (4 KHz each)
5 groups of 12 phone channels each (48 KHz each)
1 jumbo group of 3600 phone channels (14.400 MHz total)

Figure 11.6 Illustration of the FDM hierarchy used in the telephone system.

As the figure illustrates, the primary input consists of a set of twelve analog telephone signals, which each occupy frequencies 0 through 4 KHz. At the first stage, the signals are multiplexed into a single signal known as a *group* that uses the frequency range of 0 through 48 KHz. At the next stage, five groups are multiplexed into a single *supergroup* that uses frequencies 0 through 240 KHz, and so on. At the final stage, 3600 telephone signals have been multiplexed into a single signal. To summarize:

> *It is possible to build a hierarchy of frequency division multiplexing in which each stage accepts as inputs the outputs from the previous stage.*

†In practice, additional bandwidth is needed to carry framing bits.

11.7 Wavelength Division Multiplexing (WDM)

The term *Wavelength Division Multiplexing* (*WDM*) refers to the application of frequency division multiplexing to optical fiber; the term *Dense Wavelength Division Multiplexing* (*DWDM*) is used to emphasize that many wavelengths of light are employed. The inputs and outputs of such multiplexing are wavelengths of light, denoted by the Greek letter λ, and informally called *colors*. To understand how multiplexing and demultiplexing can work with light, recall from basic physics that when white light passes through a prism, colors of the spectrum are spread out. A prism operates in the reverse mode as well: if a set of colored light beams are each directed into a prism at the correct angle, the prism will combine the beams to form a single beam of white light. Finally, recall that what humans perceive as a color is in fact a range of wavelengths of light.

Prisms form the basis of optical multiplexing and demultiplexing. A multiplexor accepts beams of light of various wavelengths, and uses a prism to combine them into a single beam; a demultiplexor uses a prism to separate the wavelengths. Figure 11.7 illustrates the concept.

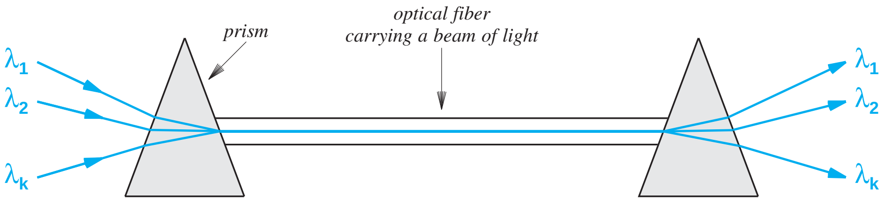


Figure 11.7 Illustration of prisms used to combine and separate wavelengths of light in wavelength division multiplexing technologies.

The point is:

> *When frequency division multiplexing is applied to optical fiber, prisms are used to combine or separate individual wavelengths of light, and the result is known as wavelength division multiplexing.*

11.8 Time Division Multiplexing (TDM)

The chief alternative to FDM is known as *Time Division Multiplexing* (*TDM*). TDM is less esoteric than FDM, and does not rely on special properties of electromagnetic energy. Instead, multiplexing in time simply means transmitting an item from one source, then transmitting an item from another source, and so on. Figure 11.8 illustrates the concept.

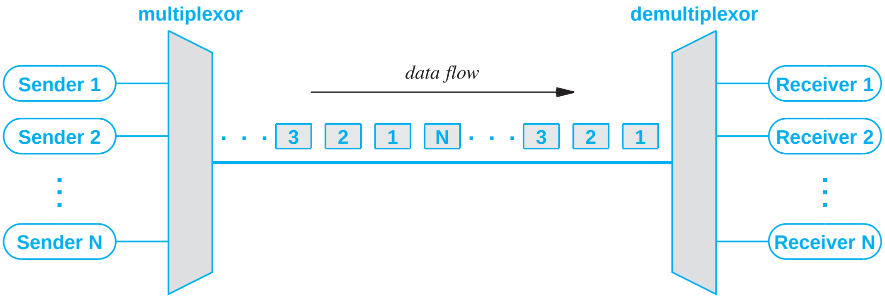


Figure 11.8 Illustration of the Time Division Multiplexing (TDM) concept with items from multiple sources sent over a shared medium.

11.9 Synchronous TDM

Time division multiplexing is a broad concept that appears in many forms and is widely used throughout the Internet. Thus, the diagram in Figure 11.8 is merely a conceptual view, and the details may vary. For example, the figure shows items being sent in a *round-robin* fashion (i.e., an item from sender 1 followed by an item from sender 2, and so on). Although some TDM systems use round-robin order, other do not.

A second detail in Figure 11.8 — the slight time gap between items — does not apply to all types of TDM. Extending the terminology of Chapter 9, we say that a *synchronous TDM* system sends items one after another with no delay. Figure 11.9 illustrates how synchronous TDM works for a system of four senders.

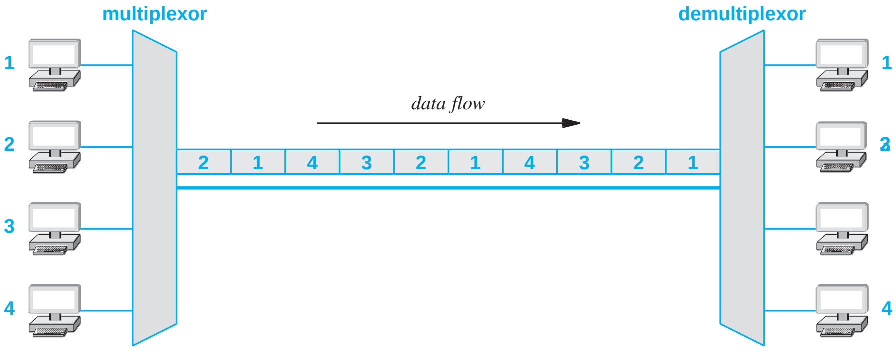


Figure 11.9 Illustration of a synchronous TDM system with four senders.

11.10 Framing Used In The Telephone System Version Of TDM

Analog telephone systems use synchronous TDM to multiplex digital streams from multiple phone calls over a single medium. In fact, telephone companies use the acronym TDM to refer to the specific form of TDM used to multiplex digital telephone calls.

The phone system standards for TDM include an interesting technique to ensure that a demultiplexor stays synchronized with the multiplexor. To understand why synchronization is needed, observe that a synchronous TDM system sends one data item after another without any indication of the output to which a given item occurs. Because a demultiplexor cannot tell where an item begins, a slight difference in the clocks used to time bits can cause a demultiplexor to misinterpret the bit stream.

To prevent misinterpretation, the version of TDM used in the phone system includes an extra *framing channel* as input. Instead of taking a complete slot, framing inserts a single bit in the stream on each round. Along with other channels, a demultiplexor extracts data from the framing channel and checks for alternating 0 and 1 bits. The idea is that if an error causes a demultiplexor to lose a bit, it is highly likely that the framing check will detect the error and allow the transmission to be restarted. Figure 11.10 illustrates the use of framing bits.

To summarize:

> *The synchronous TDM mechanism used for digital telephone calls includes a framing bit at the beginning of each round. The framing sequence of alternating 1s and 0s ensures that a demultiplexor either remains synchronized or detects the error.*

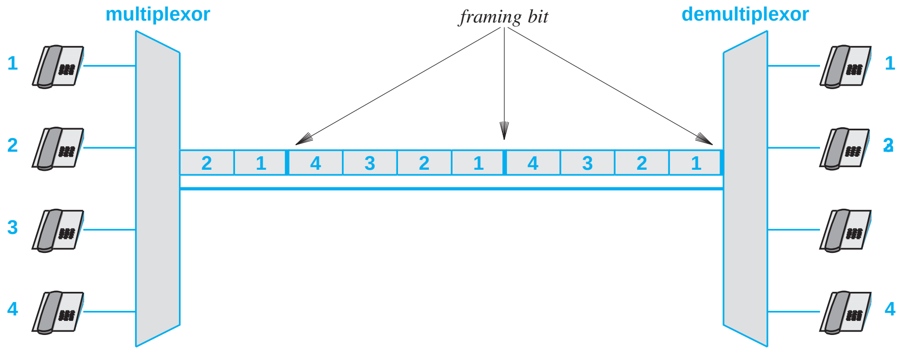


Figure 11.10 Illustration of the synchronous TDM system used by the telephone system in which a framing bit precedes each round of slots.

11.11 Hierarchical TDM

Like a frequency division multiplexing system, a TDM system can be arranged in a hierarchy. The difference is that each successive stage of a TDM hierarchy uses N times the bit rate, whereas each successive stage of an FDM hierarchy uses N times the frequencies. Additional framing bits are added to the data, which means that the bit rate of each successive layer of hierarchy is slightly greater than the aggregate voice traffic. Compare the example TDM hierarchy in Figure 11.11 with the FDM example in Figure 11.6 on page 220.

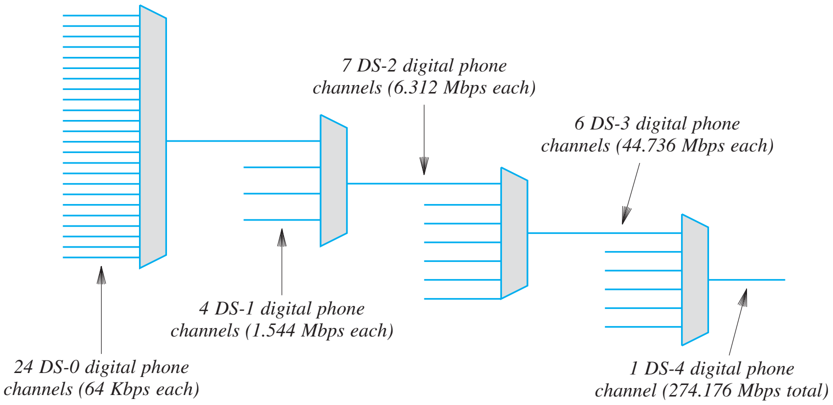


Figure 11.11 Illustration of the TDM hierarchy used in the telephone system.

11.12 The Problem With Synchronous TDM: Unfilled Slots

Synchronous TDM works well if each source produces data at a uniform, fixed rate equal to *1/N* of the capacity of the shared medium. For example, if a source corresponds to a digital telephone call, the data will arrive at a uniform rate of 64 Kbps. As Chapter 9 points out, many sources generate data in bursts, with idle time between bursts, which does not work well with a synchronous TDM system. To understand why, consider the example in Figure 11.12.

In the figure, sources on the left produce data items at random. Thus, the synchronous multiplexor leaves a slot unfilled if the corresponding source has not produced an item by the time the slot must be sent. In practice, of course, a slot cannot be empty because the underlying system must continue to transmit data. Thus, the slot is assigned a value (such as zero), and an extra bit is set to indicate that the value is invalid.

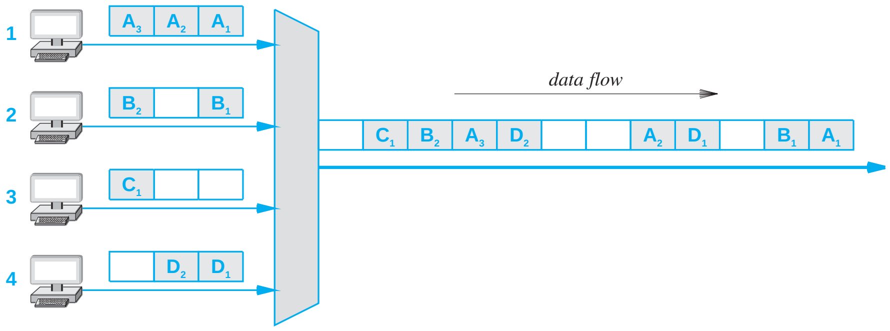


Figure 11.12 Illustration of a synchronous TDM system leaving slots unfilled when a source does not have a data item ready in time.

11.13 Statistical TDM

How can a TDM multiplexing system make more efficient use of a shared medium? One technique to increase the overall data rate is known as *statistical TDM* or *statistical multiplexing*†. The terminology is awkward, but the technique is straightforward: select items for transmission in a round-robin fashion, but instead of leaving a slot unfilled, skip any source that does not have data ready. By eliminating unused slots, statistical TDM takes less time to send the same amount of data. For example, Figure 11.13 illustrates how a statistical TDM system sends the data from Figure 11.12 in only 8 slots instead of 12.

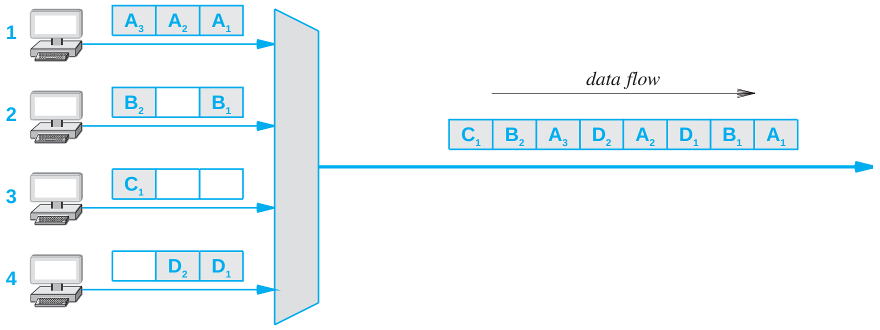


Figure 11.13 Illustration that shows how statistical multiplexing avoids unfilled slots and takes less time to send data.

†Some literature uses the term *asynchronous time division multiplexing*.

Although it avoids unfilled slots, statistical multiplexing incurs extra overhead. To see why, consider demultiplexing. In a synchronous TDM system a demultiplexor knows that every N^{th} slot corresponds to a given receiver. In a statistical multiplexing system, the data in a given slot can correspond to any receiver. Thus, in addition to data, each slot must contain the identification of the receiver to which the data is being sent. Later chapters discuss identification mechanisms that are used with statistical multiplexing in packet switching networks and the Internet.

11.14 Inverse Multiplexing

An interesting twist on multiplexing arises in cases where the only connection between two points consists of multiple transmission media, but no single medium has a bit rate that is sufficient. At the core of the Internet, for example, service providers need higher bit rates than are available. For example, an ISP may need a capacity of 100 Gbps, but can only acquire connections that operate at 10 Gbps†. To solve the problem, multiplexing is used in reverse: send data across ten 10 Gbps circuits in parallel. In general, inverse multiplexing spreads high-speed digital input over multiple lower-speed circuits, and combines the results at the receiving end. Figure 11.14 illustrates the concept.

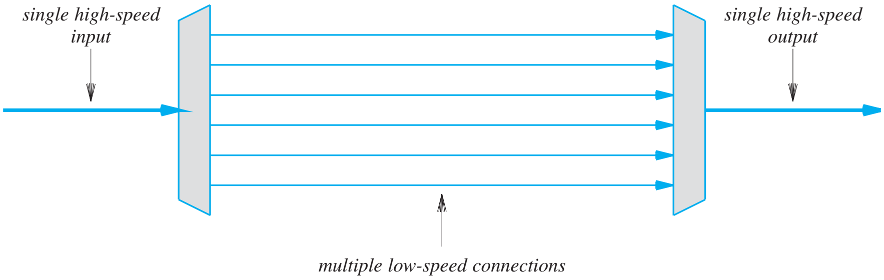


Figure 11.14 Illustration of inverse multiplexing in which a single high-capacity digital input is distributed over lower-capacity connections for transmission and then recombined to form a copy of the input.

In practice, an *inverse multiplexor* cannot be constructed merely by connecting the pieces of a conventional multiplexor backward. Instead, hardware must be designed so the sender and receiver agree on how data arriving from the input will be distributed over the lower-speed connections. More important, to ensure that all data is delivered in the same order as it arrived, the system must be engineered to handle cases where one or more of the lower-speed connections has longer latency than others. Despite its complexity, inverse multiplexing is widely used in the Internet.

†Inverse multiplexing is also used for economic reasons: in some cases, the cost of N lower-capacity circuits is less than the cost of a single high-capacity circuit.

11.15 Code Division Multiplexing

A final form of multiplexing used in parts of the cellular telephone system and for some satellite communication is known as *Code Division Multiplexing* (*CDM*). The specific version of CDM used in cell phones is known as *Code Division Multi-Access* (*CDMA*).

Unlike FDM and TDM, CDM does not rely on physical properties, such as frequency or time. Instead, CDM relies on an interesting mathematical idea: values from orthogonal vector spaces can be combined and separated without interference. The particular form used in the telephone network is easiest to understand. Each sender is assigned a unique binary code C_i that is known as a *chip sequence*. Chip sequences are selected to be orthogonal vectors (i.e., the dot product of any two chip sequences is zero). At any point in time, each sender has a value to transmit, V_i. The senders each multiply $C_i \times V_i$, and transmit the results. In essence, the senders transmit at the same time, and the values are added together. To extract value V_i, a receiver multiplies the sum by C_i.

To clarify the concept, consider an example. To keep the example easy to understand, we will use a chip sequence that is only two bits long and data values that are four bits long. We think of the chip sequence as a vector. Figure 11.15 lists the values.

Sender	Chip Sequence	Data Value
A	1 0	1 0 1 0
B	1 1	0 1 1 0

Figure 11.15 Example values for use with code division multiplexing.

The first step consists of converting the binary values into vectors of bits that use –1 to represent 0:

$$C_1 = (1, -1) \qquad V_1 = (1, -1, 1, -1) \qquad\qquad C_2 = (1, 1) \qquad V_2 = (-1, 1, 1, -1)$$

Multiplying $C_1 \times V_1$ and $C_2 \times V_2$ produces:

$$((1, -1), (-1, 1), (1, -1), (-1, 1)) \qquad\qquad ((-1, -1), (1, 1), (1, 1), (-1, -1))$$

If we think of the resulting values as a sequence of signal strengths to be transmitted at the same time, the resulting signal will be the sum of the two signals:

$$\begin{array}{crrrrrrrr} & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ + & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ \hline & 0 & -2 & 0 & 2 & 2 & 0 & -2 & 0 \end{array}$$

A receiver treats the sequence as a vector, computes the product of the vector and the chip sequence, treats the result as a sequence, and converts the result to binary by interpreting positive values as binary 1 and negative values as binary 0. Thus, Receiver *A* computes:

$$(1,\ -1)\ \bullet\ ((0,\ -2),\ (0,\ 2),\ (2,\ 0),\ (-2,\ 0))$$

to get:

$$((0+2),\ (0-2),\ (2+0),\ (-2+0))$$

Interpreting the result as a sequence produces:

$$2 \quad -2 \quad 2 \quad -2$$

which becomes the binary value:

$$1 \ \ 0 \ \ 1 \ \ 0$$

Note that 1010 is the correct value of V_1. Meanwhile, Receiver *B* will extract V_2 from the same transmission.

It may seem that CDM offers little advantage over TDM. In fact, CDM is somewhat inefficient because a large chip sequence is required, even if only a few senders transmit during a given interval. Thus, if the utilization is low, statistical TDM works better than CDM.

The advantages of CDM arise from its ability to scale and because it offers lower delay in a highly utilized network. To see why low delay is important, consider a statistical TDM system. Once a sender transmits, a TDM multiplexor allows $N-1$ other senders to transmit before giving the first sender another turn. Thus, if all senders are active, the potential delay between successive transmissions from a given sender can be high. In a CDM system, however, a sender can transmit at the same time as other senders, which means the delay is lower. CDM is especially attractive for a telephone service because low delay between transmissions is essential to delivering high-quality voice. To summarize:

> *CDM incurs lower delay than statistical TDM when a network is highly utilized.*

11.16 Summary

Multiplexing is a fundamental concept in data communications. A multiplexing mechanism allows pairs of senders and receivers to communicate over a shared medium. A multiplexor sends inputs from many senders over a shared medium, and a demultiplexor separates and delivers the items.

There are four basic approaches to multiplexing: frequency division, time division, wavelength division, and code division. Frequency Division Multiplexing (FDM) permits simultaneous communication over multiple channels, each of which corresponds to a separate frequency of electromagnetic radiation. Wavelength Division Multiplexing (WDM) is a form of frequency division multiplexing that sends frequencies of light, called wavelengths, over an optical fiber.

Time Division Multiplexing (TDM) sends one item at a time over the shared medium. A synchronous TDM system transmits items with no idle time between them, usually using round-robin selection. A statistical TDM system avoids empty slots by skipping any sender that does not have an item ready to send during its turn.

Code Division Multiplexing (CDM) uses a mathematical combination of codes that permits multiple senders to transmit at the same time without interference. The chief advantages of CDM arise from the ability to scale with low delay.

EXERCISES

11.1 What are the four basic types of multiplexing?

11.2 Give an example of multiplexing in a non-electronic communications system.

11.3 What is a guard band?

11.4 What are the general characteristics of FDM?

11.5 Explain how a range of frequencies can be used to increase data rate.

11.6 An FDM system may assign each channel a range of frequencies. Using a range is essential when which type of modulation is used for each carrier?

11.7 What is the key mechanism used to combine or separate wavelengths of light in a WDM system?

11.8 In a hierarchical FDM system, explain how a high-capacity channel is divided into subchannels.

11.9 Explain why framing and synchronization are important in a TDM system.

11.10 Is a TDM system required to use round-robin service?

11.11 Suppose N users compete using a statistical TDM system, and suppose the underlying physical transport can send K bits per second. What is the minimum and maximum data rate that an individual user can experience?

11.12 In a hierarchical TDM system, at what bit rate does the output of a given level need to operate? (Express the answer in terms of the number and bit rate of inputs.)

11.13 Of the four basic multiplexing techniques, is CDM always the best? Explain.

11.14 Suppose an OC-12 circuit is twenty percent the cost of an OC-48 circuit. What multiplexing technology can an ISP use to lower the cost of sending data at the OC-48 rate. Explain.

11.15 How does code division multiplexing work?

Chapter Contents

12

Access And Interconnection Technologies

12.1 Introduction

Chapters in this section each examine one of the fundamental aspects of data communications. The previous chapter discusses multiplexing and the concept of a multiplexing hierarchy. The chapter describes the time and frequency division multiplexing schemes that phone companies use for digital telephony.

This chapter concludes the discussion of data communications by examining two facilities used in the Internet. First, the chapter discusses access technologies, such as DSL, cable modems, and T1 lines that are used to connect individual residences and businesses to the Internet. Second, the chapter considers high-capacity digital circuits used in the core of the Internet. The chapter expands the discussion of the telephone system multiplexing hierarchy, and gives examples of circuits that common carriers offer to businesses and Internet Service Providers. The discussion focuses on the data communications aspects of the technologies by considering multiplexing and data rates.

12.2 Internet Access Technology: Upstream And Downstream

Internet access technology refers to a data communications system that connects an Internet *subscriber* (typically a private residence or business) to an *Internet Service Provider* (*ISP*), such as a telephone company or cable company. To understand how access technology is designed, one must know that most Internet users follow an *asymmetric* pattern. A typical residential subscriber receives more data from the Internet than they send. For example, to view a web page, a browser sends a URL that comprises a few bytes. In response, a web server sends content that may consist of thousands of bytes of text or an image that can comprise tens of thousands of bytes. A business that runs a web server may have the opposite traffic pattern — the business sends more data than it receives. The point is:

> *Because a typical residential subscriber receives much more information than the subscriber sends, Internet access technologies are designed to transfer more data in one direction than the other.*

The networking industry uses the term *downstream* to refer to data traveling from a service provider in the Internet to a subscriber, and *upstream* to refer to data traveling from a subscriber to a service provider. Figure 12.1 illustrates the definitions.

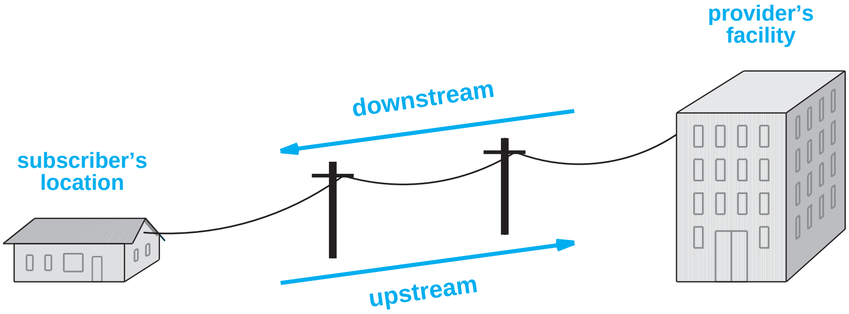


Figure 12.1 Definition of upstream and downstream directions as used in Internet access technologies.

12.3 Narrowband And Broadband Access Technologies

A variety of technologies are used to provide Internet access. They can be divided into two broad categories based on the data rate they provide:

- Narrowband
- Broadband

Although Chapter 6 explains the difference between the bandwidth of a transmission medium and the data rate, the terminology used for access networks does not observe the distinction. Instead, the networking industry generally uses the term *network bandwidth* to refer to data rate. Thus, the terms *narrowband* and *broadband* reflect industry practice.

12.3.1 Narrowband Technologies

Narrowband generally refers to technologies that deliver data at up to 128 Kbps. For example, the maximum data rate that can be achieved over an analog telephone connection with the most sophisticated modem technology and the least noisy phone lines is 56 Kbps. Thus, dialup is classified as a narrowband technology. Similarly, analog circuits that use modems, slower-speed digital circuits, and some of the data services that have been offered by telephone companies (e.g., ISDN) are narrowband. Figure 12.2 summarizes the main narrowband access technologies.

Narrowband
Dialup telephone connections
Leased circuit using modems
Fractional T1 data circuits
ISDN and other telco data services

Figure 12.2 The main narrowband technologies used for Internet access.

12.3.2 Broadband Technologies

The term *broadband* generally refers to technologies that offer high data rates, but the exact boundary between broadband and narrowband is blurry. Many professionals suggest that broadband technologies deliver more than 1 Mbps. However, providers such as telephone companies use the term *broadband* to refer to any service that offers a higher rate than dialup. Thus, phone companies claim any technology that offers 128 Kbps or higher can be classified as broadband. Figure 12.3 summarizes the main broadband access technologies.

Broadband
DSL technologies
Cable modem technologies
Wireless access technologies
Data circuits at T1 speed or higher

Figure 12.3 The main broadband Internet access technologies.

12.4 The Local Loop And ISDN

The term *local subscriber line* or *local loop* describes the physical connection between a telephone company *Central Office* (*CO*) and a subscriber's location. To understand how a local loop can be used, it is important to think of the local loop as independent from the rest of the phone system. Although the overall phone system is engineered to provide each dialup call with 4 KHz of bandwidth, the local loop portion consists of a pair of copper wires, and often has much higher potential bandwidth. In particular, the local loop for a subscriber close to a CO may be able to handle frequencies above 1 MHz.

As data networking became important, telephone companies explored ways to use the local loop to provide higher-speed data communication. One of the first phone company efforts to provide large-scale digital services to subscribers is offered under the name *Integrated Services Digital Network* (*ISDN*). From a subscriber's point of view, ISDN offers three separate digital channels, designated *B*, *B*, and *D* (usually written $2B + D$). The two *B* channels, which each operate at a speed of 64 Kbps, are intended to carry digitized voice, data, or compressed video; the *D* channel, which operates at 16 Kbps, is used as a control channel. In general, a subscriber uses the *D* channel to request services, which are then supplied over the *B* channels (e.g., a phone call that uses digital voice). Both of the *B* channels can be combined or *bonded* to produce a single channel with an effective data rate of 128 Kbps. When ISDN was first proposed, 128 Kbps seemed much faster than dialup modems. Newer local loop technologies provide higher data rates at lower cost, relegating ISDN to a few special cases.

12.5 Digital Subscriber Line (DSL) Technologies

Digital Subscriber Line (*DSL*) is one of the main technologies used to provide high-speed data communication services over a local loop. Figure 12.4 lists DSL variants. Because the names differ only in the first word, the set is collectively referred to by the acronym *xDSL*.

Name	Expansion	General Use
ADSL	Asymmetric DSL	Residential customers
ADSL2	Asymmetric DSL version 2	Approximately three times faster
SDSL	Symmetric DSL	Businesses that export data
HDSL	High bit rate DSL	Businesses up to 3 miles away
VDSL	Very-high bit rate DSL	Proposed version for 52 Mbps

Figure 12.4 Main variants of DSL that are collectively known as xDSL.

ADSL is the most widely deployed variant, and the one that most residential customers use. ADSL uses frequency division multiplexing to divide the bandwidth of the local loop into three regions. One of the regions corresponds to traditional analog phone service, which is known in the industry as *Plain Old Telephone Service* (*POTS*), and two regions provide data communication. The point is:

> *Because it uses frequency division multiplexing, ADSL and traditional analog phone service (POTS) can use the same wires simultaneously.*

Figure 12.5 illustrates how ADSL divides bandwidth.

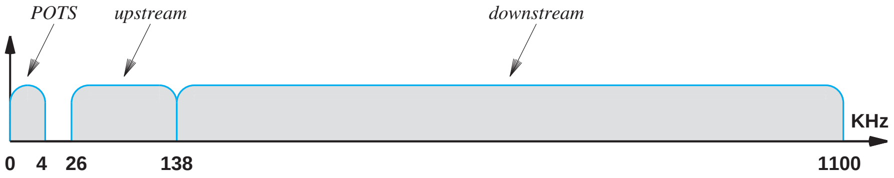


Figure 12.5 An illustration of how ADSL divides the available bandwidth of the local loop.

In the figure, the x-axis is not linear. If it were, the 4 KHz region reserved for POTS would not be visible, nor would the 22 KHz guard band between POTS and the upstream region.

12.6 Local Loop Characteristics And Adaptation

ADSL technology is complex because no two local loops have identical electrical characteristics. Instead, the ability to carry signals depends on the distance, the diameter of the wiring used, and the level of electrical interference. For example, consider two subscribers who live in different parts of a town. If the telephone line leading to the first subscriber passes near a commercial radio station, the station's signal will cause interference at the frequency the station uses. If the second subscriber does not live near the same radio station, the frequency the radio station uses may work well for data on that subscriber's line. However, the second subscriber can experience interference on another frequency. Thus, the ADSL designers could not pick a particular set of carrier frequencies or modulation techniques that would work well in all local loops.

To accommodate differences in local loop characteristics, ADSL is *adaptive*. That is, when a pair of ADSL modems are powered on, they probe the line between them to find its characteristics, and then agree to communicate using techniques that are optimal for the line. In particular, ADSL uses a scheme known as *Discrete Multi Tone modula-*

tion (*DMT*) that combines frequency division multiplexing and inverse multiplexing techniques.

Frequency division multiplexing in DMT is implemented by dividing the bandwidth into *286* separate frequencies called *subchannels*†, with *255* subchannels allocated for downstream data transmission and *31* allocated for upstream data transmission. Two of the upstream channels are reserved for control information. Conceptually, there is a separate "modem" running on each subchannel, which has its own modulated carrier. Carriers are spaced at 4.1325 KHz intervals to keep the signals from interfering with one another. Furthermore, to guarantee that its transmissions do not interfere with analog phone signals, ADSL avoids using the bandwidth below 26 KHz. When ADSL starts, both ends probe the available frequencies to determine which frequencies work well and which experience interference. In addition to selecting frequencies, the two ends assess the signal quality at each frequency, and use the quality to select a modulation scheme. If a particular frequency has a high signal-to-noise ratio, ADSL selects a modulation scheme that encodes many bits per baud; if the quality on a given frequency is low, ADSL selects a modulation scheme that encodes fewer bits per baud. We can summarize:

> *Because the electrical characteristics of local loops vary, ADSL uses an adaptive technology in which a pair of modems probe many frequencies on the line between them, and select frequencies and modulation techniques that yield optimal results on that line.*

12.7 The Data Rate Of ADSL

How fast can ADSL operate? ADSL can achieve a downstream rate of 8.448 Mbps on short local loops, and an upstream rate of 640 Kbps. Because the mandatory network control channel requires 64 Kbps, the effective upstream rate for user data is 576 Kbps. Under the best conditions, ADSL2 can download at close to 20 Mbps.

From a user's point of view, adaptation has an interesting property: ADSL does not guarantee a data rate. Instead, ADSL can only guarantee to do as well as line conditions allow its techniques to operate. Subscribers who live farther from a Central Office or whose local loop passes near sources of interference experience lower data rates than subscribers who live near the Central Office and whose local loop does not pass near sources of interference. Thus, the downstream rate varies from 32 Kbps to 8.448 Mbps, and the upstream rate varies from 32 to 640 Kbps.

It is important to understand that the ADSL data rate only applies to the local loop connection between a subscriber and the telephone Central Office. Many other factors affect the overall data rates that a user experiences. For example, when a user contacts a web server, the effective data rate can be limited by: the speed or current load on the server, the access technology used to connect the server's site to the Internet, or intermediate networks between the user's Central Office and the provider that handles the server.

†The term *subchannel* arises because some DSL variants divide bandwidth into 1.544 Mbps "channels" that each correspond to a T1 circuit as described later in the chapter.

12.8 ADSL Installation And Splitters

Although traditional analog phones operate at frequencies below 4 KHz, lifting a receiver can generate noise that interferes with DSL signals. To provide complete isolation, ADSL uses an FDM device known as a *splitter* that divides the bandwidth by passing low frequencies to one output and high frequencies to another. Interestingly, a splitter is *passive*, which means that it does not require power. A splitter is usually installed at the location where the local loop enters a residence or business. One side of the splitter connects to the POTS wiring and the other side connects to an ADSL modem. Figure 12.6 illustrates the connection.

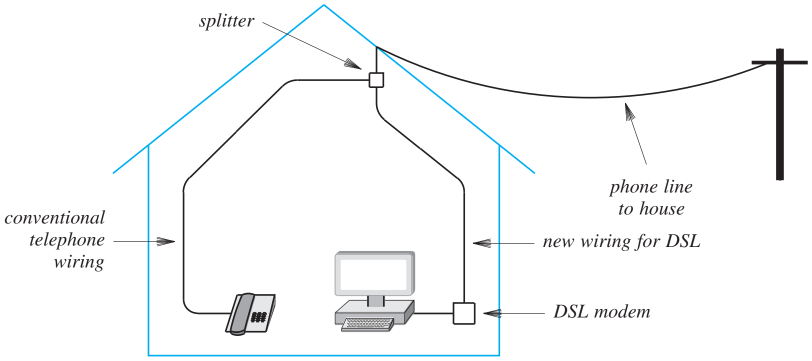


Figure 12.6 Illustration of a splitter and the wiring used with ADSL.

An interesting variation of ADSL wiring has become popular. Sometimes called *DSL lite*, the alternative approach does not require a splitter to be installed on the incoming phone line. Instead, existing house wiring is used for DSL, and a splitter must be installed between each telephone and the house wiring. The advantage of the alternative approach is that a subscriber can install DSL by plugging a splitter into a wall jack and plugging a telephone into the splitter.

12.9 Cable Modem Technologies

Although technologies like ADSL provide data rates that are much higher than originally thought possible, telephone local loop wiring has inherent limitations. The chief problem lies in the electrical characteristics of twisted pair wiring. The lack of shielding makes the wiring susceptible to interference that substantially degrades performance for some subscribers. As demand for higher bit rates increases, alternative wiring schemes have become important. Consequently, a variety of wireless and wired technologies are being developed for use in the local loop.

An alternative access technology that stands out as particularly attractive uses the wiring already in place for *cable television*†. The medium used in cable systems is coaxial cable, which has high bandwidth and is less susceptible to electromagnetic interference than twisted pair. Furthermore, cable television systems use frequency division multiplexing (FDM) to deliver many channels of entertainment simultaneously.

One might assume that with many channels available, a cable provider could use a separate channel to deliver digital information to each subscriber. That is, configure a pair of *cable modems*, one in the CATV center and the other at a subscriber's site, to use a given channel (i.e., carrier frequency) for communication, and multiplex the channel onto the cable along with television signals.

Despite the large bandwidth available in CATV systems, the bandwidth is insufficient to handle a frequency division multiplexing scheme that extends a channel to each user. To understand why, observe that in a dense metropolitan area, a single cable supplier can have millions of subscribers. As a result, using a separate channel per subscriber does not scale.

To solve the problem, cable systems combine FDM and statistical multiplexing by allocating a channel for digital communication for a set of subscribers (typically, everyone in a neighborhood). Each subscriber is assigned a unique *address*, and each message sent over the channel contains the address to which it has been sent. A subscriber's modem listens to the assigned frequency, but before accepting a message, the modem verifies that the address in the message matches the address assigned to the subscriber.

12.10 The Data Rate Of Cable Modems

How fast can a cable modem operate? In theory, a cable system can support data rates of 52 Mbps downstream and 512 Kbps upstream. In practice, the rate can be much less. First, the data rate of a cable modem only pertains to communication between the local cable office and the subscriber's site. Second, the bandwidth is shared among a set of N subscribers, where the size of the set is controlled by the cable provider. From a subscriber's point of view, sharing the bandwidth with other subscribers can be a disadvantage because the effective data rate available to each individual subscriber varies over time. In the worst case, if N subscribers share a single frequency, the amount of capacity available to an individual subscriber will be $1/N$.

12.11 Cable Modem Installation

Because cable systems use FDM, cable modem installation is straightforward. Unlike xDSL technologies that require the use of splitters, cable modems attach to the cable wiring directly. The FDM hardware in existing cable boxes and in cable modems

†Cable television, formally known as *Community Antenna TeleVision* (*CATV*), uses FDM to deliver broadcast television signals to subscribers over coaxial cable. CATV is not available in all countries.

guarantees that data and entertainment channels will not interfere with one another. The point is:

> *Because cable systems use frequency division multiplexing, a cable modem can be attached directly to existing cable wiring without the need for a splitter.*

12.12 Hybrid Fiber Coax

One of the most expensive aspects of access technologies arises from the cost that is incurred in running copper or optical cables between each customer and a provider's facility. Therefore, providers look for technologies that allow them to offer higher-speed services while minimizing the number of physical wires that must be changed. One technology that offers higher data rates without replacing all the wires is known by the general name of *Hybrid Fiber Coax* (*HFC*). As the name implies, a Hybrid Fiber Coax system uses a combination of optical fibers and coaxial cables, with fiber used for the central facilities and coax used for connections to individual subscribers. In essence, an HFC system is hierarchical. It uses fiber optics for the portions of the network that require the highest bandwidth, and uses coax for parts that can tolerate lower data rates. To implement such a system, a provider places devices in each neighborhood that can convert between optical and coaxial cable. Each device connects back to the provider over an optical fiber, and connects to houses in the neighborhood via coaxial cable. If coaxial cables are already in place for cable TV, the cost is minimized. Figure 12.7 illustrates the architecture.

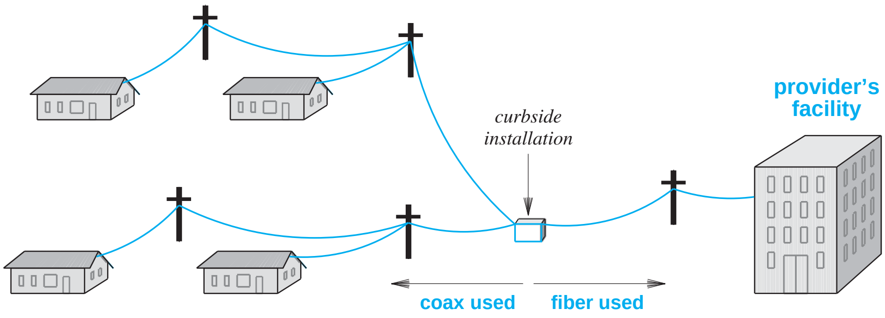


Figure 12.7 Illustration of a Hybrid Fiber Coax access system.

The cable industry uses the term *trunk* to refer to the high-capacity connections between the cable office and each neighborhood area, and the term *feeder circuit* to refer to the connection to an individual subscriber. Trunk connections can be up to *15* miles long; feeder circuits are usually less than a mile.

12.13 Access Technologies That Employ Optical Fiber

Cable companies have proposed a variety of technologies that either employ optical fiber in a hybrid system or deploy optical fiber all the way to each subscriber. Figure 12.8 summarizes names of key technologies.

Name	Expansion
FTTC	Fiber To The Curb
FTTB	Fiber To The Building
FTTH	Fiber To The Home
FTTP	Fiber To The Premises

Figure 12.8 Names of additional access technologies that use optical fiber.

Fiber To The Curb (FTTC). As the name implies, FTTC is similar to HFC because it uses optical fiber for high capacity trunks. The idea is to run optical fiber close to the end subscriber, and then use copper for the feeder circuits. FTTC differs from HFC because it uses two media in each feeder circuit to allow the cable system to provide an additional service, such as voice. The technology is being deployed in some areas, especially in the U.S. and Canada.

Fiber To The Building (*FTTB*). A fundamental question concerns the bandwidth that will be needed by businesses, and whether access technologies using copper (even coaxial cable) will suffice. FTTB is a technology that will use optical fiber to allow high upstream data rates.

Fiber To The Home (*FTTH*). A counterpart of FTTB, FTTH is an access technology that uses optical fiber to deliver higher downstream data rates to residential subscribers. Although FTTH also provides higher upstream data rates, the emphasis is on many channels of entertainment and video.

Fiber To The Premises (*FTTP*). A generic term, FTTP, encompasses both FTTB and FTTH.

12.14 Head-End And Tail-End Modem Terminology

An access technology (using copper wires or optical fibers) requires a pair of modems, with one at the subscriber's site and one at the provider's site. The industry uses the term *head-end modem* to refer to a modem used at the provider's site, and the term *tail-end modem* to refer to a modem used at the subscriber's location.

Head-end modems are not individual devices. Instead, a large set of modems is built as a unit that can be configured, monitored, and controlled together. A set of head-end modems used by a cable provider is known as a *Cable Modem Termination System* (*CMTS*). A set of industry standards known as the *Data Over Cable System In-*

terface Specifications (*DOCSIS*) specifies both the format of data that can be sent as well as the messages used to request services (e.g., on-demand movies).

12.15 Wireless Access Technologies

Although technologies such as ADSL or HFC can deliver digital services to most subscribers, they do not handle all circumstances. The primary problems arise in rural areas. For example, imagine a farm or remote village many miles from the nearest city. The twisted pair wiring used to deliver telephone service to such locations exceeds the maximum distance for technologies like ADSL. In addition, rural areas are least likely to have cable television service.

Even in suburban areas, technologies like ADSL may have technical restrictions on the type of line they can use. For example, it may be impossible to use high frequencies on telephone lines that contain *loading coils*, *bridge taps*, or *repeaters*. Thus, even in areas where a local loop technology works for most subscribers, it may not work on all lines.

To handle special cases, a variety of wireless access technologies have been explored. Figure 12.9 lists a few examples, and Chapter 16 discusses several technologies.

Technology	Description
3G and 4G services	Third and fourth generation cellular telephone data services (e.g., EVDO and LTE)
WiMAX	Wireless access technology up to 155 Mbps using radio frequencies
Satellite	Various commercial vendors offer Internet access services over satellite

Figure 12.9 Examples of wireless access technologies.

12.16 High-Capacity Connections At The Internet Core

Networking professionals say that access technologies handle the *last mile problem*, where the last mile is defined as the connection to a typical residential subscriber or a small business. An access technology provides sufficient capacity for a residential subscriber or a small business (industry uses the term *Small Office Home Office* or *SOHO*). Connections to large businesses or connections among providers require substantially more bandwidth. To differentiate high-speed connections from those found at the edge of the Internet, professionals use the term *core* and refer to high-speed technologies as *core technologies*.

To understand the data rates needed for the core, consider the data transferred by a provider that has 5000 customers. Assume the provider uses an access technology that can provide up to 2 Mbps per customer, and consider what happens if all subscribers attempt to download data at the same time. Figure 12.10 shows the aggregate traffic from the Internet to the provider.

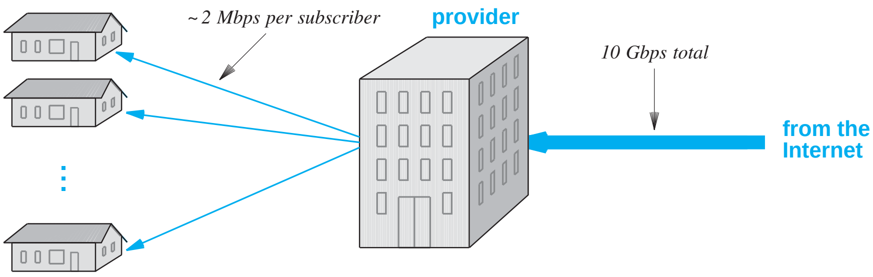


Figure 12.10 Aggregate traffic from the Internet to a provider assuming the provider has 5000 customers each downloading 2 Mbps.

The question arises: what technology can a provider use to move data a long distance at a rate of 10 Gbps? The answer lies in a *point-to-point digital circuit* leased from a telephone company. Although originally designed to be used internally in the phone system, high-capacity digital circuits are available for a monthly fee, and can be used to transfer data. Because telephone companies have the authority to install wiring that crosses municipal streets, a circuit can extend between two buildings, across a city, or from a location in one city to a location in another. The fee charged depends on the data rate of the circuit and the distance spanned. To summarize:

> *Digital circuits leased from common carriers form the fundamental building blocks for long distance data communications. The cost depends on the circuit capacity and distance.*

12.17 Circuit Termination, DSU / CSU, And NIU

To use a leased digital circuit, one must agree to follow the rules of the telephone system, including adhering to the standards that were designed for transmitting digitized voice calls. It may seem that following standards for digitized information would be trivial because computers are digital as well. However, because the computer industry and the telephone industry developed independently, the standards for telephone system digital circuits differ from those used in the computer industry. Thus, a special piece of hardware is needed to interface a computer to a digital circuit provided by a telephone company. Known as a *Data Service Unit/Channel Service Unit* (*DSU/CSU*) the device contains two functional parts, usually combined into a single chassis. The CSU portion of the DSU/CSU device handles line termination and diagnostics. For example, a CSU contains diagnostic circuitry that can test whether the line has been disconnected. A CSU also contains a *loopback* test facility that allows the CSU to transmit a copy of all data that arrives across the circuit back to the sender without further processing.

A CSU provides a service that computer engineers find surprising — it prohibits excessive consecutive *1* bits. The need to prevent excessive *1*s arises from the electrical signals used. In particular, because the telephone company originally designed their digital circuits to work over copper cables, engineers were concerned that having too many contiguous *1* bits would mean excessive current on the cable. To prevent problems, a CSU can either use an encoding that guarantees a balance (e.g., a differential encoding) or a technique known as *bit stuffing*.

The DSU portion of a DSU/CSU handles the data. It translates data between the digital format used on the carrier's circuit and the digital format required by the customer's computer equipment. The interface standard used on the computer side depends on the rate that the circuit operates. If the data rate is less than 56 Kbps, the computer can use RS-232. For rates above 56 Kbps, the computer must use interface hardware that supports higher speeds (e.g., hardware that uses the *RS-449* or *V.35* standards).

The phone company provides one additional piece of equipment, known as a *Network Interface Unit* (*NIU*†), that forms a boundary between equipment owned by the telephone company and equipment provided by the subscriber. The telephone company refers to the boundary as the *demarc*.

> *A digital circuit needs a device known as a DSU/CSU at each end. The DSU/CSU translates between the digital representation used by phone companies and the digital representation used by the computer industry.*

12.18 Telephone Standards For Digital Circuits

A digital circuit leased from a telephone company follows the same digital transmission standards that the phone company uses to transport digital phone calls. In the U.S., standards for digital telephone circuits were given names that consist of the letter *T* followed by a number. Engineers refer to them collectively as the *T-series standards*. One of the most popular is known as T1; many small businesses use a T1 circuit to carry data.

Unfortunately, T-standards are not universal. Japan adopted a modified version of the T-series standards, and Europe chose a slightly different scheme. European standards can be distinguished because they use the letter *E*. Figure 12.11 lists the data rates of several digital circuit standards.

†Although the term *Smartjack* is sometimes used as a synonym for NIU, Smartjack refers to a specific type of NIU manufactured by Westell Corporation.

Name	Bit Rate	Voice Circuits	Location
basic rate	0.064 Mbps	1	
T1	1.544 Mbps	24	North America
T2	6.312 Mbps	96	North America
T3	44.736 Mbps	672	North America
E1	2.048 Mbps	30	Europe
E2	8.448 Mbps	120	Europe
E3	34.368 Mbps	480	Europe

Figure 12.11 Examples of digital circuits and their capacity.

12.19 DS Terminology And Data Rates

Recall from Chapter 11 that the telephone companies use a multiplexing hierarchy that combines multiple voice calls into a single digital circuit. Thus, the data rates of T standards have been chosen so they can each handle a multiple of voice calls. The important thing to note is that the capacity of circuits does not increase linearly with their numbers. For example, the T3 standard defines a circuit with much more than three times the capacity of T1. Finally, it should be noted that phone companies do lease circuits with lower capacity than those listed in the figure; they are known as *fractional T1* circuits.

To be technically precise, one must distinguish between the T-standards, which define the underlying carrier system, and the standards that specify how to multiplex multiple phone calls onto a single connection. The latter are known as *Digital Signal Level standards* or *DS standards*. The names are written as the letters *DS* followed by a number, analogous to the T-standards. For example, DS1 denotes a service that can multiplex 24 phone calls onto a single circuit, and T1 denotes a specific standard that does so. Because DS1 defines the effective data rate, it is technically more accurate to say, “a circuit running at DS1 speed” than to refer to “T1 speed.” In practice, few engineers bother to distinguish between T1 and DS1. Thus, one is likely to hear someone refer to “T1-speed”.

12.20 Highest Capacity Circuits (STS Standards)

Telephone companies use the term *trunk* to denote a high-capacity circuit, and have created a series of standards for digital trunk circuits. Known as the *Synchronous Transport Signal* (*STS*) standards, they specify the details of high-speed connections. Figure 12.12 summarizes the data rates associated with various STS standards. All data rates in the table are given in Mbps, making it easy to compare. It should be noted that data rates for STS-24 and above are greater than 1 Gbps.

Copper Name	Optical Name	Bit Rate	Voice Circuits
STS-1	OC-1	51.840 Mbps	810
STS-3	OC-3	155.520 Mbps	2430
STS-12	OC-12	622.080 Mbps	9720
STS-24	OC-24	1,244.160 Mbps	19440
STS-48	OC-48	2,488.320 Mbps	38880
STS-192	OC-192	9,953.280 Mbps	155520

Figure 12.12 Data rates of digital circuits according to the STS hierarchy of standards.

12.21 Optical Carrier Standards

In addition to STS standards, the phone company defines an equivalent set of *Optical Carrier* (*OC*) standards. Figure 12.12 gives the names for optical standards as well as for copper standards. To be precise, one should observe a distinction between the STS and OC terminology: the STS standards refer to the electrical signals used in the digital circuit interface (i.e., over copper), while the OC standards refer to the optical signals that propagate across the fiber. As with other network terminology, few professionals make the distinction. Thus, one often hears networking professionals use the term *OC-3* to refer to a digital circuit that operates at 155 Mbps, independent of whether the circuit uses copper or optical fiber.

12.22 The C Suffix

The Synchronous Transport Signal and Optical Carrier terminology described above has one additional feature not shown in Figure 12.12: an optional suffix of the letter *C*, which stands for *concatenated*. The presence of the suffix denotes a circuit with no inverse multiplexing. That is, an OC-3 circuit can consist of three OC-1 circuits operating at 51.840 Mbps each, or it can consist of a single OC-3C (STS-3C) circuit that operates at 155.520 Mbps.

Is a single circuit operating at full speed better than multiple circuits operating at lower rates? The answer depends on how the circuit is being used. In general, having a single circuit operating at full capacity provides more flexibility and eliminates the need for inverse multiplexing equipment. More to the point, data networks are unlike voice networks. In a voice system, high-capacity circuits are used as a way of aggregating smaller voice streams. In a data network, however, there is a single stream of data traffic. Thus, if given a choice, most network designers prefer an OC-3C circuit over an OC-3 circuit.

12.23 Synchronous Optical Network (SONET)

In addition to the STS and OC standards described above, the phone companies defined a broad set of standards for digital transmission. In North America, the standards are known by the term *Synchronous Optical NETwork* (*SONET*), while in Europe they are known as the *Synchronous Digital Hierarchy* (*SDH*). SONET specifies details such as how data is framed, how lower-capacity circuits are multiplexed into a high-capacity circuit, and how synchronous clock information is sent along with data. Because carriers use SONET extensively, when someone leases an STS-1 circuit, the carrier is likely to require them to use SONET encoding on the circuit. For example, Figure 12.13 shows the SONET frame format used on an STS-1 circuit.

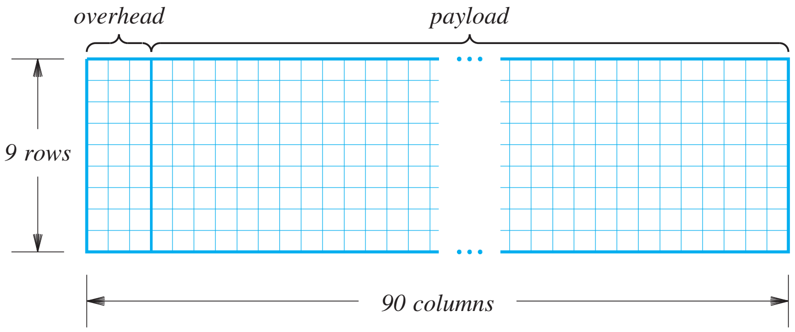


Figure 12.13 Illustration of a SONET frame when used over an STS-1 circuit.

Each frame is 810 octets long. According to SONET terminology, octets in the frame are divided into 9 “rows”, with 90 “columns” in each row. Interestingly, the size of a SONET frame depends on the bit rate of the underlying circuit. When used on an STS-3 circuit, each SONET frame holds 2430 octets. How do the numbers arise? To understand the difference, recall that digital telephony takes 8000 PCM samples per second, which means that a sample is taken every 125 μ seconds. SONET uses the time to define frame size. At the STS-1 transmission rate of 51.840 Mbps, exactly 6480 bits are transferred in 125 μ seconds, which means that a frame consists of 810 8-bit octets. Similarly, at the STS-3 rate, 2430 octets can be transmitted in 125 μ seconds. The chief advantage of making the frame size depend on the bit rate of the circuit is that it makes synchronous multiplexing trivial — retaining synchronization while combining three STS-1 SONET streams into one STS-3 SONET stream is straightforward.

Although most data networks use SONET as an encoding scheme on a single point-to-point circuit, the standard provides more possibilities. In particular, it is possible to build a high-capacity counter rotating ring network using SONET technology that handles single-point failures. Each station on the ring uses a device known as an *add/drop mux*. In addition to passing received data around the ring, the add/drop mux can be configured to accept additional data from a local circuit and add it to frames passing across the ring or to extract data and deliver it to a local computer. If the ring is broken, the hardware detects the loss of framing information and uses the counter rotating ring to reconnect. To summarize:

> *Although the SONET standard defines a technology that can be used to build a high-capacity ring network with multiple data circuits multiplexed across the fibers that constitute the ring, most data networks only use SONET to define framing and encoding on a leased circuit.*

12.24 Summary

Access technologies provide Internet connections to individual residences or small businesses. A variety of access technologies exist, including dialup telephone connections, wireless (using radio frequency or satellite), and wired. Two current access technologies are Digital Subscriber Line (DSL) and cable modems. DSL uses FDM techniques to allow digital communication and a traditional analog voice call to proceed simultaneously on the local loop between a telephone company Central Office and a subscriber. Cable modem service uses FDM to multiplex digital communication over the same coaxial cable system used to carry entertainment channels. When using cable modem technology, cable modems in each neighborhood employ statistical multiplexing to share a single data communications channel.

Technologies like Hybrid Fiber Coax (HFC) and Fiber To The Curb (FTTC) use optical fibers to distribute data to each neighborhood and use coaxial cable to reach an individual subscriber. Future technologies have been proposed that will use optical fiber to deliver higher data rates to each individual residence.

Although they suffice for individual residences and small businesses, access technologies do not provide sufficient capacity for use in the core of the Internet. To achieve the highest data rates over long distances, service providers and large businesses lease point-to-point circuits from common carriers. Digital circuits use time division multiplexing standards (T-standards in North America or E-standards in Europe). High-speed circuits are defined using the Synchronous Transport Signal (North America) or Synchronous Digital Hierarchy (Europe). A parallel set of Optical Carrier standards exists for use with optical fiber; many professionals use the OC standard names, independent of whether the circuit uses fiber or copper.

A telephone company standard known as SONET defines framing for use on a digital circuit. The size of a SONET frame depends on the bit rate of the circuit; one frame always takes 125 μseconds to send. In addition to its use on point-to-point circuits, SONET can be configured into a ring, which allows hardware to determine if the ring is broken and automatically reconfigure around the malfunction.

EXERCISES

12.1 Why do service providers distinguish between upstream and downstream communication?

12.2 What is an *access technology*?

12.3 Telephone companies once promoted ISDN as a high-speed access technology. Why has use of ISDN declined?

12.4 Give examples of narrowband and broadband access technologies.

12.5 What type of multiplexing does ADSL use?

12.6 If a customer intends to transmit more data than they receive, which forms of DSL would be appropriate? If they intend to receive more data than they transmit?

12.7 Why is a splitter used with DSL?

12.8 Two neighbors, who live on the same street, both use ADSL service, but measurements show that one subscriber can download at approximately 1.5 Mbps and the other can download at 2.0 Mbps. Explain.

12.9 Why would a service provider choose Hybrid Fiber Coax instead of Fiber To The Premises?

12.10 If you had a choice between DSL and cable modem, which would provide the highest potential data rate?

12.11 What is the advantage of WiMAX access technology compared to satellite? What is the advantage of satellite?

12.12 Where is a head-end modem located? A tail-end modem?

12.13 Use the Web to find the approximate size of a movie on DVD. How long does it take to download a movie over a T1 line? Over a T3 line? (Ignore overhead.)

12.14 If you lease a T1 circuit, what equipment will be installed between the circuit and a computer at your site?

12.15 Explain how the size of a SONET frame is computed.

12.16 Why did the designers of the Synchronous Digital Hierarchy choose unusual values for data rates instead of exact powers of ten?

12.17 If someone shows you a copper cable and claims that it is an “OC-12 circuit”, what error have they made? What is the correct name they should have used?