Exercises

1. Do the following:
   - Download Wireshark.
   - Start Wireshark.
   - Turn on Wireshark capture.
   - Type a URL in your browser window (not Wikipedia.org).
   - After a few seconds, stop the capture.
   - Answer the following questions:
     1a. What URL did you use? What was the IP address of the webserver?
     1b. Find the frame in which your PC sent the SYN packet. List the source and destination IP address, the source and destination port numbers, and the header checksum.
     1c. Select the SYN/ACK packet. List the source and destination IP address, the source and destination port numbers, and the header checksum.
     1d. Select the packet that acknowledges the SYN/ACK segment. List the source and destination IP address, the source and destination port numbers, and the header checksum.
2. Change the options so that only packets you send are recorded. Do a capture. Click on the window containing Wireshark and hit *Alt-Enter*. This captures the window to your clipboard. Paste it into your homework.

3 NETWORK SECURITY

LEARNING OBJECTIVES

By the end of this chapter, you should be able to:

- Describe the threat environment, including types of attackers and types of attacks.
- Explain the Plan–Protect–Respond cycle for security management.
- Explain in detail planning principles and policy-based security.
- Describe protection.
  - Evaluate authentication mechanisms, including passwords, smart cards, biometrics, digital certificate authentication, and two-factor authentication.
  - Describe firewall protection, including stateful inspection.
  - Explain in detail the protection of dialogues by cryptography, including symmetric key encryption for confidentiality, electronic signatures, and cryptographic system standards.
- Describe response: Reacting according to plan for successful compromises and disasters.

STEUBEN ARC

Steuben ARC is a nonprofit organization in Bath, New York, that provides care for developmentally disabled adults. In September 2009, cyberthieves stole nearly $200,000 from the company.[1]

The attack began when a cybercriminal sent a fake invoice in an e-mail message to one of the company's accountants. The message had an attachment, dhlinvoice.zip. When the accountant opened the attachment, it installed a very sophisticated keystroke logger on the accountant's computer. This program captured the accountant's username and password on the company's accounting server and sent it to the attacker.

Armed with this information, the thieves transferred the money out of the company's bank accounts in two batches. Instead of sending it to themselves, the thieves had the banks send the money to 20 money mules around the country. These money mules forwarded the money to offshore accounts controlled by the attackers. For each transaction,

[1]Brian Krebs, "Cyber Gangs Hit Healthcare Providers," *Washington Post*, September 28, 2009. voices.washingtonpost.com/securityfix/2009/09/online_bank_robbers_target_hea.html?wprss=securityfix.Mary Pernham, "Alleged cyber-theft: Hackers take $50K from Arc," *The Corning Leader*, October 1, 2009. www.the-leader.com/news/x1699607673/Alleged-cyber-theft-Hackers-take-50K-from-Arc.

the mules received a fee. Using money mules allowed the attackers to avoid shipping the money directly to offshore accounts, which could have raised the bank's suspicions.

The bank actually did become suspicious. It blocked some of the transfers to money mules and by money mules to offshore banks. However, only some of the money was recovered. Overall, it was a successful attack.

Test Your Understanding

1. a) How did the attacker get the credentials for the company's bank account? b) Why were money mules used? c) List indications that this was a sophisticated attack. d) How might the company have been able to avoid this compromise? e) What motivated the attacker? f) What would you say to executives in small companies who believe that they are too little to be attacked?

INTRODUCTION

This is the third of four introductory chapters. The fact that it deals entirely with security tells you how important security has become in networking. In the 1990s, the Internet blossomed, allowing billions of people to reach hundreds of millions of servers around the world. Unfortunately, the Internet also made all of these users potential victims. Security quickly became one of the most important IT management issues.

Figure 3-1 shows security threats and the plan–protect–respond cycle that companies follow to deal with the threat environments. In this chapter, we will begin looking at the threat environment. Sun Tzu's *The Art of War* admonishes defenders to "know your enemy." We will see that the **threat environment**—the attacks and attackers that companies face—is complex and rapidly changing.

The **Plan–Protect–Respond cycle** that companies follow to deal with the threat environment begins with the planning that companies must do to defend against these threats. This is followed by the Protect phase, in which companies implement the

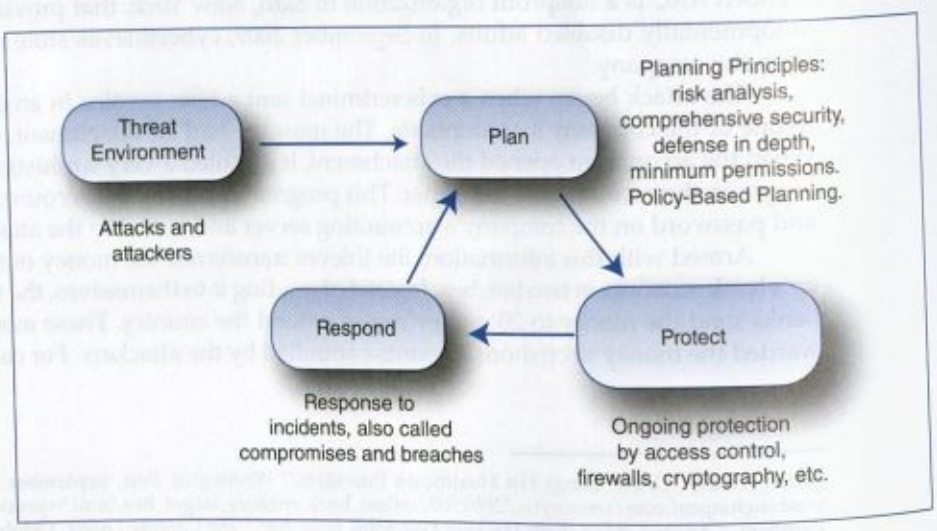

FIGURE 3-1 Threats and the Plan–Protect–Respond Cycle

protections they have planned. The Protect box is larger than the others to emphasize that this phase accounts for most of the work in security. The Respond stage is needed when protections fail and attacks succeed. Successful attacks are called **compromises**, **incidents**, or **breaches**.

The main thing that sets security management apart from other aspects of IT management is that the company must battle against intelligent adversaries, not simply against errors and other forms of unreliability. Companies today are engaged in an escalating arms race with attackers, and security threats and defenses are mutating at a frightening rate.

In this chapter, we will look at security broadly. We will focus on network security, but network security is impossible to separate from general IT security.

Test Your Understanding

2. a) What are the two elements of the threat environment? b) Briefly explain each of the three stages in the plan–protect–respond cycle. c) Which of these three stages consumes the most corporate effort? d) Give three names for *successful attack*. e) What is the main thing that separates security from other aspects of IT?

TYPES OF ATTACKS

As just noted, we will begin by looking at the threat environment that corporations face. In this section, we will look at types of attacks. Later, we will look at the types of attackers.

Malware

A general name for evil software

Vulnerability-Specific versus Universal Malware

Vulnerabilities are flaws in specific programs
Vulnerabilities allow attacks against these specific programs
Vulnerability-specific malware requires a specific vulnerability to be effective
Vendors release patches to close vulnerabilities
However, users do not always install patches promptly or at all and so continue to be vulnerable
Also, zero-day attacks occur before the patch is released for the vulnerability
Universal malware does not require a specific vulnerability to be effective
Universal malware often requires risky human actions

Viruses

Pieces of code that attach themselves to other programs
Virus code executes when an infected program executes
The virus then infects other programs on the computer
Propagation vectors
E-mail attachments
Visits to websites (even legitimate ones)

(continued)

FIGURE 3-2 Malware (Study Figure)

- Social networking sites
- Many others (USB RAM sticks, peer-to-peer file sharing, etc.)

Stopping viruses

- Antivirus programs are needed to scan arriving files for viruses
- Antivirus programs also scan for other malware
- Patching vulnerabilities may help

Worms

- Stand-alone programs that do not need to attach to other programs
- Can propagate like viruses through e-mail, etc.
  - This requires human gullibility, which is slow
- Directly-propagating worms jump to victim hosts directly
  - Can do this if target hosts have a specific vulnerability
- Directly-propagating worms can spread with amazing speed
- Directly-propagating worms can be thwarted by firewalls and by installing patches
  - *Not* by antivirus programs

Mobile Code

- HTML webpages can contain scripts
  - Scripts are snippets of code that are executed when the webpage is displayed in a browser
  - Scripts enhance the user experience and may be required to see the webpage
  - Scripts are called mobile code because they are downloaded with the webpage
- Scripts are normally benign but may be damaging if the browser has a vulnerability
  - The script may do damage or download a program to do damage

Payloads

- After propagation, viruses and worms execute their payloads
- Payloads erase hard disks or send users to pornography sites if they mistype URLs
- Often, the payload downloads another program
  - An attack program with such a payload is called a downloader
- Many downloaded programs are Trojan horses
  - Trojan horses are programs that disguise themselves as system files
- Spyware Trojans collect sensitive data and send the data to an attacker
  - Website activity trackers
  - Keystroke loggers
  - Data mining software

Getting Infected

- E-mail from infected machines or spammers
- Visiting websites
  - Even normally legitimate websites can be seeded with pages containing mobile malware
- Peer-to-peer file transfers
- Downloading "free" software
- Etc.

FIGURE 3-2 Continued

Malware Attacks

We will begin with malware attacks. **Malware** is a name for any evil software. This includes viruses, worms, Trojan horses, and other dangerous attack software. Malware attacks are the most frequent attacks on companies. Nearly every firm has one or more significant malware compromises each year.

Malware is evil software.

Test Your Understanding

3. What is malware?

VULNERABILITY-SPECIFIC VERSUS UNIVERSAL MALWARE A **vulnerability** is a flaw in a program that permits a specific attack or set of attacks against this program. If the vulnerability is not present in the program, a **vulnerability-specific** attack aimed at that vulnerability will fail.

A vulnerability is a flaw in a program that permits a specific attack or set of attacks against this program.

When a vulnerability is discovered, the software vendor usually issues a **patch**, which is a small program designed to correct the vulnerability. After patch installation, the program is safe from attacks based on that particular vulnerability. Too often, however, users fail to install patches, and their programs continue to be vulnerable. Even if they do install patches, furthermore, they may delay doing so. This creates a long window of opportunity for attackers.

Of course, if attacks begin before the program vendor creates a patch (or even learns about the attack), the attacks will succeed. A vulnerability-specific attack that occurs before a patch is available is called a **zero-day attack**.

A vulnerability-specific attack that occurs before a patch is available is called a zero-day attack.

Not all malware is vulnerability-specific. **Universal malware** works whether or not the computer has a security vulnerability. In general, universal malware programs require the human victim to do something risky, such as downloading "free" software, pornography, or an electronic greeting card.

Test Your Understanding

4. a) What is a vulnerability? b) How can users eliminate a vulnerability in one of their programs? c) What name do we give to attacks that occur before a patch is available? d) What type of malware does not require a vulnerability?

VIRUSES Pieces of executable code that attach themselves to other programs are called **viruses**. Within a computer, whenever an infected program runs (executes), the virus attaches itself to other programs on that computer.

Viruses are pieces of executable code that attach themselves to other programs on that computer.

Propagation Vectors The virus spreads between computers when an infected program is transferred to another computer via a USB RAM stick, an e-mail attachment, a webpage download, a peer-to-peer file-sharing transfer, a social networking site, or some other **propagation vector** (method for malware to move to a victim computer). Once on another machine, if the infected program is executed, the virus spreads to other programs on that machine.

More than 90 percent of viruses today spread via e-mail. Viruses find addresses in the infected computer's e-mail directories. They then send messages with infected attachments to all of these addresses. If a receiver opens the attachment, the infected program executes, and the receiver's programs become infected.

Another popular propagation vector is visiting a website and having the website download a virus (or other type of malware) to your computer. Obviously, the risk is greatest if you visit a high-risk website, such as a site for "free" software or pornography. However, even if you visit a known legitimate website daily, you may become infected one day if an attacker has planted malware on its webpages. (In 2009, this happened to subscribers who went to the *New York Times* website.) A substantial fraction of all infected websites are legitimate websites.

Social networking sites are already popular with virus writers. By their nature, social networking sites are designed for sharing, and if a malware writer can inject malware into the sharing process, spread can be very rapid. USB RAM sticks and peer-to-peer file transfers are two examples of these other propagation vectors.

Stopping Viruses To stop viruses, a company must protect its computers with **antivirus programs** that scan each arriving e-mail message or file for patterns that identify viruses. Antivirus programs today also scan for other types of malware, but we still call them antivirus programs.

Antivirus programs also scan for other types of malware.

For some viruses (but not for all), it is also useful to patch security vulnerabilities. However, patching does not work for most viruses.

Firewalls are devices that examine each packet passing through a certain part of a network. Client PCs can have firewalls. These firewalls have major benefits, but they do not stop normally propagating worms and viruses.

Test Your Understanding

5. a) What is a propagation vector? b) How do viruses propagate within computers? c) How do viruses propagate between computers? d) In what two ways can viruses be stopped? e) Do firewalls usually stop viruses?

WORMS Another important type of malware is worms. We have just seen that viruses are pieces of code that must attach themselves to other programs. In contrast, **worms** are full programs that operate by themselves. Both viruses and worms can create mass epidemics that infect hundreds or even millions of computers.

Worms are full programs that operate by themselves.

Worms are capable of propagating, like viruses, through e-mail attachments, USB RAM sticks, and similar propagation vectors. These methods typically require human gullibility to succeed. Although human gullibility is sadly reliable, it is rather slow. Until someone opens an e-mail attachment, inserts an infected USB RAM stick, or takes some other action, nothing happens.

Unlike viruses, some (but not all) worms have another propagation vector. A **directly propagating** worm tries to jump from the infected computer to many other computers. Target computers that have a specific vulnerability will accept the directly propagating worm. They then become sites from which the worm spreads further.

A directly propagating worm tries to jump from the infected computer to many other computers.

Freed from the need for human intervention, directly propagating worms can spread with incredible speed. In 2003, the Blaster worm infested 90 percent of all vulnerable hosts on the entire Internet within 10 minutes. The nightmare scenario for security professionals is the prospect of a fast-spreading worm that exploits a vulnerability in a large percentage of all computers on the Internet.

Antivirus programs do nothing to stop directly propagating worms. However, firewalls and patching vulnerabilities can stop them.

Antivirus programs do nothing to stop directly propagating worms. However, firewalls and patching vulnerabilities can stop them.

Figure 3-3 summarizes the differences between how viruses and worms can be stopped. Note that directly propagating worms cannot be stopped using techniques used to stop traditionally propagating worms.

Propagation Vector	Antivirus Program	Firewall	Patching Vulnerabilities
Normally propagating virus or worm (e-mail, visiting website, etc.)	Yes	No	Sometimes
Directly-propagating worm	No	Yes	Yes

FIGURE 3-3 Stopping Viruses and Worms

Test Your Understanding

6. a) How do viruses and worms differ? b) Distinguish how directly propagating worms and e-mail worms spread. c) Which can spread faster—viruses or directly propagating worms? Explain. d) How can directly propagating worms be stopped? e) Can antivirus programs usually stop directly propagating worms?

MOBILE CODE An HTML webpage can contain a **script**, which is a group of commands written in a simplified programming language. Scripts are executed when the webpage is loaded. Scripts can enhance the user's experience, and many webpages will not work unless script execution is enabled, which it usually is by default.

Scripts are referred to as **mobile code** because they travel with the downloaded webpage from the webserver to the browser. Mobile code normally is safe and beneficial. However, if the user's browser has a vulnerability, a script may be able to do harm. A script may do damage itself or may download a more complex program to do damage.

Test Your Understanding

7. a) What is a script? b) Are scripts normally bad? c) Under what circumstances are scripts likely to be dangerous? d) Why are scripts on webpages called mobile code?

PAYLOADS In war, when a bomber aircraft reaches its target, it releases its payload of bombs. Similarly, after they spread, viruses, worms, and other types of malware may execute pieces of code called **payloads**. Malicious payloads can completely erase hard disks and do other significant damage. In some cases, they can take the victim to a pornography site whenever the victim mistypes a URL. In other cases, they can turn the user's computer into a spam generator or a pornography download site. Not all malware has malicious payloads or payloads at all.

TROJAN HORSES Often, the payload installs another program on the computer. A program that does this is called, as you might suspect, **downloader**.

Often, the downloader retrieves and installs a **Trojan horse**, which is a program that disguises itself as a legitimate system file. This makes it difficult to detect. A Trojan horse cannot spread from one computer to another by itself. Rather, it relies on a virus, worm, hacker, or gullible user to install it on a computer. Once installed, the Trojan horse continues to exploit the user indefinitely.

A Trojan horse cannot spread from one computer to another by itself.

SPYWARE An especially problematic category of Trojan horses is **spyware**—a name given to Trojan horses that **surreptitiously** (without your knowledge) collect information about you and send this information to the attacker.

- Some spyware Trojans collect information about your Web surfing habits and send this information to advertisers.
- More dangerous are keystroke loggers, which record your keystrokes. Within these keystrokes, they look for passwords, social security numbers, and

other information that can help the person who receives the keystroke logger's data.
- Data mining spyware, in contrast, searches through files on your hard drive for potentially useful information and sends this information to the attacker.

Test Your Understanding

8. a) What are payloads? b) What are Trojan horses? c) How do Trojan horses propagate to computers? d) What is spyware? e) What is a keystroke logger? f) What does data mining software do?

Attacks on Individuals

SOCIAL ENGINEERING As technical defenses have improved, malware writers have focused more heavily on **social engineering**, which is a fancy name for tricking the victim into doing something against his or her interests. Viruses and worms have long tried to do this with e-mail attachments—say, by telling the user that he or she has won a lottery and needs to open the attachment for the details. The range of social engineering attacks has expanded greatly in the last few years.

Social engineering is tricking the victim into doing something against his or her interests.

SPAM The most annoying type of malware on a day-in, day-out basis is **spam**,[2] which is unsolicited commercial e-mail. Spammers send the same solicitation e-mail message to millions of e-mail addresses in the hope that a small percentage of all recipients will respond.

Spam is unsolicited commercial e-mail.

FRAUD Spam is not merely annoying. Attackers use spam to perpetrate damaging attacks. Few spam messages are really designed to sell legitimate products. Most are fraudulent attempts to get someone to send money for "investment opportunities" and goods that will not be delivered or that are effectively worthless. **Fraud** is lying to get victims to do something against their financial self-interest. *Social engineering* is the more general term; fraud is social engineering applied to financial interests.

Fraud is lying to get victims to do something against their financial self-interest.

E-MAIL ATTACHMENTS Some spam messages have damaging e-mail attachments. For instance, a spam message may say that it is an electronic greeting card. The user is told that a program must be downloaded to read the greeting card. The "reader" program, of course, is malware.

[2]Except at the beginnings of sentences, e-mail *spam* is spelled in lowercase. This distinguishes unsolicited commercial e-mail from the Hormel Corporation's meat product, Spam, which should always be capitalized. In addition, Spam is *not* an acronym for "spongy pink animal matter."

Social Engineering

Tricking the victim into doing something against his or her interests

Spam

Unsolicited commercial e-mail

Fraud

Spam often asks the victim to send money for products that will not be delivered

Or for false investment opportunities

E-Mail Attachments

Including a Link to a Website That Has Malware

The website may complete the fraud or download software to the victim

Phishing Attacks

A sophisticated social engineering attack in which an authentic-looking e-mail or website entices the user to enter his or her username, password, or other sensitive information

Credit Card Number Theft

Uses stolen credit card numbers in unauthorized transactions

Performed by carders

Identity Theft

Involves collecting enough data to impersonate the victim in large financial transactions

- Purchase a house or car
- Obtain a loan
- Commit fraud or other crime

Can result in much greater financial harm to the victim than carding

May take a long time to restore the victim's credit rating

In corporate identity theft, the attacker impersonates an entire corporation

- Accept credit cards in the company's name
- Commit other crimes in the name of the firm
- Can seriously harm a company's reputation

FIGURE 3-4 Attacks on Individuals (Study Figure)

INCLUDING A LINK TO A WEBSITE THAT HAS MALWARE Spam can also take victims to dangerous websites. One way for spam to create problems is for messages to include a link to a website. If the receiver clicks on the link, he or she will be taken to a website that will complete the fraud or download malware into the victim's computer.

PHISHING ATTACKS An especially effective form of spam does phishing,[3] which is the use of authentic-looking e-mail or websites to entice the user to send his or her

[3]IT attackers often replace *f* with *ph*. For example, *phone freaking* (dialing long-distance numbers illegally) became *phone phreaking* and later just *phreaking*.

username, password, or other sensitive information to the attacker. One typical example of phishing is an e-mail message that appears to be from the person's bank. The message asks the person to "confirm" his or her username and password in a return message. Another typical example is an e-mail message with a link to what appears to be the victim's bank website but that is, in fact, an authentic-looking fake website.

Phishing is the use of authentic-looking e-mail or websites to entice the user to send his or her username, password, or other sensitive information to the attacker.

CREDIT CARD NUMBER THEFT In fraudulent spam, the message may convince the user to type a credit card number to purchase goods. The attacker will not deliver the goods. Instead, the **carder** (credit card number thief) will use the credit card number to make unauthorized purchases. Most credit card firms will refund money spent by the carder, but this can be a painful process, and the victim must notify the credit card firm promptly to get a refund.

IDENTITY THEFT In other cases, thieves collect enough data about a victim (name, address, social security number, driver's license number, date of birth, etc.) to impersonate the victim during complex crimes. This impersonation is called **identity theft**. Thieves commit identity theft in order to purchase expensive goods, take out major loans using the victim's assets as collateral, commit crimes, obtain prescription medicines, get a job, enter the country illegally, and do many other things. Identity theft is more damaging than credit card theft because it can involve large monetary losses and because restoring the victim's credit rating can take months. Some victims have even been arrested for crimes committed by the identity thief.

In identity theft, thieves collect enough data about a victim to impersonate the victim during complex crimes.

Test Your Understanding

9. a) What is social engineering? b) What is fraud? c) What is the definition of spam? d) How can spam be used to harm people who open spam messages? e) What is phishing? f) Distinguish between credit card number theft and identity theft. g) What are carders? h) Which tends to produce more damage—credit card theft or identity theft? Explain your answer.

Human Break-Ins (Hacking)

A virus or worm typically has a single method. If that method fails, the attack fails. However, human attackers often break into a specific company's computers manually. A human adversary can attack a target company with a variety of approaches until one succeeds. This flexibility makes human break-ins much more likely to succeed than malware break-ins.

WHAT IS HACKING? **Hacking** is defined as intentionally using a computer resource without authorization or in excess of authorization. The key issue is authorization.[4] If you see a password written on a note attached to a computer screen, this does not mean

Human Break-Ins

Viruses and worms only have a single attack method

Humans can keep trying different approaches until they succeed

Hacking

Informally, hacking is breaking into a computer

Formally, hacking is intentionally using a computer resource without authorization or in excess of authorization

Scanning Phase

Send attack probes to map the network and identify possible victim hosts (Figure 3-6)

Scan for IP addresses with active hosts

Scan IP addresses that reply for programs for which the attacker has an attack method

The Break-In

Uses an exploit—a tailored attack method that is often a program (Figure 3-6)

Normally exploits a vulnerability on the victim computer

The act of breaking in is called an exploit

The hacker tool is also called an exploit

After the Break-In

The hacker downloads a hacker tool kit to automate hacking work

The hacker becomes invisible by deleting log files

The hacker creates a backdoor (way to get back into the computer)

- Backdoor account—account with a known password and full privileges
- Backdoor program—program to allow reentry; usually Trojanized

The hacker can then do damage at his or her leisure

- Download a Trojan horse to continue exploiting the computer after the attacker leaves
- Manually give operating system commands to do damage

FIGURE 3-5 Human Break-Ins (Study Figure)

that you have authorization to use it. Also, note that it is hacking even if a person is given an account but uses the computer for unauthorized purposes.

Hacking is intentionally using a computer resource without authorization or in excess of authorization.

THE SCANNING PHASE When a hacker attacks a firm, he or she usually begins by **scanning** the network. Figure 3-6 shows that this involves sending **probe packets** into the firm's network. Responses to these probe packets tend to reveal information about the

[4]Note also that the unauthorized access must be intentional. Proving intentionality is almost always necessary in criminal prosecution, and hacking is no exception.

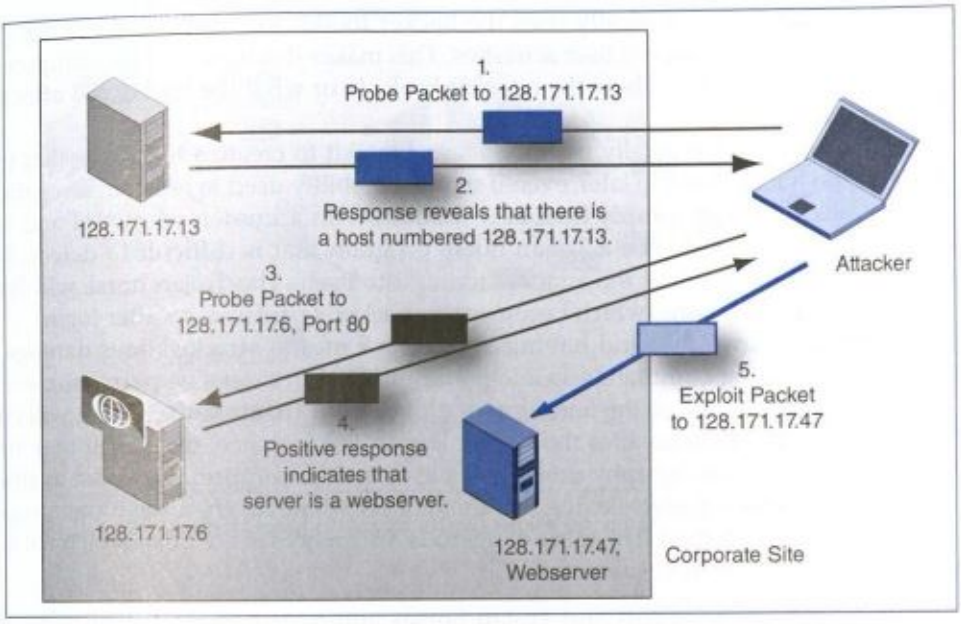

FIGURE 3-6 Scanning Probes and Exploit Packets

firm's general network design and about its individual hosts. Usually there are two phases to these probe attacks.

- The first probe packet in the figure is an IP address probe. It is sent to the IP address 128.171.17.13. If the host at that IP address responds, this means that there is a potential victim at that IP address. The attacker typically probes a large range of IP addresses to get a list of potential victims.
- The attacker then sends port number probes to previously identified IP addresses. This second round of probes is sent to particular ports on these hosts. In the figure, the probe packet is sent to Port 80. As we saw in Chapter 2, this is the well-known port number for webservers. If the server responds, the attacker knows that Host 128.171.17.6 is a webserver.

THE BREAK-IN The colored server is a webserver that the attacker has previously probed. The attacker has an **exploit** (attack method) for webservers. He or she uses this exploit to take over the host by sending exploit packets. Confusingly, the act of breaking into a computer is also called an exploit, as is the program the attacker uses during the break-in.

AFTER THE BREAK-IN After the break-in, the real work begins.

- Typically, the first thing a hacker does is download[5] a **hacker toolkit** to the victim computer. The toolkit is a collection of tools that automate some tasks the hacker will have to perform after the break-in.

[5]Some students find the use of the term *download* to be confusing. Look at it this way. The hacker is now logged into the victim computer. So he or she downloads software from a toolkit server to the victim computer and installs the software on the victim computer.

- Second, the hacker typically uses the hacker toolkit to erase the operating system's log files that record user activities. This makes it difficult for the computer's rightful owner to trace how the attacker broke in or what the hacker did after the break-in.
- Third, the hacker typically uses the hacker toolkit to create a **backdoor** that will allow the hacker back in later, even if the vulnerability used to break in is repaired. The backdoor may simply be a new account with a known password and full privileges. It can also be a Trojan horse program that is difficult to detect. The Trojan horse will allow the attacker to log into itself. The Trojan horse will have extensive permissions, which become the attacker's permissions after login.
- Fourth, once invisible and having a way back in, the attacker does damage at leisure by giving commands as a logged-in user with extensive permissions. For long-term exploitation, the hacker may download a Trojan horse, which will continue to cause damage after the hacker leaves. For instance, the Trojan may turn the host into a pornography download site or use the compromised host to attack other computers. Keystroke loggers that collect whatever the user types are also popular Trojan horses. The most dangerous Trojan horses are bots, which we will learn about in the next subsection.

Although hacker toolkits and Trojan horses automate a great deal of what the hacker wishes to do, hackers also work manually. With full access to the computer, the attacker can give ordinary operating system commands to read any file on the computer, change files, delete them, or do anything else that a legitimate user with extensive permissions can do.

Test Your Understanding

10. a) List the three main phases in human break-ins (hacks). b) What is hacking? c) What are the two purposes of probe packets? d) What is an exploit? e) What steps does a hacker usually take immediately after a break-in? f) What software does the hacker download to help him or her do work after compromising a system? g) After breaking in, what does a hacker do to avoid being caught? h) What is a backdoor? i) What are the two types of backdoors?

Denial-of-Service (DoS) Attacks Using Bots

Another type of attack, the denial-of-service attack, does not involve breaking into a computer, infecting it with a virus, or infesting it with a worm. Rather, the goal of **denial-of-service (DoS)** attacks is to make a computer or entire network unavailable to its legitimate users.

The goal of denial-of-service (DoS) attacks is to make a computer or entire network unavailable to its legitimate users.

As Figure 3-7 shows, most DoS attacks involve flooding the victim computer with attack packets. The victim computer becomes so busy processing this flood of attack packets that it cannot process legitimate packets. The overloaded host may even fail.

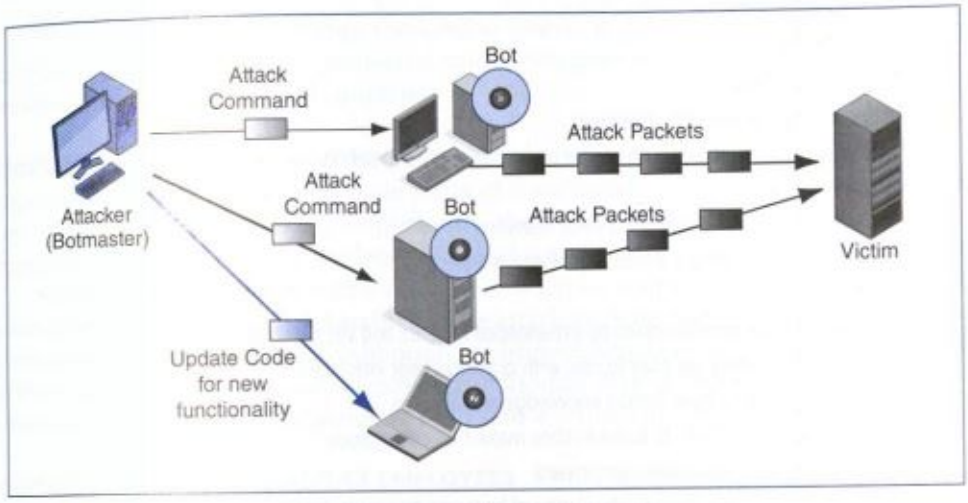

FIGURE 3-7 Distributed Denial-of-Service (DDoS) Attack Using Bots

More specifically, the attack shown in the figure is a **distributed DoS (DDoS)** attack. In this type of DoS attack, the attacker first installs programs called bots on hundreds or thousands of PCs or servers. When the user sends these bots an attack command, they all begin to flood the victim with packets.

Bots are not limited to DDoS attacks. **Bots** are general-purpose exploitation programs that can be remotely controlled after installation. As Figure 3-7 shows, the attacker can send attack commands to the bots and can even upgrade them remotely with new capabilities.

Bots are general-purpose exploitation programs that can be remotely controlled after installation and can even be upgraded remotely with new capabilities.

Bots are extremely dangerous because they can engage in massive attacks that were previously possible only with relatively dumb and inflexible viruses and worms. Through upgrades, bots bring the flexibility of human thought into the attack, making them very dangerous.

Test Your Understanding

11. a) What is the purpose of a denial-of-service attack? b) What are bots? c) What gives bots flexibility? d) How do distributed DoS attacks work?

TYPES OF ATTACKERS

The threat environment consists of types of attacks and types of attackers. As Figure 3-8 shows, there are many different types of attackers facing organizations today.

Traditional Attackers

- Traditional hackers
  - Hackers break into computers
  - Driven by curiosity, a desire for power, and peer reputation
- Malware writers
  - It usually is not a crime to write malware
  - It is almost always a crime to release malware
- Script kiddies
  - Use attack scripts written by experienced hackers and virus writers
  - Scripts usually are easy to use, with graphical user interfaces
  - Script kiddies have limited knowledge and abilities
  - But large numbers of script kiddies make them dangerous
- Disgruntled employees and ex-employees
  - Have extensive access, knowledge of how systems work, and knowledge of how to avoid detection

Criminal Attackers

- *Most* attacks are now made by criminals
- Crime generates funds that criminal attackers need to increase attack sophistication
- Large and complex black markets for attack programs, attacks for hire services, bot rentals and sales, money laundering, and other activities

On the Horizon

- Cyberterror attacks by terrorists
- Cyberwar by nations
- Potential for massive attacks far larger than conventional attacks

FIGURE 3-8 Types of Attackers

Traditional Attackers

When most people think of attackers, they normally have three pictures in their minds: hackers driven by curiosity, virus writers, and disgruntled employees and ex-employees. Indeed, these used to be the three most important types of attackers.

HACKERS Traditionally, some **hackers** have been motivated primarily by curiosity and the sense of power they get from breaking into computers. In many cases, they are also motivated by a desire to increase their reputation among their hacker peers by boasting about their exploits. This typically is the image of hackers presented in Hollywood movies. However, these are not the typical hackers today.

MALWARE WRITERS **Malware writers**, as the name suggests, create malware. Malware writers appear to enjoy the excitement of seeing their programs spread rapidly. These malware writers tend to be blind to the harm that they do to people.

In most countries, including the United States, it generally is not illegal to *write* malware. These activities are protected under freedom of speech. However, *releasing* malware is illegal in nearly all countries.

SCRIPT KIDDIES Experienced hackers and virus writers often developed small programs, called **scripts**, to automate parts of their attacks. Over time, these programs grew more sophisticated. More importantly, they grew easier to use. Many now have graphical user interfaces and the look, feel, and reliability of commercial programs.

Some hackers and virus writers release or sell their scripts. This has led to the emergence of relatively nontechnical **script kiddie** attackers, who use these scripts developed by more experienced attackers. Although traditional attackers disparage script kiddies for their lack of skills, there are far more script kiddies than traditional hackers and virus writers, and script kiddies collectively represent a serious threat to corporations.

DISGRUNTLED EMPLOYEES AND EX-EMPLOYEES Other traditional types of attackers are **disgruntled employees** and **disgruntled ex-employees** who attack their own or their former firms. Employee attackers tend to do extensive damage when they strike because they typically already have access to systems, have broad knowledge of how the systems work, often know how to avoid detection, and tend to be trusted because they are part of the corporate "family."

The most dangerous employees of all are IT staff members and especially IT security staff members. They typically have far more access than other employees, have much better knowledge of corporate systems, and have extensive knowledge of how to avoid detection. In fact, they may even be in charge of identifying attackers. The ancient Roman question, "Quis custodiet ipsos custodes?" means "Who guards the guardians?" It is a serious question in security.

Criminal Attackers

Today, there are still many traditional attackers of the types we have just seen. However, even collectively they do not make up the majority of attackers today. Today, *most* attackers are **career criminals**, who steal credit card numbers to commit credit card fraud, who extort firms, and who steal trade secrets to sell to competitors.

Today, most attackers are career criminals.

Funded by their crimes, many criminals can afford to hire the best hackers and to enhance their own security-breaking skills. Consequently, criminal attacks are not just growing in numbers; they also are growing very rapidly in technical sophistication.

Cyberterrorists and National Governments

On the horizon is the danger of far more massive **cyberterror** attacks by terrorists and even worse **cyberwar** attacks by national governments. These could produce unprecedented damages in the hundreds of billions of dollars.

Cyberwar is not a theory. The United States has acknowledged that it has long had cyberwar capabilities, and it established a consolidated Cyberwar Command in

2009. It is clear that several other countries have these capabilities as well (especially China). Countries could use IT to do espionage to gather intelligence, conduct attacks on opponents' financial and power infrastructures, or destroy enemy command and control facilities during physical attacks.

A 2009 article in the *New York Times*[6] reported that before the 2003 invasion of Iraq, the United States considered an attack that would shut down Iraq's entire financial infrastructure. This attack was not approved, but this was not because it was unfeasible. It was not approved because its impact might have spread beyond Iraq and might even have damaged the U.S. financial system.

Cyberterror is also likely. During physical attacks, terrorists might disable communication systems to thwart first responders and to spread confusion and terror among the population. Cyberterrorists could also conduct purely IT-based attacks. While the United States was afraid of side effects of cyberwar attacks on Iraq, terrorists would have no such qualms.

Test Your Understanding

12. a) Are most attackers today driven by curiosity and a sense of power? b) Is it generally illegal to write malware? c) For what four reasons are employees dangerous? d) What are the most dangerous types of employees? e) What type of attacker are most attackers today? f) What are cyberterror and cyberwar attacks? g) Why are cyberwar and cyberterror serious security concerns?

PLANNING

Security Is a Management Issue

People tend to think of security as a technological issue, but security professionals agree unanimously that security is primarily a management issue. Unless a firm does excellent planning, implementation, and day-to-day execution, the best security technology will be wasted. As Bruce Schneier, a noted security expert, has often said, "Security is a process, not a product."[7] Unless firms have good security processes in place, the most technologically advanced security products will do little good.

Security is primarily a management issue, not a technology issue.

Test Your Understanding

13. Why is security primarily a management issue, not a technology issue?

Planning Principles

Perhaps more than any other aspect of IT, effective security depends on effective planning. Security planning is a complex process that we can discuss only briefly. We will note four key principles that must be used in planning.

[6]Markoff, John, and Shanker, Thom, " '03 Plan Displays Cyberwar Risk," *New York Times*, August 1, 2009. www.msnbc.msn.com/id/3032619/%2328368424.

[7]Schneier, Bruce, *Crypto-Gram Newsletter*, May 15, 2000. www.schneier.com/crypto-gram-0005.html.

Security Is a Management Issue, Not a Technical Issue

- Without good management, technology cannot be effective
- A company must have good security processes

Security Planning Principles

- Risk analysis
  - Risk analysis is the process of balancing threats and protection costs for individual assets
  - Annual cost of protection should not exceed the expected damage
  - If the probable annual damage is $10,000 and the annual cost is $200,000, the protection is not worth the cost
  - Goal is not to eliminate risk but rather to reduce it in an economically rational level
- Comprehensive security
  - An attacker has to find only one weakness
  - A firm needs comprehensive security to close all avenues of attack
  - This requires very good planning
- Defense in depth
  - Every protection breaks down sometimes
  - An attacker should have to break through several lines of defense to succeed
  - Providing this protection is called defense in depth
- Minimal permissions
  - Access control is limiting who can use resources and limiting their permission while using resources
  - Permissions are things they can do with the resource
  - People should be given minimum permissions—the least they need to do their jobs—so that they cannot do unauthorized things

FIGURE 3-9 Security Planning (Study Figure)

RISK ANALYSIS In contrast to military security, which often makes massive investments to stop threats, corporate security planners have to ask whether applying a countermeasure against a particular threat is economically justified. For example, if the probable annual loss due to the threat is $10,000 and the security measures needed to thwart the threat will cost $200,000 per year, the firm should not spend the money. Instead, it should accept the probable loss. **Risk analysis** is the process of balancing threats and protection costs for individual assets.

Risk analysis is the process of balancing threats and protection costs for individual assets.

Figure 3-10 shows a simple risk analysis. Without a countermeasure, the damage per successful attack is expected to be $1,000,000, and the annual probability of a successful attack is 20 percent. Therefore, the annual provable damage is $200,000 without a countermeasure. The net annual probable outlay therefore is $200,000.

Countermeasure A is designed to cut the seriousness of a successful attack in half. So the damage per successful attack is expected to be $500,000 instead of a million

Countermeasure	None	A
Damage per successful attack	$1,000,000	$500,000
Annual probability of a successful attack	20 percent	20 percent
Annual probable damage	$200,000	$100,000
Annual cost of countermeasure	$0	$20,000
Net annual probable outlay	$200,000	$120,000
Annual value of countermeasure	$0	$80,000

FIGURE 3-10 Risk Analysis Example

dollars. The countermeasure will not reduce the probability of a successful attack, so that continues to be 20 percent. With Countermeasure A, then, the annual probable damage is reduced to $100,000. However, the countermeasure is not free. It will cost $20,000 per year. So the net annual probable outlay is $120,000 with the countermeasure.

Countermeasure A, then, will reduce the net annual probable outlay from $200,000 to $120,000. The countermeasure has a value of $80,000 per year. This is positive, so Countermeasure A is justified economically.

Note that *the goal of security is not to eliminate risk*. That would be impossible. Despite efforts throughout human history, we still have theft, and we still have murder. Despite strong security efforts, in turn, there will still be some risk of a compromise. People with little to steal who live in gated communities with bars on their windows, expensive alarm systems, and a permanent armed security guard are not doing rational risk analysis. The goal of security is to reduce risk to a degree that is economically rational.

The goal of security is not to eliminate risk. The goal of security is to reduce risk to a degree that is economically rational.

COMPREHENSIVE SECURITY To be safe from attack, a company must close off *all* vectors of attack. In contrast, an attacker only needs to find one unprotected attack vector to succeed. Although it is difficult to achieve **comprehensive security**, in which all avenues of attack are closed off, it is essential to come as close as possible.

Comprehensive security is closing off all avenues of attack.

DEFENSE IN DEPTH Another critical planning principle is defense in depth. Every protection will break down occasionally. If attackers have to break through only one line of defense, they will succeed during these vulnerable periods. However, if an attacker has to break through two, three, or more lines of defense, the breakdown of a single defense technology will not be enough to allow the attacker to succeed. Having successive lines of defense that all must be breached for an attacker to succeed is called **defense in depth**.

Having several lines of defense that all must be breached for an attacker to succeed is called defense in depth.

MINIMUM PERMISSIONS IN ACCESS CONTROL Security planners constantly worry about access to resources. People who get access to resources can do damage. Not surprisingly, companies work very hard to control access to their resources. **Access control** is limiting who may have access to each resource and limiting his or her permissions when using the resource.

Access control is limiting who may have access to each resource and limiting his or her permissions when using the resource.

One aspect of access control that we will see later is authentication, that is, requiring users requesting access to prove their identity. However, just because you know who someone is does not mean that he or she should have unfettered access to your resources. (There undoubtedly are several people you know who you would not let drive your car.)

Permissions are the actions that a person given access to a resource is allowed to take. For example, although everyone is permitted to view the U.S. Declaration of Independence, no one is allowed to add his or her own signature at the bottom.

Permissions are the actions that a person given access to a resource is allowed to take.

An important principle in assigning permissions is to give each person **minimum permissions**—the least permissions that the user needs to accomplish his or her job. In the case of access to team documents, for example, most team members may be given only read-only access, in which the user can read team documents but not change them. It is far less work to give the user extensive or full permissions so that he or she does not have to be given additional permissions later. However, it is not safe to do so.

Minimum permissions are the least permissions that the user needs to accomplish his or her job.

Test Your Understanding

14. a) List the four major planning principles. b) What is risk analysis? c) Repeat the risk analysis described in this section, this time with Countermeasure B that does not affect damage severity but that reduces the likelihood of an attack by 75 percent. The annual cost of Countermeasure B is $175,000. Show the full table. d) Comment on the statement, "The goal of security is to eliminate risk." e) What is comprehensive security? f) Why is comprehensive security important? g) What is defense in depth? h) Why is defense in depth necessary? i) What is access control? j) What are permissions? k) Why should people get minimum permissions?

Policy-Based Security

POLICIES The heart of security management is the creation and implementation of security policies. Figure 3-11 illustrates how policies should be used. **Policies** are broad statements that specify what should be accomplished. For example, a policy might be, "All information on USB RAM sticks should be encrypted."

Policies are broad statements of what should be accomplished.

POLICY VERSUS IMPLEMENTATION Note that the policy does not specify what encryption technology should be used or other implementation details. Put another way, policies describe *what* (should be done), not *how* (to do it).

Policies describe what (should be done), not how (to do it).

This separation of policy from implementation permits the implementers to implement the policy in the best way possible. Policymakers have the overview knowledge that operational people may not have. For instance, policymakers may know that new laws create serious liability unless USB RAM sticks are encrypted. However, people who do implementation are likely to know more about the specific technologies and the local situation than do policymakers. They have the specific knowledge that policymakers do not, including technical knowledge.

The separation of policy from implementation does not mean that policy is irrelevant to implementation. It is easy to get lost in implementation details. Having a clear

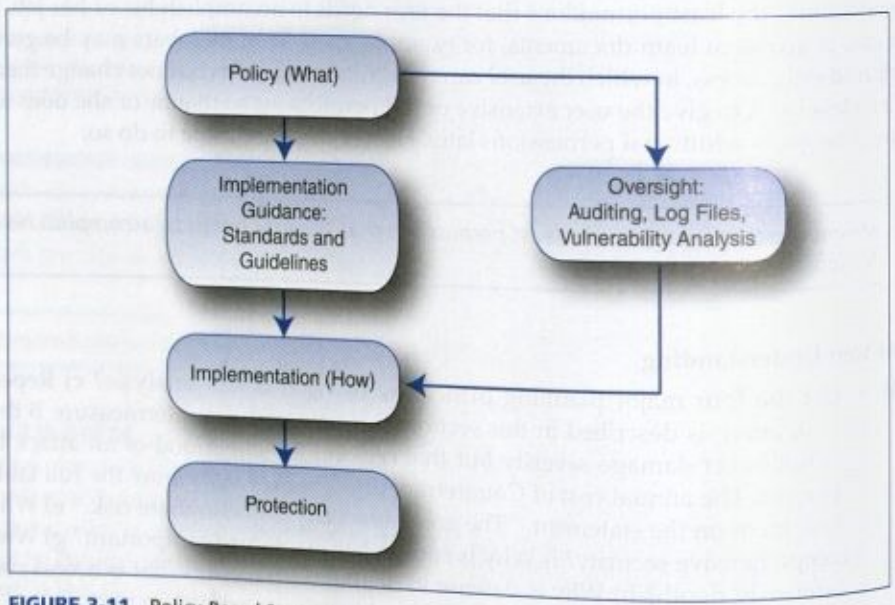

FIGURE 3-11 Policy-Based Security

policy permits everybody involved in implementation to stay synchronized by checking frequently whether what they are doing will lead to the successful implementation of the policy.

IMPLEMENTATION GUIDANCE In many cases, the policymaker will only specify the policy. However, in some cases, the policymaker will also create some implementation guidance. **Implementation guidance** consists of instructions that are more specific than policies but less specific than implementation.

Implementation guidance consists of instructions that are more specific than policies but less specific than implementation.

For example, after establishing a policy that USB RAM sticks must be encrypted, implementation guidance might be added in the form of a directive that the encryption must be strong encryption. This ensures that implementers will not have the latitude to choose weak encryption that can be defeated by an attacker.

There are two general forms of implementation guidance: standards and guidelines. **Standards** are mandatory directives that *must* be followed. Requiring strong encryption is a standard. It is mandatory for implementers to follow the directive.

Standards are mandatory directives that must be followed.

In turn, **guidelines** are directives that *should* be followed but that need not be followed, depending on the context.[8] For example, a directive that security staff members should have three years of security work experience indicates that someone hiring a security staff member must consider that three years of experience is a good indicator of competence. If the person doing the hiring selects someone with only two years of work experience, it would be legitimate to ask the person doing the hiring if he or she felt that less than three years of work experience was acceptable. Following guidelines is optional, but seriously considering guidelines is mandatory.

Guidelines are directives that should be followed but that need not be followed, depending on the context.

OVERSIGHT Figure 3-11 also shows that policymakers cannot merely toss policies and implementation guidance out and ignore how implementation is done. It is essential for management to exercise **oversight**, which is a collection of methods to ensure that policies have been implemented properly.

[8] In the *Pirates of the Caribbean* movies, there was a running joke that the Pirate's Code is "more like a guideline, really."

Oversight is a collection of methods to ensure that policies have been implemented properly.

One form of oversight is an audit. An **audit** samples actions taken within the firm to ensure that policies are being implemented properly. Note that an audit only *samples* actions. It does not look at everything, which would be impossible to do. However, if the sampling is done well, the auditor can issue an opinion on whether a policy is being carried out appropriately based on well-considered data.

An audit samples actions taken within the firm to ensure that policies are being implemented properly.

Another form of oversight is reading **log files**. Whenever users take actions, their actions should be recorded in log files. Reading log files can reveal improper behavior. Of course, if these log files are not read, they are useless. Consequently, it is critical to read log files frequently. Important log files should be read daily or even several times each day.

Reading log files can reveal improper behavior.

Another important oversight mechanism is vulnerability testing. Simply put, **vulnerability testing** is attacking your own systems before attackers do, so that you can identify weaknesses and fix them before they are exploited by attackers.

Vulnerability testing is attacking your own systems before attackers do, so that you can identify weaknesses and fix them before they are exploited by attackers.

Note that the policy drives both implementation and oversight. Implementers who attempt to implement the policy must interpret the policy. Auditors and other oversight professionals must also interpret the policy. If the implementers are lax, the auditors should be able to identify this. However, if oversight practitioners and implementers disagree, this may simply mean that they are interpreting the policy differently. Policymakers may find that one or the other has made a poor choice in interpreting the policy. They may also find that the policy itself is ambiguous or simply wrong. The important thing is to identify problems and then resolve them.

Policies drive both implementation and oversight.

EFFECTIVE PROTECTION Policies certainly do not give protection by themselves. Neither may unexamined implementations. Protection is most likely to be effective when excellent implementation is subject to strong oversight.

Test Your Understanding

15. a) What is a policy? b) Distinguish between policy and implementation. c) Why is it important to separate policies from implementation? d) Why is oversight important? e) Compare the specificity of policies, implementation guidance, and implementation. f) Distinguish between standards and guidelines. g) Must guidelines be considered? h) List the three types of oversight listed in the text. i) What is vulnerability testing, and why is it done? j) Why is it important for policy to drive both implementation and oversight?

AUTHENTICATION

The most complex element of access control is authentication. Figure 3-12 illustrates the main terminology and concepts in authentication. The user trying to prove his or her identity is the **supplicant**. The party requiring the supplicant to prove his or her identity is the **verifier**. The supplicant tries to prove his or her identity by providing **credentials** (proofs of identity) to the verifier.

The type of authentication tool that is used with each resource must be appropriate for the risks to that particular resource. Sensitive personnel information should be protected by very strong authentication methods. However, strong authentication is expensive and inconvenient. For relatively nonsensitive data, weaker but less expensive authentication methods may be sufficient. Strength of authentication, like everything else in security, is a matter of risk management.

Test Your Understanding

16. a) What is authentication? b) Distinguish between the supplicant and the verifier. c) What are credentials? d) Why must authentication be appropriate for risks to an asset?

Reusable Passwords

The most common authentication credential is the **reusable password**, which is a string of characters that a user types to gain access to the resources associated with a certain **username** (account) on a computer. These are called **reusable** passwords because the user will type the password each time he or she needs access to the resource. The reusable password is the weakest form of authentication, and it is appropriate only for the least sensitive assets.

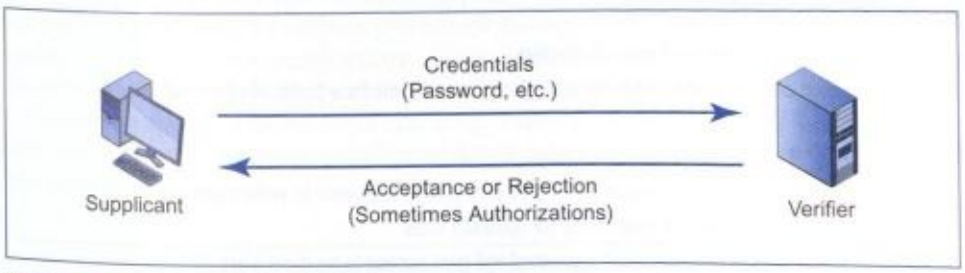

FIGURE 3-12 Authentication

The reusable password is the weakest form of authentication, and it is appropriate only for the least sensitive assets.

EASE OF USE AND LOW COST The popularity of password authentication is hardly surprising. For users, passwords are familiar and relatively easy to use. For corporate IT departments, passwords add no additional cost because operating systems and many applications have built-in password authentication.

WORD/NAME PASSWORDS AND DICTIONARY ATTACKS The main problem with passwords is that most users pick very weak passwords.

The main problem with passwords is that most users pick very weak passwords.

For example, they often pick ordinary **dictionary words** or the names of family members, pets, sports teams, or celebrities. Dictionary-word and name passwords

Reusable Passwords

Passwords are strings of characters
They are typed to authenticate the use of a username (account) on a computer
They are used repeatedly and so are called reusable passwords

Benefits

Ease of use for users (familiar)
Inexpensive because they are built into operating systems

Often Weak (Easy to Crack)

Word and name passwords are common
They can be cracked quickly with dictionary attacks

Passwords Should Be Complex

Should mix case, digits, and other keyboard characters ($, #, etc.)
Complex passwords can be cracked only with brute force attacks (trying all possibilities)

Passwords Also Should Be Long

Should have a minimum of eight characters
Each added character increases the brute force search time by a factor of about 70

Other Concerns

If people are forced to use long and complex passwords, they tend to write them down
People should use different passwords for different sites
Otherwise, a compromised password will give access to multiple sites

FIGURE 3-13 Password Authentication (Study Figure)

often can be cracked (guessed) in a few seconds if the attacker can get a copy of the password file (which contains an encrypted list of account names and passwords). The attacker uses a **dictionary attack**, trying all words or names in a standard or customized dictionary. There are only a few thousand dictionary words and names in any language, so dictionary attacks can crack dictionary-word and name passwords almost instantly.

Dictionary attacks also have **hybrid modes**, in which they look for simple variations on words, such as a word with the first letter capitalized, followed by a single digit (e.g., Dog1). Hybrid word or name passwords are cracked almost as quickly as passwords made of simple words and names.

Names, words, and simple variants of words and names that can be cracked by hybrid mode dictionary attacks are never adequately strong, regardless of how long they are. They can always be cracked too quickly for safety.

Names, words, and simple variants of words and names that can be cracked by hybrid mode dictionary attacks are never adequately strong, regardless of how long they are.

COMPLEX PASSWORDS AND BRUTE FORCE ATTACKS Dictionary and hybrid dictionary attacks fail if passwords are more complex than dictionary words, names, and simple variations. Good complex passwords have all of the following:

- Lowercase letters.
- Uppercase letters, not simply at the start of the password.
- The digits from 0 to 9, not simply at the end of the password.
- Other keyboard symbols, such as & and #, which serve as swear words in cartoons—not simply at the end of the password.

Complex passwords can be cracked only by **brute force attacks** that try all possible combinations of characters. First, all combinations of a single character are tried, all combinations of two characters, all combinations of three characters, and so forth. Brute force attacks take far longer than dictionary attacks.

Unfortunately, complex passwords are difficult for users to remember, so they tend to write them on a sheet of paper that they keep next to their computers. This makes passwords easy to steal so that there is no need to crack them by dictionary or brute force attacks.

COMPLEX PASSWORD LENGTH Increasing **password length** (the number of characters in the password) makes a complex password stronger. If the password has a combination of uppercase and lowercase letters, digits, and other keyboard characters, then each additional character increases cracking time by a factor of about 70.

Given the speed of brute force cracking today, passwords should be complex and at least eight characters long to be considered adequate. Even longer passwords are highly desirable.

Only passwords that are complex and at least eight characters long should be considered to be adequately strong.

Test Your Understanding

17. a) Distinguish between usernames and reusable passwords. b) Why are passwords widely used? c) What types of passwords are susceptible to dictionary attacks? d) What types of passwords are susceptible to dictionary attacks in hybrid mode? e) Can a password that can be broken by a dictionary attack or a dictionary attack in hybrid mode be adequately strong if it is very long? f) What is a brute force attack? g) What types of passwords can be broken only by brute force attacks? h) Why is password length important? i) How long should passwords be?
18. Critique each of the following passwords. First, describe the type of attack that would be used to crack it, justifying your answer. Second, say whether or not it is of adequate strength, justifying your answer. a) velociraptor; b) Viper1; c) NeVeR; d) R7%t&.

Other Forms of Authentication

Companies are beginning to look for stronger types of authentication for most of their resources. This will allow them to replace most or all of their reusable password access systems. We have space to mention only the few types of authentication shown in Figure 3-14.

ACCESS CARDS To get into your hotel room, you may have to swipe your **access card** through a card reader before being allowed through. For door and computer access, many companies also use these handy access cards, including **proximity access cards** that use radio signals and can be read with a simple tap against a reader. Companies need to control the distribution of access cards, and they need to rapidly disable any access card that has been lost or stolen.

BIOMETRICS Access cards are easy to use, but if you lose your access card, you cannot get entry. In hotels, of course, you simply walk down to the front desk. They disable the code on your room card reader and give you a new card that will open your room. In corporate environments, the process takes a good deal longer.

In biometrics, in contrast, access control is granted based on something you always have with you—your body. **Biometrics** is the use of body measurements to authenticate you.

Biometrics is the use of body measurements to authenticate you.

There are several types of biometrics that differ in cost, precision, and susceptibility to deception by someone wishing to impersonate a legitimate user.

- At the low end on price, precision, and the ability to reject deception is **fingerprint scanning**, which looks at the loops, whorls, and ridges in your fingerprint. Although fingerprint scanning is not the strongest form of authentication, its low price makes it ideal for low-risk applications. Even for protecting laptop computers and smart phones, fingerprint scanning may be preferred to reusable passwords, given the tendency of people to pick poor passwords and forget them.

Perspective

- Goal is to replace reusable passwords

Access Cards

- Permit door access
- Proximity access cards do not require scanning
- Need to control distribution and disable lost or stolen access cards

Biometrics

- Biometrics uses body measurements to authenticate you
- Vary in cost, precision, and susceptibility to deception
- Fingerprint scanning
  - Inexpensive but poor precision, deceivable
  - Sufficient for low-risk uses
  - On a notebook, may be better than requiring a reusable password
- Iris scanning
  - Based on patters in the colored part of your eye
  - Expensive but precise and difficult to deceive
- Facial scanning
  - Based on facial features
  - Controversial because can be done surreptitiously—without the supplicant's knowledge

Digital Certificate Authentication

- Components
  - Everyone has a private key that only he or she knows
  - Everyone also has a public key that is not secret
  - Public keys are available in unalterable digital certificates
  - Digital certificates are provided by trusted certificate authorities
- Operation
  - Supplicant does a calculation with his or her private key
  - Verifier checks this calculation with a public key in a digital certificate
  - Verifier uses the digital certificate of the true party—the person the supplicant claims to be
  - If the calculation check works, the supplicant must have the true party's private key, which only the true party should know. The supplicant must be the true party.

Two-Factor Authentication

- Supplicant needs two forms of credentials
- Example: debit card and pin
- Strengthens authentication
- Fails if attacker controls user's computer or intercepts authentication communication

FIGURE 3-14 Other Forms of Authentication

- At the high end of the scale on price, precision, and the ability to reject deception is **iris scanning**,[9] which looks at the pattern in the colored part of your eye. Although extremely precise, iris scanners are too expensive to use for computer access. They are normally used for access to sensitive rooms.
- One controversial form of biometrics is **facial scanning**, in which each individual is identified by his or her facial features. This is controversial because facial scanning can be done **surreptitiously**—without the knowledge of the person being scanned. This raises privacy issues.

DIGITAL CERTIFICATE AUTHENTICATION The strongest form of authentication is digital certificate authentication. Figure 3-15 illustrates this form of authentication.

- In this form of authentication, each person has a secret **private key** that only he or she knows.
- Each person also has a **public key**, which anyone can know.
- A trusted organization called a **certificate authority** distributes the public key of a person in a document called a **digital certificate**. A digital certificate cannot be changed without this change being obvious.

First, the supplicant claims to be someone we will call the **true party**. To prove this claim, the supplicant does a calculation[10] with his or her private key and sends this calculation to the verifier.

Second, the verifier gets the true party's digital certificate, which contains the true party's public key. The verifier tests the calculation with the public key of the true party—the person the supplicant claims to be.[11] If the test works, then the supplicant must know the true party's private key, which only the true party should know. The supplicant must be the true party.

Note that the verifier uses the public key of the true party—not the supplicant's public key. If the verifier used the supplicant's public key, the test would always succeed—even if the supplicant is an impostor.

Note that the verifier uses the public key of the true party—not the supplicant's public key.

TWO-FACTOR AUTHENTICATION Debit cards are potentially dangerous because if someone finds a lost debit card, the finder might be able to use it to make purchases. So possession of the debit card is not enough to use it. To use a debit card, the user must type a **personal identification number (PIN)**, which usually is four or six digits long. Requiring two credentials for authentication is called **two-factor authentication**. Two-factor authentication increases the strength of authentication.

---

[9]In science fiction movies, eye scanners are depicted as shining light into the supplicant's eye. This does not really happen. Iris scanners merely require the supplicant to look into a camera. In addition, science fiction movies use the term *retinal scanning*. The retina is the back part of the eye and has distinctive vein patterns. Retinal scanning is not used frequently because the supplicant must press his or her face against the scanner.

[10]To be more specific, the verifier sends the supplicant a challenge message that is a random stream of bits. The calculation that the supplicant does is to encrypt the challenge message with the supplicant's private key. The result is the response message, which the supplicant sends to the verifier.

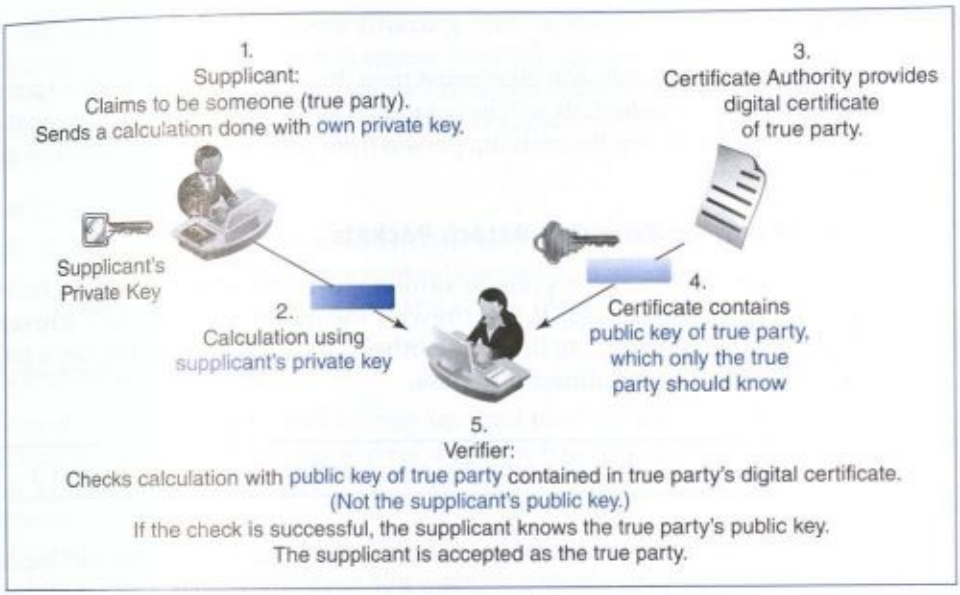

FIGURE 3-15 Digital Certificate Authentication

However, if a user's computer is compromised, the attacker typically controls both aspects of communication. Two-factor authentication may also break down if an eavesdropper can intercept authentication communication between the two parties.

Two-factor authentication requires two forms of authentication.

Test Your Understanding

19. a) What security problem do access cards have? b) What is biometrics? c) By what three criteria should biometric methods be judged? d) Why may fingerprint scanning be used to authenticate access to a laptop? e) Why is iris scanning desirable? f) Why is face recognition controversial?
20. a) In digital certificate authentication, what does the supplicant do? b) What does the verifier do? c) Does the verifier use the true party's public key or the supplicant's public key? d) How does the verifier get the public key? e) From what type of organization does the verifier get the digital certificate?
21. a) Why is two-factor authentication desirable? b) Will two-factor authentication still be strong if the attacker controls the supplicant's computer? c) Will two-factor authentication still be strong if the attacker can intercept all authentication communication?

[11] To be more specific, the verifier decrypts the response message with the true party's public key. If someone encrypts something with his or her private key, this can be decrypted with the true party's public key. If the supplicant is the true party, the true party's public key will decrypt the response message back to the challenge message. The supplicant will be authenticated as the true party. If the supplicant is an impostor, when the verifier decrypts the response message, the result will not be the challenge message that the verifier originally sent to the supplicant. The supplicant will then be rejected as an impostor.

FIREWALLS

In hostile military environments, travelers must pass through one or more checkpoints. At each checkpoint, their credentials will be examined. If the guard finds the credentials insufficient, the guard will stop the arriving person from proceeding and note the violation in a checkpoint log.

Dropping and Logging Provable Attack Packets

Figure 3-16 shows that firewalls operate in similar ways. Whenever a packet arrives, the **firewall** examines the packet. If the firewall identifies a packet as a **provable attack packet**, the firewall discards it. On the other hand, if the packet is not a provable attack packet, the firewall allows it to pass.

If a firewall identifies a packet as a provable attack packet, the firewall discards it.

The firewall copies information about the discarded packet into a **firewall log file**. Firewall managers should read their firewall log files every day to understand the types of attacks coming against the resources that the firewall is protecting.

Note that firewalls pass *all* packets that are not provable attack packets. Some attack packets will not be provable attack packets. Consequently, some attack packets inevitably get through the firewall to reach internal hosts. It is important to harden all internal hosts against attacks by adding firewalls, adding antivirus programs, installing all patches promptly, and taking other precautions. This chapter focuses on network security, rather than IT host security, so we will not consider host hardening.

Ingress and Egress Filtering

When most people think of firewalls, they think of filtering packets arriving at a network *from the outside*. Figure 3-16 illustrates this **ingress filtering**.

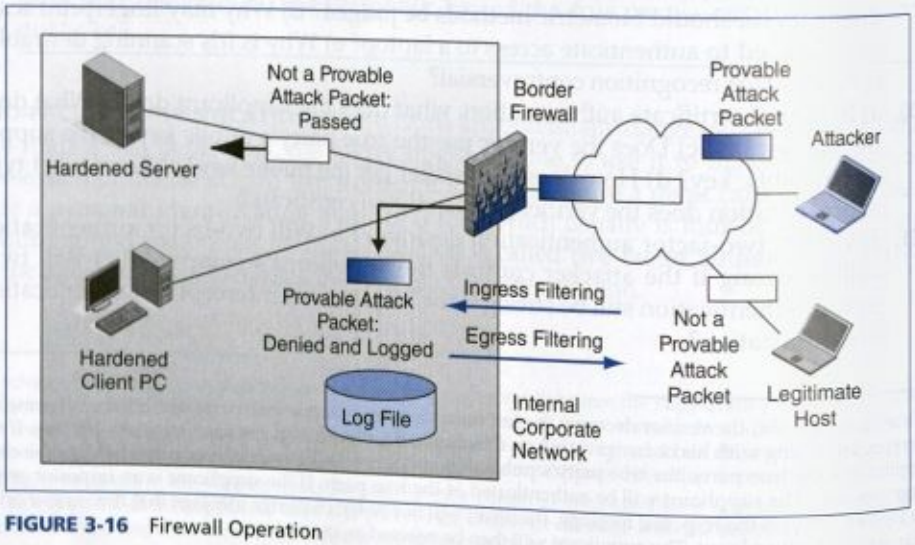

FIGURE 3-16 Firewall Operation

Most firms also do **egress filtering**, that is, they filter packets going from the network *to the outside*. By doing egress filtering, the corporation is acting as a good citizen, ensuring that its computers are not used in attacks against outside firms. Egress filtering also attempts to prevent sensitive corporate information from being sent outside the firm.

Test Your Understanding

22. a) What does a firewall do when a packet arrives? b) Does a firewall drop a packet if it probably is an attack packet? c) Why is it important to read firewall logs daily? d) Distinguish between ingress and egress filtering.

Static Packet Filtering

We have used the term *firewall filtering* up until now without explaining it. We did this because different firewalls use several different **filtering methods**. In this section, we will look at three: static packet filtering, stateful firewall filtering, and deep packet inspection.

Static packet filtering was the first type of filtering used by firewalls. As Figure 3-17 illustrates, static packet filter firewalls examine packets one at a time, in isolation. In addition, they look at only the internet and transport layer fields in the packet.

Despite these limits, static packet filtering can handle many types of packet attacks. For example, in Chapter 2, we saw that TCP segments that have their SYN bit set indicate that the sender wishes to open a connection. In Chapter 8, we will see that the sender sets the FIN bit in a TCP segment to indicate that it wishes to close a connection. Attackers soon discovered that if they set both the SYN bit and the FIN bit in the same segment, they could confuse the receiving transport layer process. Often, the receiving computer would crash. In Chapter 8, we will see that there are many fields in IP, TCP, and UDP headers. Static packet inspection firewalls look for any suspicious patterns in the contents of the field. They can also check for many other things; for instance, if a packet arrives from outside a site that has an address in the range restricted for IP addresses within the site, the static packet filter can drop the packet.

Unfortunately, examining single packets in isolation means that static packet filtering firewalls cannot detect many types of attack. For example, if an incoming

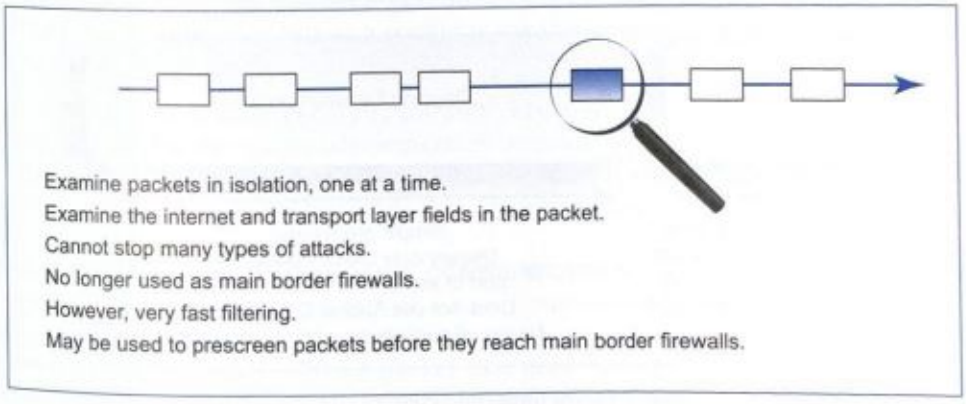

FIGURE 3-17 Static Packet Filtering

packet has its ACK bit set, this means that it is sending an acknowledgement for a TCP segment from the inside to the outside. However, the static packet filter cannot tell if this is a legitimate acknowledgment because it only examines the packet in isolation. It cannot tell if there was a previous outgoing packet to which this incoming packet is a legitimate acknowledgment.

As a consequence of not being able to stop many types of attacks, static packet filter firewalls are no longer used as main border firewalls in firms. However, static packet filtering is inexpensive to do, so some used static packet firewalls to screen out simple attacks before packets reach the main border firewall.

Test Your Understanding

23. a) What is the limitation of static packet filtering? b) Why is static packet filtering still done despite its weakness, and how is it used?

Stateful Packet Inspection (SPI) Firewalls

The most widely used *firewall filtering method* today is stateful firewall inspection, which treats different types of packets differently, spending the most resources on the most risky packets, which are relatively few, and spending less time on less risky packets.

STATES AND FILTERING INTENSITY When you talk with someone on the telephone, there are two basic stages to your conversation.

- At the beginning of a call, you need to identify the other party and decide whether you are both willing to have a conversation.
- Afterward, if you do decide to talk, you usually don't have to constantly worry about whether the conversation should go on with this person.

The key point here is that you do different things in different stages of a conversation. In the first stage, you have to pay careful attention to identifying the caller and making a decision about whether it is wise to talk. After that, you simply talk and normally do not have to spend much time thinking about whether to talk to the person.

Most firewalls today use **stateful packet inspection (SPI)** filtering, which uses the insight that there are also stages in network conversations and that not all stages require the same amount of firewall attention. At the simplest level, Figure 3-18 shows that

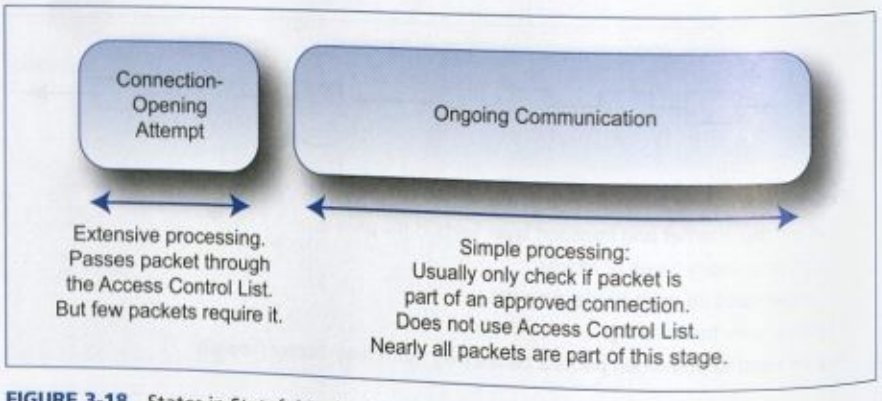

FIGURE 3-18 States in Stateful Packet Inspection (SPI)

Rule	Destination IP Address or Range	Service	Action
1	ALL	25	Allow connection
2	10.47.122.79	80	Allow connection
3	ALL	ALL	Do not allow connection
Note: ACLs are applied only to packets that attempt to open a connection.			

FIGURE 3-19 Stateful Inspection Firewall Access Control List (ACL) for Connection-Opening Attempts

there are two stages, which SPI firewalls call **states**: opening a connection (conversation) and ongoing communication.

SPI FILTERING IN THE CONNECTION-OPENING STATE SPI firewalls focus heavily on the opening state. They have complex rules to tell them whether or not to allow the conversation (connection). If they decide to allow a connection, however, they give minimal attention to packets in the ongoing communication state. This makes sense because the decision to allow a connection is the most complex and dangerous stage in the connection.

For example, suppose that a packet arriving at a firewall contains a TCP SYN segment. This clearly is a connection-opening request to the destination host. So the firewall compares the features of the packet to the rules in its **access control list (ACL)**. Figure 3-19 shows a very simplified access control list. This ACL has only three rules.

- Rule 1 allows connections to all hosts (all IP addresses) on Port 25. We saw in Chapter 2 that Port 25 is the well-known port number for SMTP. This rule permits connections to all internal mail servers.
- Rule 2 permits connections to a single internal host, 10.47.122.79, on Port 80. This rule allows access to a single internal webserver—the webserver at IP address 10.47.122.79. This is safer than Rule 1 because Rule 1 opens the firewall to *every* internal mail server, while Rule 2 opens the firewall only to connections to *a single* server.
- The last rule is the default rule for incoming packets that try to open a connection. (The default is what you get if you do not explicitly specify something else.) This last rule ensures that unless a packet is explicitly allowed by an earlier rule, it is dropped and logged.

Although ACL rules generally are not very complex, there tend to be many of them in real ACLs. Running each connection-opening attempt through the access control list can be fairly time consuming. Fortunately, only a very small percentage of all packets arriving at a firewall are connection-opening attempts.

HANDLING PACKETS DURING ONGOING COMMUNICATION If a packet does not attempt to open a connection or is not part of a connection-opening attempt, then either the packet must be part of the ongoing communication state of an approved connection or the packet is spurious. When a packet that does not attempt to open a connection arrives, then the stateful firewall does the following (see Figure 3-20).

If the packet does not attempt to open a connection,

Then if the packet is part of an established connection

It is passed without further inspection

(However, these packets can be given additional filtering if desired)

Otherwise, it is dropped and logged

Nearly all packets are not part of connection-opening attempts

This simplicity makes the cost of processing most packets minimal

FIGURE 3-20 Stateful Inspection for Packets That Do Not Attempt to Open a Connection

- If the packet is part of an established connection, it is passed without further inspection. (However, these packets can be further filtered if desired.)
- If the packet is not part of an established connection, then it must be spurious. It is dropped and logged.

These rules for ongoing communication are very simple to implement. Consequently, most packets are handled with very little processing power. This makes stateful firewalls very inexpensive.

PERSPECTIVE Although the simple operation of stateful inspection makes it inexpensive, stateful filtering provides a great deal of protection against attacks coming from the outside. This combination of low cost and strong security is responsible for the dominance of stateful inspection today.

Test Your Understanding

24. a) Why are states important? b) Why are ACLs needed for stateful firewalls? c) When a packet that is part of an ongoing connection arrives at a stateful inspection firewall, what does the firewall usually do? d) When a packet that is not part of an ongoing connection and that does not attempt to open a connection arrives at a stateful inspection firewall, what does the firewall do? e) Why are stateful firewalls attractive? f) What type of firewalls do most corporations use for their main border firewalls?

25. a) How will an SPI firewall handle a packet containing a TCP segment which is an acknowledgement? b) How will an SPI firewall handle a packet containing a TCP SYN segment? c) How will an SPI firewall handle a packet containing a TCP FIN segment? d) How will the access control list (ACL) in Figure 3-19 handle a packet that attempt to open a connection to an FTP server? Explain.

Deep Inspection Firewalls

There is a relatively new type of firewall inspection method that we will call **deep inspection**. (Different vendors use different terminology.) Figure 3-21 shows that deep inspection firewalls do two things: they examine patterns in *streams* of packets and they look for attack patterns at all layers in a packet, including the application message.

Examine Streams of Messages

Stateful inspection firewalls know packet context (connection-opening or not) but still examine only individual packets.

Deep inspection firewalls look at streams of packets for patterns

For example, reconstruct application messages from TCP segments in different packets

Read All Packet Layers, Including Application Messages

Stateful packet inspection packets do not read application messages in detail

Deep inspection firewalls examine application messages in detail

This allows them to tell when a message to Port 80 is not an HTTP message

These may use Port 80 for illegal file sharing and other attacks

Some deep inspection packets are application-aware, allowing administrators to set up filtering rules for many specific applications

Intrusion Detection Systems (IDSs)

Deep inspection firewalls began as intrusion detection systems (IDSs)

Found suspicious patterns in traffic and notified the firewall administrators

Evolved to the point where there was enough confidence to let them actively stop traffic

Requires Extensive Processing Power

Far more than SPI

Made possible by application-specific integrated circuits (ASICs)

ASICs handle specific deep firewall inspection tasks in specialized hardware, which is very fast

This is finally making deep inspection feasible

FIGURE 3-21 Deep Inspection Firewalls

Although stateful packet inspection firewalls understand the context of each packet (whether or not the packet is a connection-opening attempt), they still only examine individual packets in detail. There are some attacks that are not discoverable unless the firewall can examine a stream of packets to pick out tell-tale patterns indicating that this is an attack stream. Most obviously, a large application message will be fragmented and transmitted in multiple packets. Unless the packets containing the application message fragments are combined, there is no way to discover malicious behavior in the application message. Deep inspection firewalls collect series of packets and analyze their contents as a group.

In addition, SPI firewalls do not read application messages.[12] Instead, they use server port number to be able to create filtering rules for certain applications. For example, a firm may allow all incoming packets with TCP segments with the destination port number 80. This allows all external connections to internal webservers. However, attackers know this, so they often use Port 80 for other purposes, such as illegal file

[12]This is a slight lie. They do look for certain things in application messages. For example, in voice over IP, a connection is set up on a particular port number. Once the connection is set up, the voice over IP gateway tells the VoIP client to switch to another port number. Stateful packet inspection firewalls look for such things and allow communicate over the new port number for the duration of the call. Deep inspection firewalls go far beyond such things.

sharing or communicating with bots. Deep inspection firewalls do not have this limitation. They examine the actual application messages in packets to identify the application. For many deep inspection firewalls, the firewall administrators can create filtering rules for individual applications, such as BitTorrent, which is often used for illegal file sharing. Firewalls that do this are called **application-aware firewalls**.

Even when deep inspection firewalls cannot tell beyond a reasonable doubt that a certain conversation is forbidden, they can often classify it as suspicious and warn firewall administrators about the traffic. Deep inspection firewalls that do this are called **intrusion detection systems (IDSs)**. In fact, deep inspection devices were first used exclusively as intrusion detection systems. It was only later that their capabilities were expanded so that they had enough sophistication to be allowed to stop conversations.

If deep inspection sounds expensive, it is. Deep inspection firewalls require far more processing power than stateful packet inspection. They have been feasible to use only recently, thanks to the evolution of application-specific integrated circuits (ASICs). In the past, firewalls had to use general-purpose microprocessors and write programs to do inspection. ASICs are designed for specific purposes, in this case for many deep inspection tasks. ASICs handle these tasks in hardware, which is much faster than handling them in software. This speed advantage has allowed deep inspection firewalls to be useful in many situations today.

Test Your Understanding

26. a) What two things do deep inspection firewalls do that SPI firewalls do not? b) Why is the first useful? c) Why is the second useful? d) What is an application-aware firewall, and how it useful? e) For what type of devices was deep inspection first used? f) What is the main problem with deep inspection firewalls? g) What technology is overcoming this problem, and how is it doing so?

PROTECTING DIALOGUES CRYPTOGRAPHY

We now continue our discussion of the protection phase in the plan–protect–respond cycle by looking at cryptographic protections for dialogues involving the exchange of many messages. Cryptography is the use of mathematics to protect dialogues.

Cryptography is the use of mathematics to protect dialogues.

Symmetric Key Encryption for Confidentiality

ENCRYPTION FOR CONFIDENTIALITY When most people think of cryptographic protection, they think of encryption for confidentiality, which Figure 3-22 illustrates. **Confidentiality** means that an eavesdropper intercepting the message will not be able to read it. The sender uses an encryption method, called a **cipher**, to create a message that an eavesdropper cannot read. However, the receiver **decrypts** the message in order to read it.

SYMMETRIC KEY ENCRYPTION Most encryption for confidentiality uses **symmetric key encryption** ciphers, in which the two sides use the same key to encrypt messages to each other and to decrypt incoming messages. As Figure 3-22 shows, symmetric key encryption ciphers use only a single key for encryption by Party A and decryption by

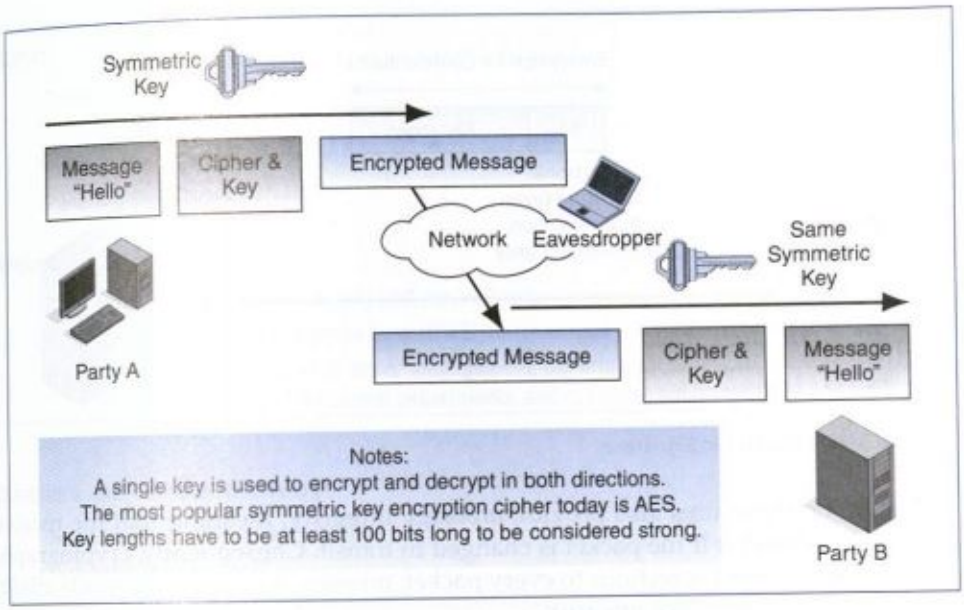

FIGURE 3-22 Symmetric Key Encryption for Confidentiality

Party B. When Party B sends to Party A, in turn, Party B uses the single key to encrypt, while Party A uses the single key to decrypt. The dominant symmetric key encryption cipher today is the **Advanced Encryption Cipher (AES)**.

KEY LENGTH Earlier, we looked at brute force password guessing. Symmetric and keys also can be guessed by the attacker's trying all possible keys. This is called **exhaustive search**. The way to defeat exhaustive key searches is to use long keys, which are merely binary strings. For symmetric key ciphers, symmetric key lengths of 100 bits or greater are considered to be strong. AES supports multiple strong key lengths up to 256 bits.

Keys are long strings of bits.

Test Your Understanding

27. a) What is a cipher? b) What protection does confidentiality provide? c) In two-way dialogues, how many keys are used in symmetric key encryption? d) What is the minimum size for symmetric keys to be considered strong?

Electronic Signatures

AUTHENTICATION AND MESSAGE INTEGRITY In addition to encrypting each packet for confidentiality, cryptographic systems normally add **electronic signatures** to each packet. This is illustrated in Figure 3-23. Electronic signatures are small bit strings that provide message-by-message authentication, much as people use signatures to authenticate individual written letters. An electronic signature allows the receiver to detect a message added to the dialogue by an impostor.

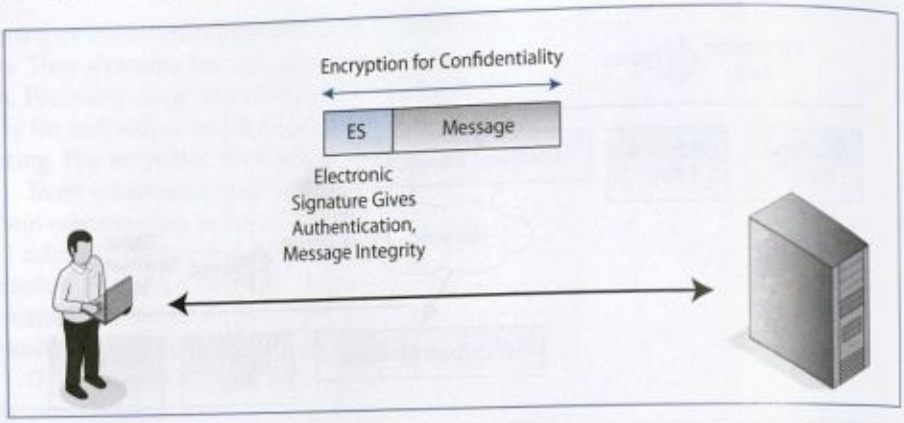

FIGURE 3-23 Electronic Signature

Electronic signatures also provide **message integrity**, meaning that the receiver will be able to detect it if the packet is changed in transit. Consequently, cryptographic systems provide three protections to every packet: message-by-message confidentiality, authentication, and message integrity.[13]

Test Your Understanding

28. What two protections do electronic signatures provide?

RESPONDING

The last stage in the plan–protect–respond cycle is **responding**. Inevitably, some attacks will succeed and will require corrective action. **Response** is reacting to a security incident according to plan. The "according to plan" part is crucial. The amount of damage done in these compromises depends heavily on how quickly and how well the organization responds. Without strong and well-rehearsed response plans, response will take far too long and is likely to be less effective than it should be. For example, the case study at the end of this chapter shows that Walmart has a well-rehearsed response plan for natural disasters. In Hurricane Katrina, while the Federal Emergency Management Agency stumbled badly, Walmart responded quickly and effectively.

Response is reacting to a security incident according to plan.

Stages

There are four general stages in responding to an attack.

DETECTING THE ATTACK The first stage is detecting the attack. Detection can be done by technology or simply by users reporting apparent problems. Obviously, until an attack is detected, the attacker will be able to continue doing damage. Companies need to develop strong procedures for identifying attacks quickly.

[13]There are two common types of electronic signatures—digital signatures and key-hashed message authentication codes. Many writers focus on digital signatures, but digital signatures are not very common because they are expensive to implement.

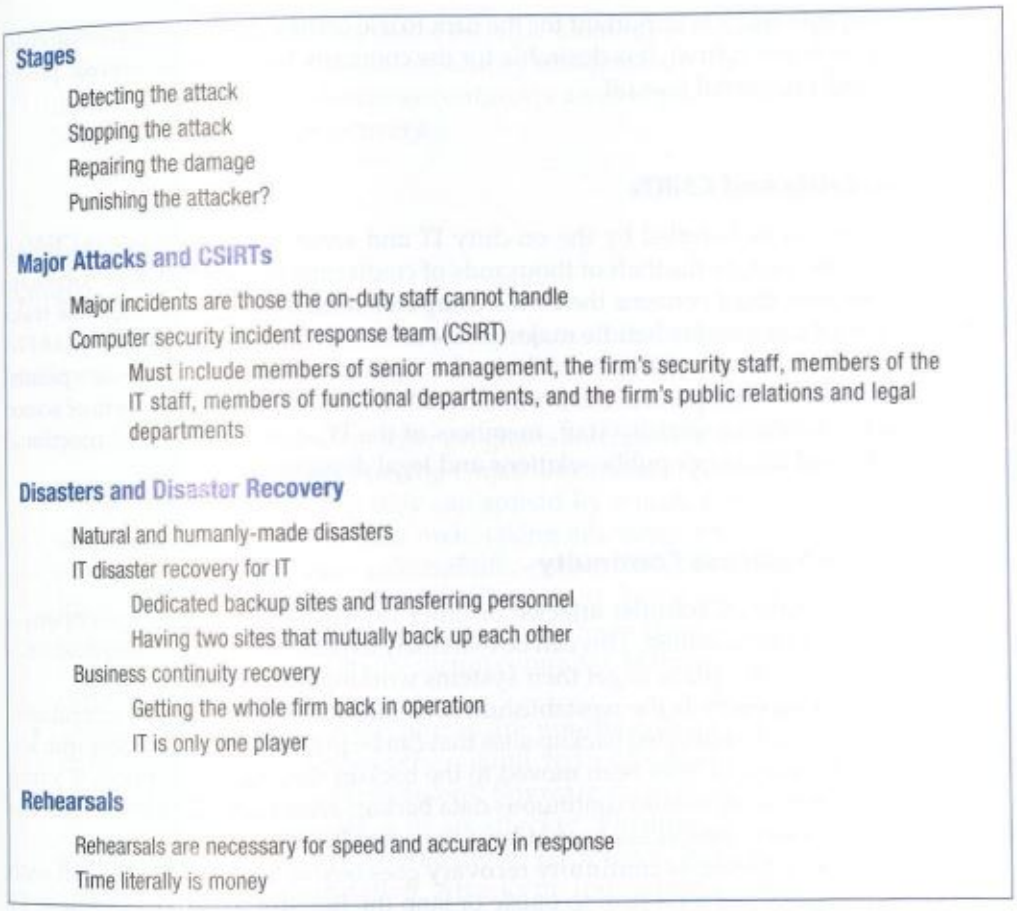

FIGURE 3-24 Incident Response (Study Figure)

STOPPING THE ATTACK The second stage is stopping the attack. The longer an attack has to get into the system, the more damage the hacker can do. Reconfiguring corporate firewall ACLs may be able to end the attack. In other cases, attack-specific actions will have to be taken.

REPAIRING THE DAMAGE The third stage is repairing the damage. In some cases, this is as simple as running a cleanup program or restoring files from backup tapes. In other cases, it may involve the reformatting of hard disk drives and the complete reinstallation of software and data.

PUNISHING THE ATTACKER? The fourth general stage is punishing the attacker, but this stage often is skipped. Punishing the attacker is relatively easy if the attacker is an employee. In general, however, attackers are extremely difficult to track down. Even if they are found, prosecution may be difficult or impossible.

If legal prosecution is to be pursued, it is critical for the company to use proper forensic procedures to capture and retain data. **Forensic** procedures are ways to capture and safeguard data in ways that fit rules of evidence in court proceedings. These rules

are very complex, and it is important for the firm to use certified forensics professionals. Even if an employee is fired, it is desirable for the company to use good forensic procedures to avoid a potential lawsuit.

Major Incidents and CSIRTs

Minor attacks can be handled by the on-duty IT and security staff. However, during **major incidents**, such as the theft of thousands of credit card numbers from a corporate host, the company must convene the firm's **computer security incident response team (CSIRT)**, which is trained to handle major incidents.

The key to creating CSIRTs is to have the right mix of talents and viewpoints. Major attacks affect large parts of the firm, so the CSIRT must include members of senior management, the firm's security staff, members of the IT staff, members of functional departments, and the firm's public relations and legal departments.

Disasters and Business Continuity

When natural disasters, terrorist attacks, or other catastrophes occur, the company's basic operations may be halted. This can be extremely expensive. Companies must have active disaster recovery plans to get their systems working quickly.

IT disaster recovery is the reestablishment of information technology operations. Many large firms have dedicated backup sites that can be put into operation very quickly, after data and employees have been moved to the backup site. Another option, if a firm has multiple server sites, is to do continuous data backup across sites. If one site fails, the other site can take over immediately or at least very rapidly.

More broadly, **business continuity recovery** goes beyond IT disasters to deal with events that affect enough of a firm to pause or stop the functioning of the business. IT security is only one player in business continuity recovery teams.

Rehearsals

"Practice makes perfect" is time-honored advice. It certainly is true for major attacks that must be handled by CSIRTs, and it is doubly true for disaster recovery. It is important for the company to establish CSIRT and disaster teams ahead of time and to have them rehearse how they will handle major attacks and disasters. Although practice does not really make perfect, it certainly improves response speed and quality. During the first two or three rehearsals, team members will work together awkwardly, and there will be many mistakes. Rehearsals will also reveal flaws in the company's major incident and disaster response plans. It is important to work through these problems before the firm is in a real crisis.

Test Your Understanding

29. a) What is the definition of response? b) What are the two benefits of a well-rehearsed response plan? c) What are the four response phases when attacks

occur? d) What is the purpose of forensic tools? e) Why are CSIRTs necessary? f) Should the CSIRT be limited to security staff personnel? g) Distinguish between disaster recovery and business continuity recovery. h) Explain how firms use backup sites in disaster recovery.

CONCLUSION

Synopsis

ATTACKS Companies today suffer compromises from many different types of attacks.

- Viruses attach themselves to other programs and need human actions to propagate, most commonly by users opening e-mail attachments that are infected programs. Worms are full programs; they can spread by e-mail, but directly propagating worms can propagate on their own, taking advantage of unpatched vulnerabilities in victim hosts. Some vulnerability-enabled worms can spread through the Internet host population with amazing speed. Many worms and viruses carry damaging payloads. Often, payloads place Trojan horse programs or other types of exploitation software on the victim computer. Malware is the general name for evil software.
- Viruses, worms, and Trojan horses are not the only attacks that are aimed at individuals. Spam deluges the victim with unsolicited commercial e-mail, and messages often are fraudulent. Spyware collects information about users and sends this information to an attacker. Phishing attacks use an official-looking e-mail message or website to trick users into divulging passwords and other special information. Attacks on individuals, including e-mail virus and worm attacks, often depend on social engineering—tricking the victim into doing something against his or her best interests. Two common goals of attacks on individuals are credit card number theft, in which a credit card number is stolen, and identity theft, in which enough private information is stolen to enable the attacker to impersonate the victim in large financial transactions.
- Hacking is the intentional use of a computer resource without authorization or in excess of authorization. Hacking break-ins typically require a prolonged series of probing actions on the part of the attacker.
- Denial-of-service (DoS) attacks overload victim servers so that they cannot serve users. Distributed DOS (DDoS) attacks use bots to carry out the attack. Bots can be updated to take on new functionality.

ATTACKERS Traditionally, most attackers were curiosity-driven hackers and disgruntled employees and ex-employees. Now, criminals dominate the attack world, and the money their crimes generate enables them to invest in new technology and hire top hackers. On the horizon, cyberterror attacks by terrorists and cyberwar attacks by national governments could do unprecedented levels of damage.

SECURITY MANAGEMENT Security is primarily a management issue, not a technical issue. Planning involves risk analysis (balancing the costs and benefits of protections), creating comprehensive security (closing all avenues of attack), using defense in depth (establishing successive lines of defense in case one line of defense fails), and access control using minimum permissions.

We looked at policy-based security in which a high-level policy group creates security policies and lower-level staff members implement the policy. Policies specify what (is to be done). Implementation focuses on how (to do it). This division of labor works because high-level policy people have a broad understanding of security risks and must create policies that will give comprehensive security. Implementation is done by lower-level staff members who know the technology and local situation in detail. Sometimes, the policy group creates intermediate implementation guidance consisting of standards (which must be followed) and guidelines that must be considered, although they do not have to be followed if there is good reason not to.

CONTROL AND AUTHENTICATION Firms need to control access to their assets. Access control normally requires authentication—proving the identity of the person wishing access. The person requesting access is the supplicant, and the device requiring proof of identity is the verifier. The supplicant sends credentials to the verifier to prove the supplicant's identity. For consistency, a central authentication server is used to do the credential checking. There are several common authentication technologies.

- Passwords are inexpensive and easy to use, but users typically choose poor passwords that are easy to crack. Passwords should be used only for low-sensitivity resources.
- Access cards are often used for door entry.
- Biometrics promises to use bodily measurements to authenticate supplicants, replacing other forms of authentication. Concerns with biometrics include cost, error rates, and the effectiveness of deliberate deception by supplicants.
- Digital certificate authentication at the other extreme gives the strongest authentication, but it is complex and expensive to implement.

FIREWALLS Firewalls examine packets passing through the firewall. If a firewall finds provable attack packets, it drops them and records information about them in a log file. If a packet is not a provable attack packet—even if it really is an attack packet—the firewall will not drop it. Ingress filtering examines packets coming into the firm; egress filtering examines packets going out of the firm.

Most firewalls use stateful inspection, which divides communication into stages called states. During the risky connection-opening state, the firewall does extensive work to decide whether to allow connections by passing packets attempting to open a connection. This requires examining every connection-opening packet against rules in an access control list (ACL).

If a packet that does not attempt to open a connection arrives, then it is checked against allowed connections. If it is part of an approved connection, it is passed—usually

with little or no additional filtering. If the packet that does not try to open a connection is not part of an approved connection, it is dropped.

Stateful packet inspection firewalls give strong security during the connection-opening state. This is processing intensive, but few packets are parts of connection-opening attempts. For most packets, SPI firewalls do simple inspection, which requires little processing power.

CRYPTOGRAPHIC SYSTEMS Cryptography is the use of mathematics to protect message dialogues. One key protection is encryption for confidentiality, which encrypts messages to prevent attackers from reading any messages that they intercept. Encryption methods are called ciphers. In symmetric key encryption, both sides encrypt and decrypt with a single key. To be strong, a symmetric encryption key needs to be more than 100 bits long.

In addition to providing message-by-message encryption, cryptographic systems also provide message-by-message authentication by adding an electronic signature to each message. Electronic signatures also give message integrity.

Cryptographic protections are organized into cryptographic system standards. SSL/TLS provides medium-strength protection; it is built into all browsers and web-servers. IPsec is a very strong cryptographic system standard. It protects IP packets and encapsulated transport layer and application layer messages.

RESPONSE Protections occasionally break down. Response is reacting to compromises according to plan. The stages in response to attack typically include identifying the attack, stopping the attack, recovering from the attack, and (sometimes) punishing the attacker. Major incidents require the convening of a computer security incident response team (CSIRT). IT disaster recovery requires getting IT back in operation at another site, while business continuity recovery involves getting the entire firm back in operation. It is important for recovery teams to conduct rehearsals before problems occur so that they can respond quickly and correctly.

END-OF-CHAPTER QUESTIONS

Thought Questions

1. a) Suppose that an attack would do $100,000 in damage and has a 15 percent annual probability of success. Spending $9,000 per year on "Measure A" would cut the annual probability of success by 75 percent. Do a risk analysis comparing benefits and costs. Show your work clearly. b) Should the company spend the money? Explain. c) Do another risk analysis if Measure A costs $20,000 per year. Again, show your work. d) Should the company spend the money? Explain.
2. a) What form of authentication would you recommend for relatively unimportant resources? Justify your answer. b) What form of authentication would you recommend for your most sensitive resources?
3. For each of the following passwords, first state the kind of attack that would be necessary to crack it. Justify your answer. Then say whether or not it is an adequate password, again giving specific reasons. a) swordfish; b) Processing1; c) SeAtTLe; d) 3R%t; and e) 4h*6tU9$^l.

4. Keys and passwords must be long. Yet most personal identification numbers (PINs) that you type when you use a debit card are only four or six characters long. Yet this is safe. Why?
5. Revise the ACL in Figure 3-18 to permit access to an FTP server with IP address 10.32.67.112.
6. In digital certificate authentication, the supplicant could impersonate the true party by doing the calculation with the true party's private key. What prevents impostors from doing this?

Online Exercise

1. Go to http://www.cybercrime.gov. Go to the section on computer crimes. Select one of the cases randomly. Describe the type of attacker and the type of attack(s).

Case Study: Patco

In 2009, the Patco Construction Company had $588,000 drained from its bank accounts at Ocean Bank. The theft involved six withdrawals on May 8, May 11, May 12, May 13, May 14, and May 15. The money in each withdrawal was sent to a group of money mules.

After thieves stole all of the company's cash, they continued to make withdrawals. Patco's bank continued to allow withdrawals, covering them with over $200,000 from Patco's line of credit. Although the bank was able to recover or block $243,406 in transfers, Patco was still out $345,400. In addition, the bank began charging Patco for interest on the money that had been withdrawn using Patco's line of credit.

Although the transactions were far larger than Patco normally made, Ocean Bank did not inform Patco of any problems until one of the account numbers entered by the thieves was invalid. It sent a notification by mail, and it did not arrive at Patco until several days later. Patco notified the bank of problems the next morning. However, the bank had already sent out $111,963 that day, some of which was recovered.

The bank used account numbers and passwords. For transactions over $1,000, Patco employees had to answer two challenge questions. Most withdrawals were over $1,000, so employees had to answer these same challenge questions many times. Patco believes that these challenge messages were too easy.

The State of Maine has stringent banking laws. The Federal Financial Institutions Examination Council in 2005 required banks to use at least two-factor authentication and specifically noted that usernames and passwords were not enough. Patco sued People's United Bank for its losses, claiming that the challenge questions were nothing more than a second set of passwords and that the bank should have required much stronger credentials.

Patco also claimed that Ocean Bank should have been suspicious when such large unprecedented withdrawals were made and when they were sent to 30 different accounts. Normally, Patco only withdrew money for payrolls on Fridays. Its previous largest single-day withdrawal had been under $37,000. Patco's complaint stated that based on belief and information from the bank, Patco assumed that antifraud monitoring was being done by the bank.

Ocean Bank did not comment on the case, but most banks in a similar situation use the defense that they were not negligent. A bank can be found negligent only if it has lower protections than are the norm in the industry.

CAUTION: The information in this case is based only on Patco's complaint.[14] Consequently, the statements made in the case have not been validated and may be disputed by Ocean Bank as being nonfactual. Analyze the case based on Patco's allegations, but do not draw firm conclusions against the bank.

1. a) According to the information in the case, do you think the bank satisfied the requirement to use two-factor authentication? b) According to the information in the case, do you think the bank was doing antifraud monitoring? c) According to the information in the case, do you think Ocean Bank was negligent? d) According to the information in the case, if you were the head of Ocean Bank, what would you do to prevent the reoccurrence of this problem?

Case Study: Walmart

In 2005, Hurricane Katrina slammed into Louisiana and Mississippi, devastating New Orleans and many other cities along the U.S. Gulf Coast. Shortly afterward, the fourth most intense Atlantic hurricane in history, Rita, added enormously to the destruction. The Federal Emergency Management Agency (FEMA) became notorious for its handling of the crisis, responding belatedly and acting ineptly when it did respond.

Many businesses collapsed because they were poorly prepared for the hurricanes. One company that *did* respond effectively was Walmart.[15] In its Brookhaven, Mississippi, distribution center, the company had 45 trucks loaded and ready for delivery even before Katrina made landfall. The company soon supplied $20 million in cash donations, 100,000 free meals, and 1,900 truckloads full of diapers, toothbrushes, and other emergency supplies to relief centers. The company also supplied flashlights, batteries, ammunition, protective gear, and meals to police and relief workers.

Although the relief effort was impressive, it was merely the visible tip of Walmart's disaster recovery program. Two days before Katrina hit, Walmart activated its business continuity center. Soon, 50 managers and experts in specific areas such as trucking were hard at work. Just before the storm knocked out the company's computer network, the center ordered the Mississippi distribution center to send out recovery merchandise such as bleach and mops to its stores. The company also sent 40 generators to its stores so that stores that lost power could open to serve their customers. It also sent out many security employees to protect stores.

After computer networks failed, the company relied on the telephone to contact its stores and other key constituencies. Most stores came back immediately, and almost all stores were able to serve their customers within a few days. Lines of customers were long, and Walmart engaged local law enforcement to help maintain order.

Walmart was successful because of intensive preparation. The company has a full-time director of business continuity. It also has detailed business continuity plans and clear lines

[14] *Patco Construction Company, Inc., plaintiff, v. People's United Bank*, d/b/a Ocean Bank, defendant. State of Maine, York SS Superior Court Civil Action, Docket No. 09-CV.

[15] Liza Featherstone, "Wal-Mart to the Rescue!" *The Nation*, September 26, 2005. www.thenation.com/doc/20050926/featherstone. Michael Barbaro and Justin Gillis, "Wal-Mart at Forefront of Hurricane Relief," *Washington Post.com*, September 6, 2005. www.washingtonpost.com/wp-dyn/content/article/2005/09/05/AR2005090501598.html. *NewsMax.com*, "Wal-Mart Praised for Hurricane Katrina Response Efforts," www.newsmax.com/archives/ic/2005/9/6/164525.shtml. Ann Zimmerman and Valerie Bauerlein, "At Wal-Mart, Emergency Plan has Big Payoff," *Wall Street Journal*, September 12, 2005, B1.

of responsibility. In fact, while the company was still responding to Katrina and Rita, it was monitoring a hurricane off Japan, preparing to take action there if necessary.

1. a) Why was Walmart able to respond quickly? b) List at least three actions that Walmart took that you might not have thought of.

Perspective Questions

1. What was the most surprising thing you learned in this chapter?
2. What was the most difficult part of this chapter for you?

4 NETWORK MANAGEMENT

LEARNING OBJECTIVES

By the end of this chapter, you should be able to:

- Explain general concepts in network management, including a focus on the system life cycle, the importance of cost, and strategic network planning.
- Use quality-of-service (QoS) criteria in product selection.
- Do basic design work, including doing traffic analysis with and without redundancy, knowing common topologies, selecting topologies, understanding how to handle momentary traffic peaks, understanding how to reduce capacity needs through traffic shaping and compression, and topologies.
- Evaluate alternatives using multicriteria decision making and specifying costs.
- Describe operational management, including OAM&P, the Simple Network Management Protocol (SNMP), and network management software.

INTRODUCTION

Today, we can build much larger networks than we can manage easily. For example, even a mid-size bank is likely to have 500 Ethernet switches and a similar number of routers. Furthermore, network devices and their users are spread out over large areas—sometimes international areas. While network technology is exciting to talk about and concrete conceptually, network management is where the rubber meets the road.

SDLC versus SLC

In programming and database courses, you focus on the systems development life cycle (SDLC), which looks at your information system from its moment of conception to its implementation. When you think about this, it is rather strange. Teaching the SDLC is like training new doctors in obstetrics and ignoring all training for illnesses after birth. Although the SDLC discusses episodic software maintenance activities, Figure 4-1 shows that the SDLC is far more limited than the **systems life cycle (SLC)**, which lasts from conception until death.

In networking, the creation of new networks and the modification of old networks are important, but the real work of networking professionals is the administration of ongoing networks. While systems administrators who manage servers have a difficult

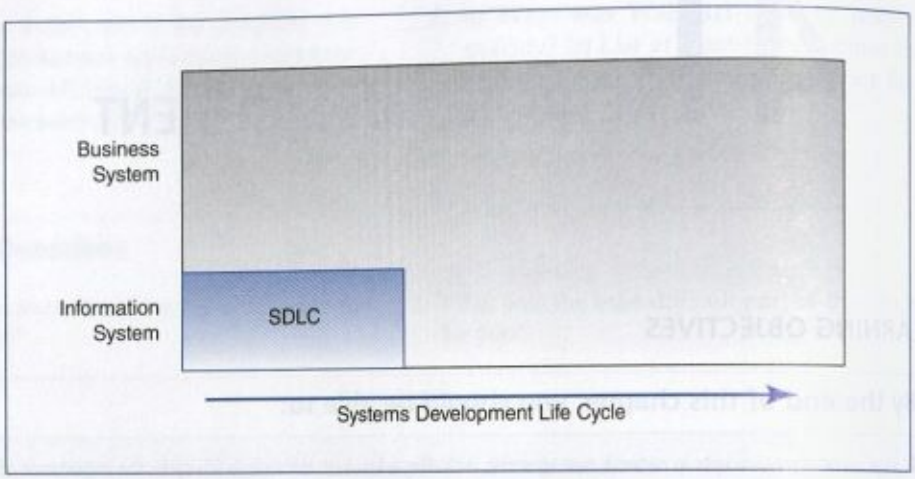

FIGURE 4-1 The Systems Development Life Cycle versus Business Systems and the Systems Life Cycle

job, especially in today's world of virtualization, network administrators face very complex tasks that require a high level of understanding of how networks work. The operational phase of a network after its creation is an enormous part of the network professional's job.

Figure 4-1 also shows that while database and programming professionals have to focus on the information system, the network professional often has to focus on the broader business system in which the information system is embedded. This is particularly true in network security.

Test Your Understanding

1. a) Why must networking professionals be concerned with the SLC rather than the SDLC? b) Why is a focus on information systems insufficient in networking?

Cost

In networking, you can never say, "Cost doesn't matter." Figure 4-2 illustrates that network demand is likely to grow rapidly in the future, just as it has always done in the past. The figure also illustrates that network budgets are growing very slowly (if they are growing at all).

Taken together, these curves mean that network budgets are always stretched thin. If the network staff spends too much money on one project, it will not have enough money left over to do another important project. Although there are many concerns beyond costs, cost is always an important consideration in network management.

Test Your Understanding

2. a) Compare trends in network demand and network budgets. b) What are the implications of these trends?

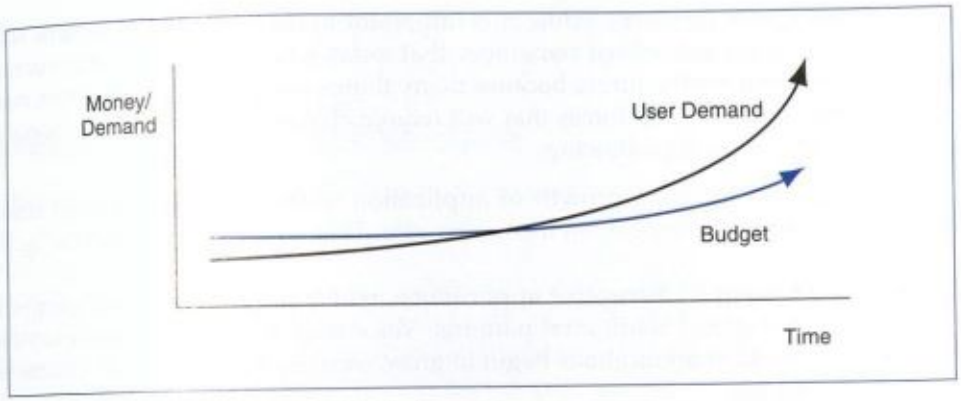

FIGURE 4-2 Network Demand and Budget

Strategic Network Planning

One of the most important things to plan in networking is the technological infrastructure—the firm's arrangement of hardware, software, and transmission lines that allows the network to carry information. Figure 4-3 shows how organizations do strategic network planning.

WHAT-IS ANALYSIS Planning for changes in the technological infrastructure must begin with the "what is," that is, the current state of the company's network. This may sound like an easy task, but most firms do not have a thorough understanding of their network components and interactions, much less of their trouble spots. What-is analysis begins with an exhaustive inventory of the network's components and their interrelationships. This sounds simple. It is not.

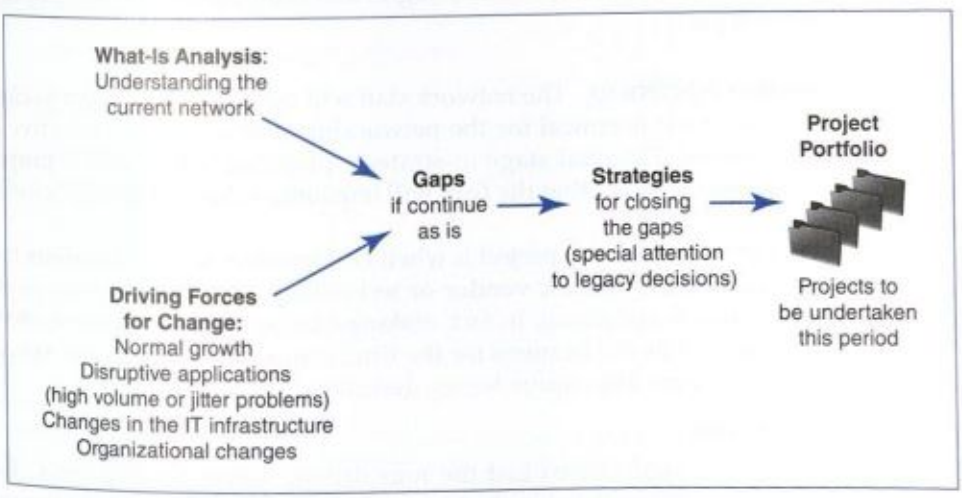

FIGURE 4-3 Strategic Network Planning

DRIVING FORCES FOR CHANGE While it is important to understand the current state of the network, companies must remember that today's technological infrastructure will not be sufficient for the future because many things will change. Companies need to consider the major driving forces that will require changes in the network. Some of these driving forces are the following:

- The normal continuing growth of application traffic demand. In most firms, traffic has been growing at an increasing rate. This will certainly continue in the future.
- The introduction of disruptive applications, which may create major surges in demand far beyond traditional patterns. Voice over IP is an obvious example. However, if video applications begin to grow, capacity planning will become extremely difficult.
- Changes in other elements of the IT technological infrastructure can also require extensive network changes. One long-term trend has been the consolidation of data centers from many to few. This can radically change traffic flows within the corporate network.
- Organizational change can be a major driving force. If a company is adding a site, not only will the site have to be served but communication between different parts of the firm will change, depending on what units are moved there. In fact, all corporate reorganizations are likely to impact network planning. At the extreme are nightmare scenarios that exist if the company is bought out or buys out another firm.

GAPS ANALYSIS Comparing driving forces to the what-is network will create inevitable gaps between what the firm will need and what the current network can provide. These gaps must be identified, characterized, and documented.

STRATEGIES FOR CLOSING THE GAPS The firm then needs to develop strategies for closing the gaps. It must consider multiple technologies and multiple topologies (physical connections) for each gap.

SELECTING A PROJECT PORTFOLIO The network staff will not have the budget to close all gaps satisfactorily. So it is critical for the networking staff to be very selective in how it spends its money. The final stage in strategic planning is to create a **project portfolio**—a selection of projects that the firm will implement during the plan's initial period.

A strong consideration for any project is whether it involves **legacy decisions** that will lock the company into a specific vendor or technology option for several years. Making legacy decisions is not wrong. In fact, making a legacy decision is often necessary. However, because of its implications for the firm, companies need to give special scrutiny to potential projects that require legacy decisions.

Test Your Understanding

3. a) What is what-is analysis? b) List the four driving forces for change. c) For each, give an example not listed in the text. d) What is gaps analysis? e) Why is it

necessary to create a project portfolio? f) What are legacy decisions, and why must projects that involve legacy decisions be judged with great care?

NETWORK QUALITY OF SERVICE (QoS)

In the early days of the ARPANET and the Internet, networked applications amazed new users. However, these users soon said, "Too bad this thing doesn't work better." Today, networking is a mission-critical service for corporations. If the network breaks down, much of the organization comes to a halt. Today, networks must work, and they must work *well*. Companies are concerned with network **quality-of-service (QoS) metrics**, that is, quantitative measures of network performance. Figure 4-4 shows that companies typically use a number of QoS metrics to quantify their quality of service so that they can set targets and determine if they have met those targets.

Test Your Understanding

4. a) What are QoS metrics? (Do not just spell out the acronym.) b) How are QoS metrics used?

Transmission Speed

There are many ways to measure how well a network is working. The most fundamental metric, as we saw in Chapter 1, is speed. While low speeds are fine for text messages, the need for speed becomes very high as large volumes of data must be sent, and video transmission requires extremely high transmission speed.

BITS PER SECOND (BPS) As we saw in Chapter 1, transmission speed[1] normally is measured in bits per second (bps). A bit is either a one or a zero. Obviously, a single bit

Quality of Service (QoS) Metric

Quantifiable measures of network performance

Examples

Speed
Availability
Error rates
...

FIGURE 4-4 Network Quality of Service (QoS)

[1]Purists correctly point out that *speed* is the wrong word to use to describe transmission rates. At faster transmission rates, bits do not physically travel faster. The sender merely transmits more bits in each second. Transmission rates are like talking faster, not running faster. However, transmission rates are called transmission speeds almost universally, so we will follow that practice in this book.

cannot convey much information. Speeds today range from thousands of bits per second to billions of bits per second. To simplify the writing of transmission speeds, professionals add metric prefixes to the base unit, bps. For example, Figure 4-5 shows that in increasing factors of 1,000 (not 1,024 as with computer memory), we have kilobits per second (kbps), megabits per second (Mbps), gigabits per second (Gbps), and terabits per second (Tbps).

Speeds are measured in factors of 1,000, not 1,024.

Consistent with metric notation, kilo is abbreviated as lowercase k instead of uppercase K. However, megabits per second is Mbps, gigabits per second is Gbps, and terabits per second is Tbps.

Speed

Normally measured in bits per second (bps)

Not bytes per second

Metric suffixed in increasing units of 1,000 (not 1,024)

The metric abbreviation for kilo is lowercase k

Abbreviation	Meaning	Name	Example
1 kbps	1,000 bps	kilobits per second	33 kbps is 33,000 bps
1 Mbps	1,000 kbps	megabits per second	3.4 Mbps is 3,400,000 bps
1 Gbps	1,000 Mbps	gigabits per second	62 Gbps is 62,000,000,000 bps
1 Tbps	1,000 Gbps	terabits per second	

Sometimes speed is measured in bytes per second, Bps, compared to bps

Bps usually only for file transfers

Expressing Speed in Proper Notation

As Written	Places before Decimal Point	Space between Number and Suffix?	Properly Written
23.72 Mbps	2	Yes	23.72 Mbps
2,300 Mbps	4	No	2.3 Gbps
0.5 Mbps	0 (leading zeros do not count)	No	500 kbps

There must be one to three spaces before the decimal point

Leading zeros do not count

(continued)

FIGURE 4-5 Transmission Speed

There must be a space between the number and the units

12 Mbps is proper; 12Mbps is improper

If the number is decreased by 1,000 (4,523 becomes 4.523), then the suffix must be increased by a thousand (kbps to Mbps)

4,523 kbps becomes 4.523 Mbps

(4,523/1000 * kbps * 1000)

If the number is increased by 1,000 (0.45 becomes 450), then the suffix must be decreased by 1,000 (Mbps to kbps)

0.45 Mbps becomes 450 kbps

Rated Speed and Throughput

Rated Speed

The speed a system *should* achieve

According to vendor claims or to the standard that defines the technology

Throughput

The data transmission speed a system *actually* provides to users

Aggregate versus Rated Throughput on Shared Lines

The aggregate throughput is the total throughput available to all users

The individual throughput is an individual's share of the aggregate throughput

FIGURE 4-5 Continued

WRITING NUMBERS IN PROPER NOTATION Networking professionals write speeds in a very specific way. The basic rule for writing speeds (and metric numbers in general) in proper notation is that there should be one to three places before the decimal point and that there should be a space between the number and the units. Figure 4-5 illustrates how to write speeds properly.

To write a speed in proper notation, there should be one to three places before the decimal point, and there should be a space between the number and the units.

- Given this rule, 23.72 Mbps is fine (two places before the decimal point and a space between the number and the metric prefix).
- However, 2,300Mbps has four places before the decimal point (2,300.00), so it should be rewritten as 2.3 Gbps (one place). Note that a space has been added between the number and its metric prefix. In turn, 0.5 Mbps has zero places to the left of the decimal point. (Leading zeros do not count.) It should be written as 500 kbps (three places).

When you look at a number in metric notation, remember that if a=b*c, then a also equals b*1000*c/1000 or b/1000*b*1000. Think of the metric prefix k as 1,000 and the metric prefix M as 1,000,000.

Suppose you have the speed 4,523 kbps. To write the number properly, you divide it by 1,000 to get 4.523. If you divide the number by 1,000, then you must multiply the

prefix by 1,000. Multiplying kbps by 1,000 gives Mbps, so the number in proper notation becomes 4.523 Mbps.

To give another example, suppose you have 0.45 Mbps. You need to multiply the number by 1,000, getting 450. You then have to divide the prefix (Mbps) by 1,000 to give you kbps. The number in proper notation, then, is 450 kbps.

RATED SPEED VERSUS THROUGHPUT

NOTE: Some students find the distinction between rated speed and throughput difficult to learn. However, we must use this distinction throughout this book, so be sure to take the time to understand it.

Talking about transmission speed can be tricky. A network's **rated speed** is the speed it *should* achieve based on vendor claims or on the standard that defines the technology. For a number of reasons, networks often fail to deliver data at their rated speeds. In contrast to rated speed, a network's **throughput** is the data transmission speed the network *actually* provides to users.

Throughput is the data transmission speed a network actually *provides to users.*

AGGREGATE VERSUS INDIVIDUAL THROUGHPUT When a transmission line on a network is multiplexed, this means that several conversations between users will share the line's throughput. Consequently, it is important to distinguish between a line's **aggregate throughput**, which is the total it provides to all users who share it, and the **individual throughput** that single users receive as their shares of the aggregate throughput. As you learned as a child, despite what your mother said, sharing is bad.

Test Your Understanding

5. a) In what units is transmission speed normally measured? b) Is speed normally measured in bits per second or bytes per second? c) Give the names and abbreviations for speeds in increasing factors of 1,000. d) What is 55,000,000,000 bps with a metric prefix? e) Write out 100 kbps in bits per second (without a metric prefix). f) Write the following speeds properly: 0.067 Mbps, 23,000 kbps, and 48.62Gbps.
6. a) Distinguish between rated speed and throughput. b) Distinguish between individual and aggregate throughput.

Other Quality-of-Service Metrics

Although network speed is important, it is not enough to provide good quality of service. We will look briefly at other important QoS metric categories.

AVAILABILITY One of these other metrics is **availability**, which is the percentage of time that the network is available for use. In contrast, **downtime** is the percentage of time that the network is not available.

Ideally, systems would be available 100 percent of the time, but that is impossible in reality. On the Public Switched Telephone Network, the availability target usually is

Availability

- The percentage of time a network is available for use
- Downtime is the amount of time a network is unavailable (minutes, hours, days, etc.)

Error Rates

- Require retransmissions
- When an error occurs, TCP assumes there is congestion and slows its rate of transmission
- Packet error rate: The percentage of packets that have errors
- Bit error rate: The percentage of packets that have error

Latency and Jitter

- Latency
  - Delay, measured in milliseconds
- Jitter (Figure 4-7)
  - Variation in latency between successive packets
  - Makes voice sound jittery

Application Response Time

- The time from when the user hits a key and when the system responds. (Figure 4-8)
- Includes the two-way network latency
- Includes contributions from the host or application program
- Often, configuration problems produce application response time problems
- Improvement requires cooperation between networking and host administration

Service Level Agreements

- Guarantees for performance
- Penalties if the network does not meet its service metrics guarantees
- Guarantees specify worst cases (no worse than)
  - Lowest speed (e.g., no worse than 1 Mbps)
  - Maximum latency (e.g., no more than 125 ms)
- Often written on a percentage basis
  - No worse than 100 Mbps 99.5% of the time

FIGURE 4-6 Quality of Service II (Study Figure)

99.999 percent. This is known as the "five nines." Data networks generally have lower availability but are under pressure to improve their availability given the cost of network downtime to firms today.[2]

[2]On a more detailed basis, availability can be discussed in terms of the mean time to failure (MTTF) and the mean time to repair (MTTR). The former asks how frequently downtime occurs. The latter asks how long service is down after a failure begins. More short failures may be preferable to infrequent but very long outages.

ERROR RATES Hosts send data in small messages called packets. Ideally, all packets would arrive intact, but this does not always happen. The **packet error rate** is the percentage of packets that are lost or damaged during delivery. The **bit error rate**, in turn, is the percentage of bits that are lost or damaged.

Most networks today have very low average error rates. However, when the network is overloaded, error rates can soar because the network has to drop the packets it cannot handle. Companies must measure error rates when traffic levels are high in order to have a good understanding of error rate risks.

The impact of even small error rates can be surprising. TCP is designed to avoid network congestion by generating TCP segments slowly at the beginning of a connection. If the segments arrive correctly, TCP generates segments more quickly. However, if there is an error or if an acknowledgment is lost, the TCP process assumes that the network is overloaded. It falls back to its initial slow start rate for creating TCP segments. Consequently, even a small error rate can produce a major drop in throughput for applications.

LATENCY AND JITTER When packets move through the network, they will encounter some delays. The amount of delay is called **latency**. Latency is measured in **milliseconds (ms)**. A millisecond is a thousandth of a second. When latency reaches about 125 milliseconds, turn-taking in telephone conversations becomes difficult.

A related concept is **jitter**, which Figure 4-7 illustrates. Jitter occurs when the latency between successive packets varies. Some packets will come too far apart in time, others too close in time. While jitter does not bother most applications, VoIP and streaming media are highly sensitive to jitter. If the sound is played back without adjustment, it will speed up and slow down. These variations often occur over millisecond time periods, and, as the name suggests, variable latency tends to make voice sound jittery.

Most networks were engineered to carry traditional data such as e-mail and database transmissions. In traditional applications, latency was only slightly important, and jitter was not important at all. However, as voice over IP and video over IP, which are sensitive to jitter and to some extent latency, have grown in importance, companies have begun to worry more about latency and jitter. They are finding that extensive network redesign may be needed to give good control over latency and jitter. This may include fork-lift upgrades for many of its switches and routers.

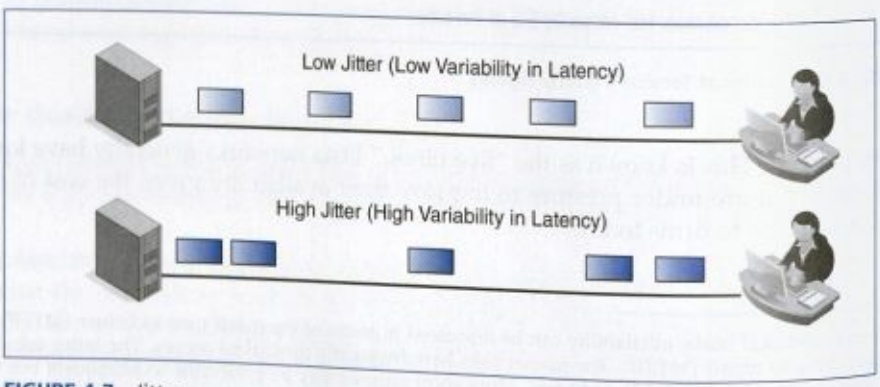

FIGURE 4-7 Jitter

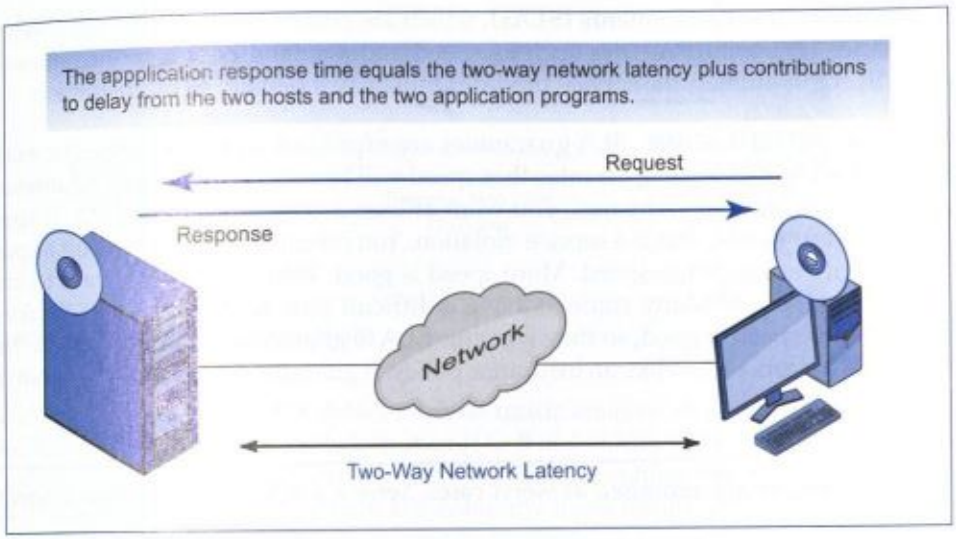

FIGURE 4-8 Application Response Time

APPLICATION RESPONSE TIME The most challenging QoS metric is **application response time.** This is the duration between the time the user presses a key (or clicks on the page) and the time he or she sees a response. Figure 4-8 shows how application response time is different from network latency.

The figure shows that network latency is only one factor in user response time. Most obviously, the delay at the client and server ends to do processing in the application software is important. However, there can be many other factors in response time. For example, poorly configured networking software in the client or server operating system may add delay. So may firewall filtering, the need for encryption (which is a heavy process), and other security matters.

Application response time planning and execution are complicated by the fact that systems administrators and application professionals often work separately from network professionals. In fact, they often know little about what the other side does. Application response time management requires strong and effective cooperation across these organizational boundaries.

Test Your Understanding

7. a) What is availability? b) What is downtime? c) What are the "five nines"? d) Does corporate network availability usually meet the five-nines expectation of the telephone network? e) What are packets? f) Distinguish between the packet error rate and the bit error rate. g) When should error rates be measured? Why? h) What is latency? i) In what units is latency measured? j) What is jitter? k) For what applications is jitter a problem? l) How does application response time differ from latency? m) Why is application response time difficult to improve?

Service Level Agreements (SLAs)

When you buy some products, you receive a guarantee promising that they will work and specifying penalties if they do not work. In networks, service providers often

provide **service level agreements (SLAs)**, which are contracts that guarantee levels of performance for various metrics such as speed and availability. If a service does not meet its SLA guarantees, the service provider must pay a penalty to its customers.

WORST-CASE SPECIFICATION SLA guarantees are expressed as **worst cases**. For example, an SLA for speed might guarantee that speed will be *no lower* than a certain amount. If you are downloading webpages, you want at least a certain level of speed. If speed falls below your needs, that is a service violation. You certainly would not want a speed SLA to specify a maximum speed. More speed is good. Why would you want to limit the maximum speed? Many students have a difficult time with worst-case thinking. This is because speed is good, so they want the SLA to guarantee them high speed. That is not the SLA's job. SLA is like an insurance policy. It guarantees that speed will not fall below a certain level.

SLA guarantees are expressed as worst cases. *Service will be no worse than a specific number.*

For latency, then, the SLA will require that latency will be *no higher* than a certain value. You might specify an SLA guarantee of 125 ms (milliseconds). This means that you will not get worse latency.

PERCENTAGE-OF-TIME ELEMENTS In addition, SLAs have percentage-of-time elements. For instance, an SLA on speed might guarantee a speed of at least 480 Mbps 99.9 percent of the time. This means that the speed will nearly always be at least 480 Mbps but may fall below that 0.1 percent of the time without incurring penalties. A smaller exception percentage might be attractive to users, but it would require a more expensive network design. Nothing can be guaranteed to work properly 100 percent of the time.

Test Your Understanding

8. a) What are service level agreements? b) Does an SLA measure the best case or the worst case? c) Would an SLA specify highest latency or lowest latency? d) Would an SLA specify a lowest availability or a highest availability? e) What happens if a carrier does not meet its SLA guarantee? f) If carrier speed falls below its guaranteed speed in an SLA, under what circumstances will the carrier not have to pay a penalty to the customers?

DESIGN

Implementing a network project requires a company to go through all phases of the systems development life cycle. In most cases, these stages are similar to those for other IT projects. One special area in the SDLC is the design of a new network or of a modified network.

Traffic Analysis

Network design always begins with traffic requirements. In network design, traffic analysis asks how much traffic must flow over the network's many individual

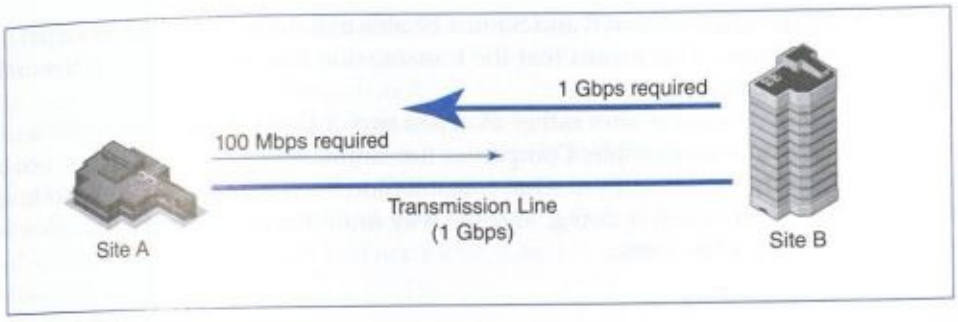

FIGURE 4-9 Two-Site Traffic Analysis

transmission lines. Figure 4-9 shows a trivial traffic analysis. A company only has two sites, A and B. A needs to be able to transmit to B at 100 Mbps. B needs to transmit to A at 1 Gbps. Transmission lines usually are symmetric, meaning that they have the same speed in both directions. Obviously, the company must install a transmission line that can handle 1 Gbps.

As soon as the number of sites becomes larger than two, traffic analysis becomes difficult. Figure 4-10 shows a three-site traffic analysis. For simplicity, we will assume that transmission is symmetric between each pair of sites.

The figure shows that Site Q attaches to Site R, which attaches to Site S. Site Q is 130 meters west of Site R. Site S is 180 meters east of Site R. Site Q needs to be able to communicate with Site R at 45 Mbps. Site R needs to be able to communicate with Site S at 2 Gbps. Site Q needs to be able to communicate with Site S at 300 Mbps.

Are you confused by the last paragraph? Anyone would be. In traffic analysis, it is critical to DTP—draw the picture. Figure 4-10 shows how the three sites are laid out. After laying out the sites, you draw the three required traffic flows.

Note that the line between Q and R must handle both Q–R traffic (45 Mbps) and the Q–S traffic (300 Mbps). It does not handle any of the traffic between R and S, however. Consequently, the line between Q and R must be able to handle 345 Mbps.

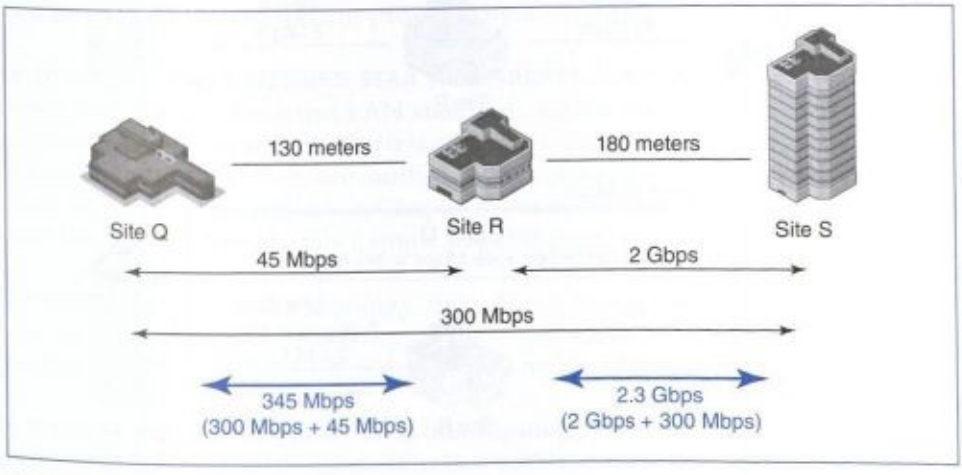

FIGURE 4-10 Three-Site Traffic Analysis

Similarly, the line between R and S must be able to handle R–S traffic (2 Gbps) and Q–S traffic (300 Mbps). This means that the transmission line between R and S must be able to handle 2.3 Gbps.

If a company has many sites rather than just two or three, then doing traffic analysis manually becomes impossible. Companies use simulation programs that determine what site pair traffic will flow over what transmission lines. However, you need to understand what the program is doing, and the way to do that is to work through a few examples with only a few sites.

Test Your Understanding

9. Do a three-site traffic analysis. Site X attaches to Site Y, which attaches to Site Z. Site X is 130 meters east of Site Y. Site Z is 180 meters west of Site Y. Site X needs to be able to communicate with Site Y at 3 Gbps. Site Y needs to be able to communicate with Site Z at 1 Gbps. Site X needs to be able to communicate with Site Z at 700 Mbps. Supply your picture giving the analysis. You may want to do this in Office Visio or Windows Draw and then paste it into your homework. a) What traffic capacity will you need between Sites X and Y? b) Between Y and Z?

Redundancy

Transmission lines sometimes fail. Suppose that the transmission line between R and S in Figure 4-10 failed. Then Q would still be able to talk to R, but Q and R would not be able to talk to S. Obviously, this is highly undesirable.

The solution is to install redundant transmission lines. Redundant transmission lines are extra transmission lines that are not necessary for the system to function but that provide backup paths in case another line fails. For example, Figure 4-11 again shows Sites Q, R, and S. This time, a redundant line has been added between Q and R.

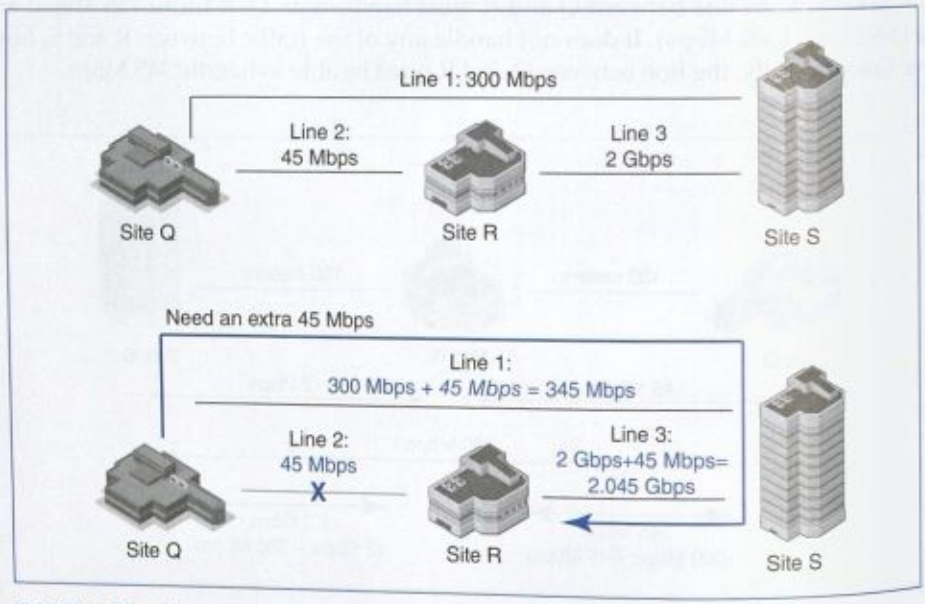

FIGURE 4-11 Three Sites with Redundancy

What happens if the line between Q and R fails? The answer is that Site Q can still talk to Site S through the direct line. Also, Q can still talk to R by sending its transmissions to S, which will send them on to R.

When redundancy is used, lines must be given extra capacity in case of failures. For instance, if the line between Q and R is only 300 Mbps, this will be enough if there are no failures. However, if the line Q–R fails, the line will need another 45 Mbps. So it will need to have 345 Mbps of capacity to handle a Q–R failure. The R–S line will also need 45 Mbps more capacity. It will need 2.045 Gbps of capacity to handle both R–S traffic and Q–R traffic.

Test Your Understanding

10. a) What is the purpose of redundancy in transmission lines? b) If the line between R and S fails in Figure 4-11, how much capacity will the line between Q and S need? c) What about the line between Q and R?

Topology

Network design focuses heavily on network topology. The term **network topology** refers to the physical arrangement of a network's computers, switches, routers, and transmission lines. Topology, then, is a physical layer concept. Different network (and internet) standards specify different physical topologies. Figure 4-12 shows the major "basic" topologies specified by network standards. Real networks often have complex topologies that involve a mixture of these basic topologies.

Network topology is the physical arrangement of a network's computers, switches, routers, and transmission lines. It is a physical layer concept.

POINT-TO-POINT TOPOLOGY The simplest network topology is the **point-to-point topology**, in which two nodes are connected directly. Although some might say that a point-to-point connection is not a network, companies often connect a pair of sites with a point-to-point private leased line provided by a telephone carrier.

STAR TOPOLOGY AND EXTENDED STAR (HIERARCHY) TOPOLOGY Modern versions of Ethernet, which is the dominant LAN standard, use the star and extended star topologies. In a simple **star topology**, all wires connect to a single switch. In an **extended star (or hierarchy) topology**, there are multiple layers of switches organized in a hierarchy. We will see Ethernet hierarchies in Chapter 5. An important characteristic of hierarchical standards is that there is only a single possible path between any two end nodes.

MESH TOPOLOGY In a **mesh topology**, there are many connections among switches or routers, so there are many alternative paths to get from one end of the network to the other. The TCP/IP standards are designed for a mesh router topology.

BUS (BROADCAST) TOPOLOGIES In a **bus topology**, when a computer transmits, it broadcasts to all other computers. Wireless LANs and WANs, which we will see in Chapters 6 and 7, broadcast their signals and so have bus topologies.

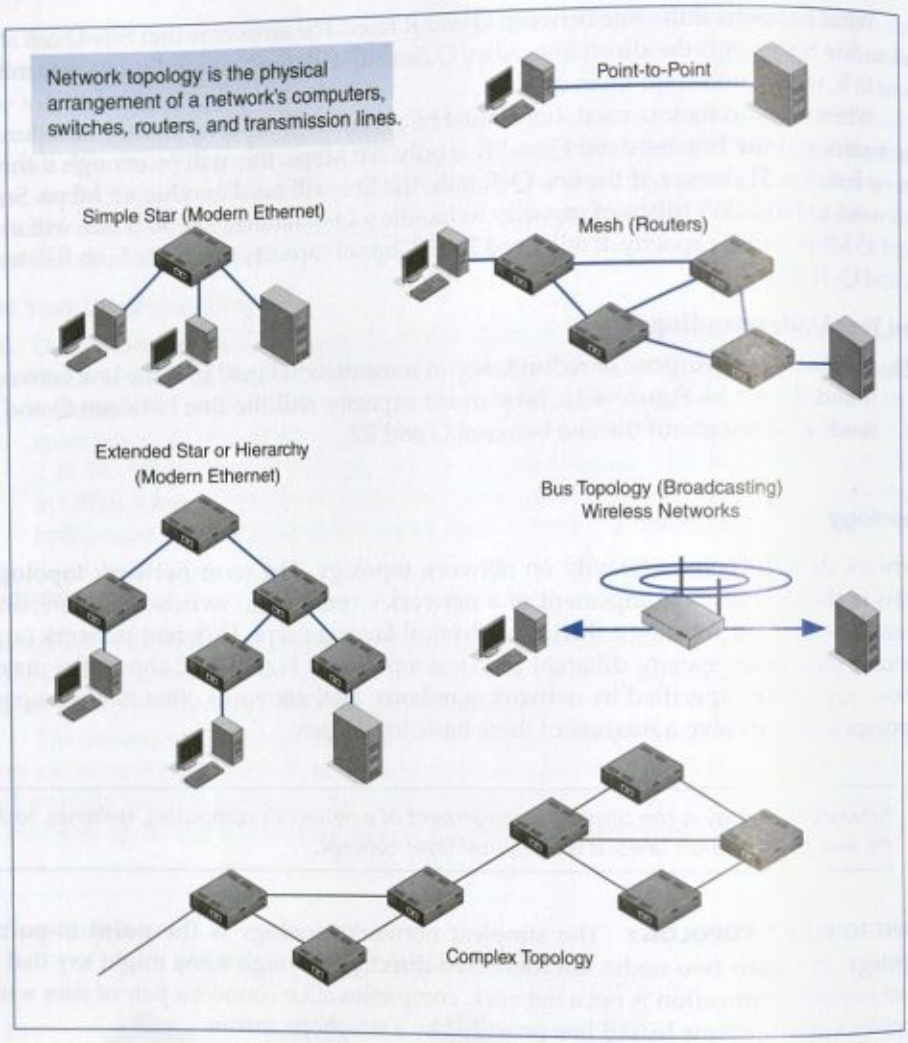

FIGURE 4-12 Major Topologies

REAL NETWORK TOPOLOGIES Some network technologies require a pure basic taxonomy. For example, the Ethernet technology we will see in Chapter 5 requires a strict hierarchy. Most networks, however, have **complex topologies** that use different basic topologies in different parts of the network.

Test Your Understanding

11. a) What is a network topology? b) At what layer do we find topologies? c) In what topology are there only two nodes? d) In what topologies is there only a single path between any two end nodes? e) In what topology are there usually many paths between any two end nodes? f) In what topology is broadcasting used? g) What topologies can be used in complex networks?

Leased Line Network Topologies

Figure 4-13 shows that companies have traditionally used leased lines to interconnect their sites. As we will discuss in Chapter 10, leased lines are high-speed, point-to-point, always-on carrier circuits. Because the telephone system only provides raw bandwidth between points, the company must design the overall data network.

FULL-MESH TOPOLOGY Should many or all pairs of sites be connected to each other, or should there be as few connections as possible? Figure 4-13 shows two topological extremes for building leased line networks.

The first is a **full-mesh topology**, which provides direct connections between every pair of sites. This provides many redundant paths so that if one site or leased line fails, communication can continue unimpeded.

Unfortunately, as the number of sites increases, the cost of a full mesh grows exponentially. For example, if there are N sites, a pure mesh will require $N * (N - 1)/2$ leased lines. So a 5-site pure mesh will require $5 * (5 - 1)/2(10)$ leased lines, a 10-site pure mesh will require 45 leased lines, and a 20-site pure mesh will require 190 leased lines. Full meshes, while reliable, are prohibitively expensive if a company has many sites.

HUB-AND-SPOKE LEASED LINE NETWORKS The second extreme topology for building leased line networks is the pure **hub-and-spoke topology**. Figure 4-13 also shows this topology. In a pure hub-and-spoke topology, all communication goes through one site. This dramatically reduces the number of leased lines required to connect all sites compared to a full mesh, and so this kind of topology minimizes cost. However, it also reduces reliability. If a line fails, there are no alternative paths for reaching an affected site. More disastrously, if the hub site fails, the entire network goes down.

MIXED DESIGNS As you might suspect, full meshes and pure hub-and-spoke topologies represent the extremes of cost and reliability. Most real networks use a mix of these two pure topologies. Real networks must trade off reliability against cost.

Test Your Understanding

12. a) What is the advantage of a full-mesh leased line network? b) What is the disadvantage? c) What is the advantage of a pure hub-and-spoke leased line network? d) What is the disadvantage? e) Do most leased line networks use a full-mesh or a pure hub-and-spoke topology? Explain.
13. A company has three sites: Micah, Mallory, and William. Micah and Mallory need 100 Mbps of transmission capacity between them. Mallory and William need 200 Mbps of transmission capacity between them. Micah and William need 300 Mbps of transmission capacity between them. a) Create a hub-and-spoke network with Micah as the hub. What links will there be, and how fast will they need to be? Explain your reasoning. b) For the same situation, create a full-mesh network. What speeds will the links need to have if you are not concerned with redundancy in case of line failure? Explain your reasoning. c) Building on the last part of this question, add redundancy so that a failure of the line between Mallory and William will not bring down the network. (Hint: DTP.)

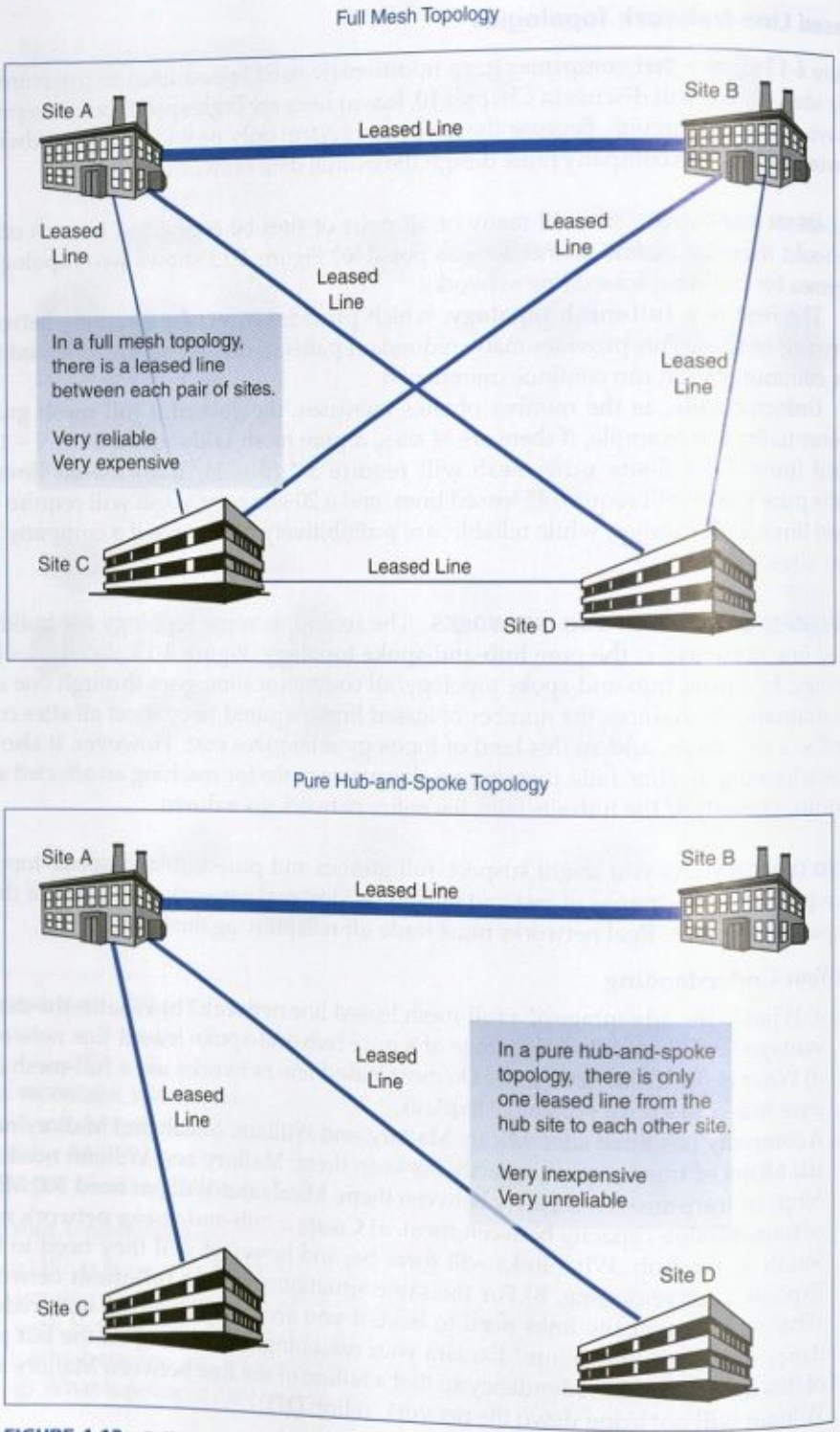

FIGURE 4-13 Full-Mesh and Hub-and-Spoke Topologies for Leased Line Data Networks

Handling Momentary Traffic Peaks

One fact of networking life that can never be ignored in designs is that traffic volume varies widely. Peak periods of traffic can overwhelm the network's switches and transmission lines. Given the statistical nature and high variability of traffic, **momentary traffic peaks** lasting a fraction of a second to a few seconds are bound to occur, and a firm must have a plan for managing momentary traffic peaks. Corporations can use several traditional traffic management methods to respond to momentary traffic peaks, as Figure 4-14 shows. Network planners must select which approach to use in the corporate network or in different parts of the corporate network.

OVERPROVISIONING One approach is overprovisioning—adding much more switching and transmission line capacity than will be needed most of the time. With overprovisioning, it will be very rare for momentary traffic peaks to exceed capacity. This means that no regular ongoing management is required. The downside of overprovisioning, of course, is that it is wasteful of capacity. Today, the simplicity of overprovisioning and the relatively low cost of overprovisioning on LANs make overprovisioning attractive

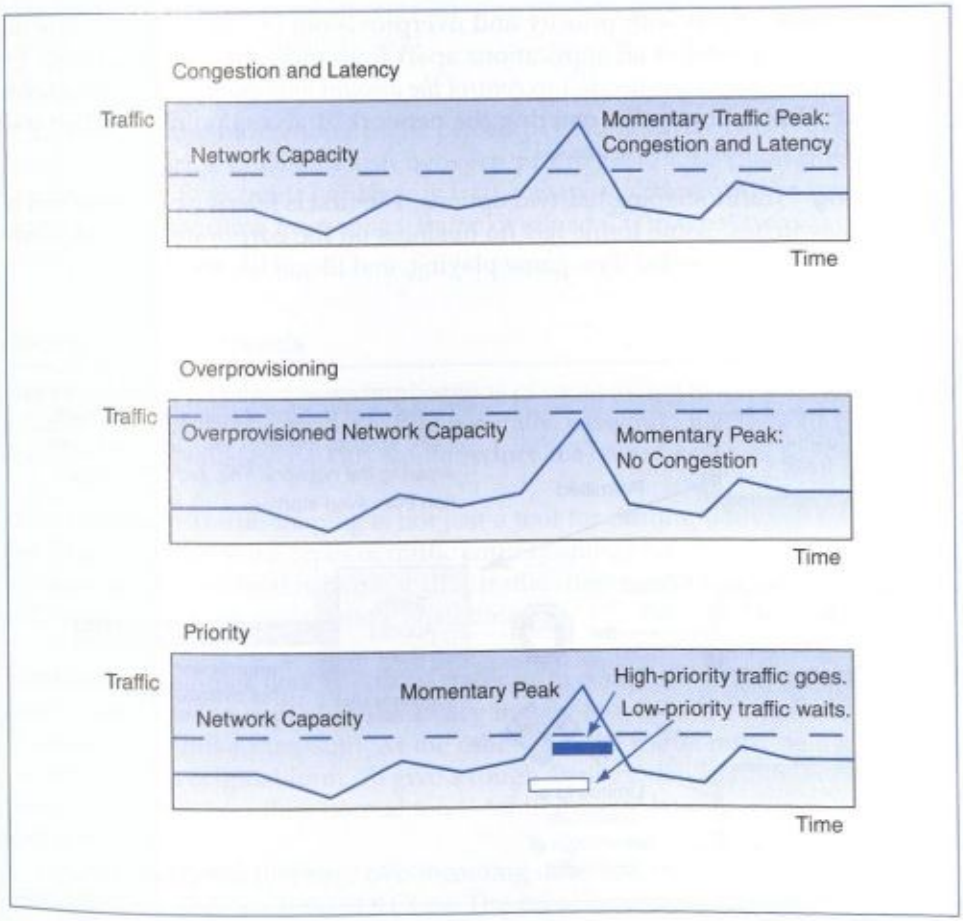

FIGURE 4-14 Handling Momentary Traffic Peaks

on LANs. On WANs, however, where the cost per bit transmitted is very high, overprovisioning is too expensive to consider.

PRIORITY Priority, in turn, assigns high priority to latency-intolerant applications, such as voice, while giving low priority to latency-tolerant applications, such as e-mail. Whenever congestion occurs, high-priority traffic is sent through without delay. Low-priority traffic must wait until the momentary congestion clears. Priority allows the company to work with lower capacity than overprovisioning but requires more management labor.

QoS GUARANTEES Quality-of-service guarantees take a step beyond priority, reserving capacity on each switch and transmission line for certain types of traffic. This allows the firm to satisfy QoS service level agreements for selected traffic by providing guarantees for minimum throughput, maximum latency, and even maximum jitter.

QoS guarantees require extremely active management. In addition, traffic with no QoS guarantees gets only whatever capacity is left over after reservations. This may be too little, even for latency-tolerant traffic.

TRAFFIC SHAPING Even with priority and overprovisioning, sufficient capacity must be provided for the total of all applications apart from momentary traffic peaks. Even more active management is needed to *control the amount of traffic entering the network in the first place*. Restricting traffic entering the network at access points is called **traffic shaping**, which is shown in Figure 4-15.

Filtering Traffic shaping has two options. The first is *filtering* out unwanted traffic at access switches. Some traffic has no business on the corporate network, such as downloading MP3 and video files, game playing, and illegal file sharing.

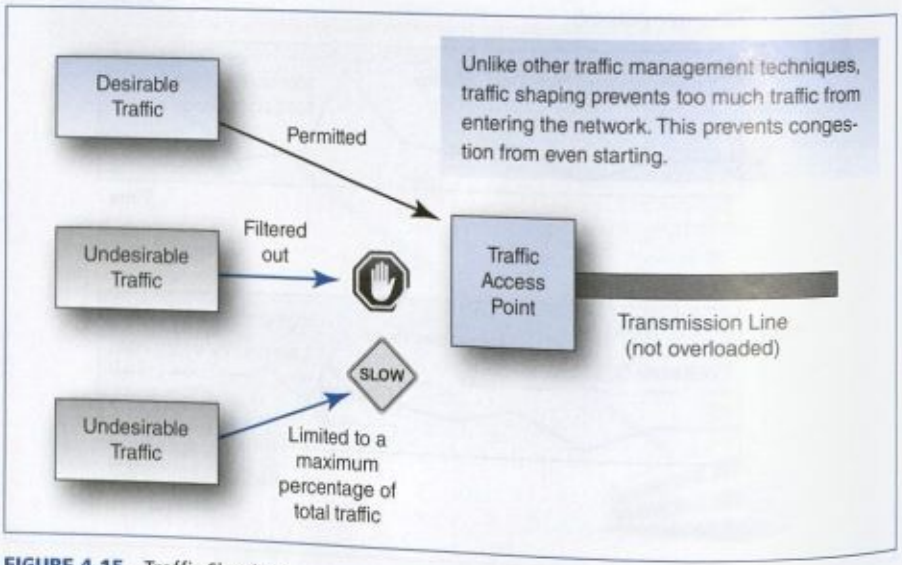

FIGURE 4-15 Traffic Shaping

Capacity Percentages The second option in traffic shaping is to assign certain percentages of capacity to certain applications arriving at access switches. Even if file sharing has legitimate uses within a firm, for instance, the firm may wish to restrict the amount of capacity that file sharing can use. Typically, each application or application category is given a maximum percentage of the network's capacity. If that application attempts to use more than its share of capacity, incoming frames containing the application messages will be rejected.

Perspective on Traffic Shaping Overprovisioning, priority, and QoS guarantees merely attempt to deal with incoming traffic. Traffic shaping actually reduces the amount of incoming traffic. Only traffic shaping can dramatically reduce network cost.

Although traffic shaping is very economical in terms of transmission capacity, it is highly labor intensive. It is used today primarily on high-cost WAN links. However, as management software costs fall in price and require less labor to operate, traffic shaping should see increasing use.

Another issue that arises when traffic shaping is used is politics. Telling a department that its traffic will be filtered out or limited in volume is not a good way to make friends. Priority and QoS reservations also raise political problems, but in traffic shaping, these problems are particularly bad.

Test Your Understanding

14. a) How long are momentary traffic peaks? b) Distinguish between overprovisioning and priority. c) Distinguish between priority and QoS guarantees. d) What problem can QoS create? e) How is traffic shaping different from traditional approaches to handling momentary traffic overloads? f) In what two ways can traffic shaping reduce traffic?

Reducing Capacity Needs

We have just looked at the design implications of momentary traffic peaks, which usually last only a fraction of a second. More generally, we would like to reduce the overall traffic the network must carry. This would reduce the cost of network services directly.

TRAFFIC SHAPING Traffic shaping is not just a tool for dealing with momentary traffic peaks. By eliminating some types of traffic entirely and by limiting other types of traffic to small percentages of total network traffic, traffic shaping can substantially reduce the overall traffic a network must handle at all times.

COMPRESSION Another way to reduce traffic is to compress traffic before it enters a network. **Compression** exploits redundancy in data to recode the data into fewer bits. This means fewer bits to transmit. At the other end, the traffic must be decompressed to put it back in its original form. To give a rough analogy, dehydrated food with water removed is much lighter than normal food. Adding water later reconstitutes the dehydrated food.

Figure 4-16 shows there are two incoming data streams. The first is 3 Gbps. The second is 5 Gbps. This is a total of 8 Gbps. The capacity of the transmission line is only 1 Gbps. We have a problem.

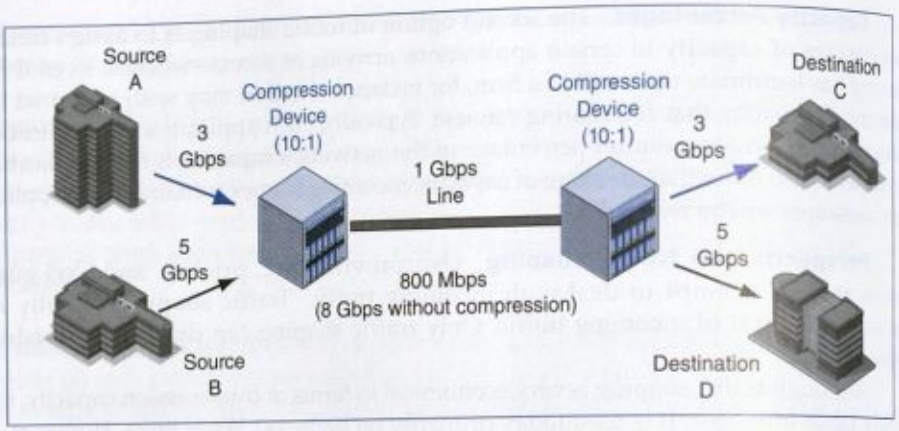

FIGURE 4-16 Compression

Before the incoming data enters the transmission line, however, the data will be compressed by 10:1, which is often possible for typical data streams. This reduces the compressed data streams to 0.3 Gbps (300 Mbps) and 0.5 Gbps (500 Mbps), for a total of only 800 Mbps. This traffic can easily fit on a 1 Gbps line with ample room for other traffic.

At the other end, another device decompresses the data streams. It sends the 3 Gbps data stream to one destination and the 5 Gbps data stream to another destination.

One requirement for compression is that you must have compatible equipment at the two ends of the network. Given a frequent lack of vendor compatibility, compression tends to lock the company into a single vendor's products.

Test Your Understanding

15. a) Why is traffic shaping a more general tool than just being a way to handle momentary traffic peaks? b) Why can compression help in traffic management? c) What makes compression possible?

Natural Designs

We have been discussing general design principles. In many cases, however, designers must choose designs that are natural for their environments. For example, Figure 4-17 shows a natural design for a building LAN. This building has multiple floors. It will simply make everyone's life easier if each floor is given an Ethernet workgroup switch that serves the hosts and wireless access points on that floor. It is also natural to place a core switch in the basement and have all communication between switches go through the core switch. The core switch can then connect to a router that acts as a gateway to the outside world.

Test Your Understanding

16. Why was the design in Figure 4-17 selected?

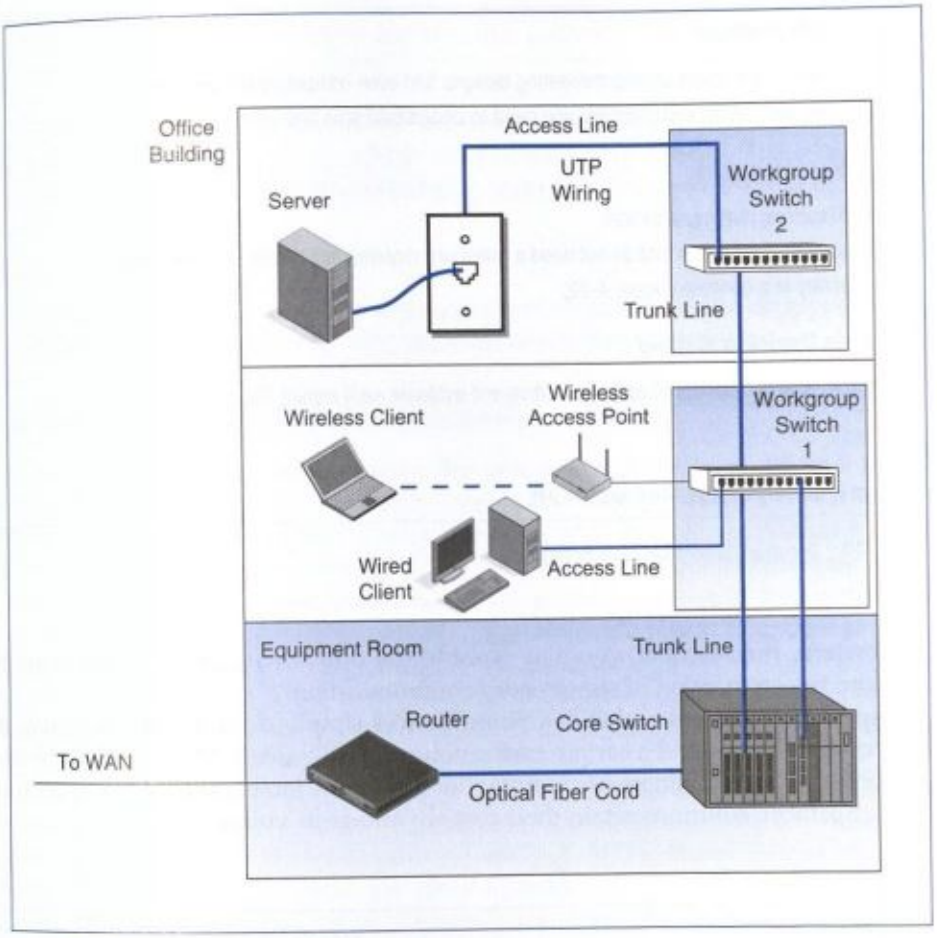

FIGURE 4-17 Natural Design for a Building LAN

EVALUATING ALTERNATIVES

When a design is completed, it is usually necessary to select between products offered by different vendors and perhaps between competing technologies. In much of this book, we will see descriptions of multiple technologies with a special focus on relative strengths and weaknesses. It is not enough to know the individual technologies. You must be able to discuss the pros and cons of competing technologies.

Minimum Requirements

Sometimes there are **minimum requirements** that will exclude a certain product or technology from final consideration. For example, if you need an e-mail server that will support at least 10,000 users in your company, you cannot consider an e-mail server product that will support only 2,000 users—even if it is very inexpensive. If security is

Comparing Alternatives

In designs, must select among competing designs and even competing technologies
When learning about technologies, you need to understand pros and cons

Minimum Requirements

Specifications that must be met
Noncompliant products that do not need a minimum requirement cannot be considered
Scalability is a concern (Figure 4-22)

Multicriteria Decision Making

Must look at all aspects of each alternative and evaluate each aspect (Figure 4-23)

Cost

Cost is difficult to measure (Figure 4-24)

FIGURE 4-18 Product Selection (Study Figure)

a high concern, furthermore, you may want to use only a wired technology or routers that require the encryption of supervisory communication.

A special concern is scalability. Some choices simply do not scale, meaning that they are not useful beyond a certain traffic volume. As Figure 4-19 shows, a technology may be cost effective when its use is small but may grow too expensive at higher traffic volumes. **Scalable** solutions retain their cost advantage as volume grows.

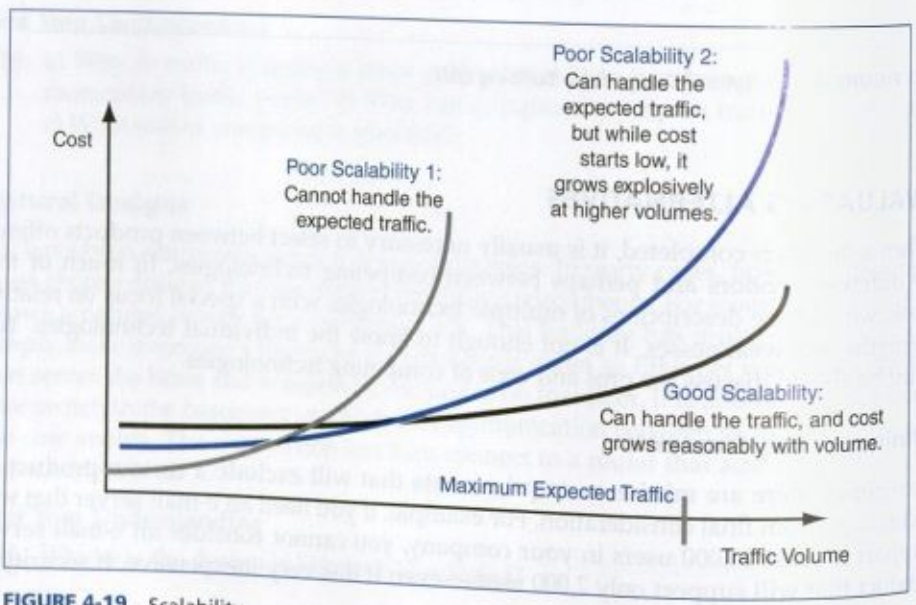

FIGURE 4-19 Scalability

Scalable solutions grow slowly in cost as traffic volume increases.

Another scalability problem is a complete inability to grow enough to meet a company's traffic volume regardless of how much a company spends. The example of the e-mail server that does not meet a minimum requirement is an example of a failure to scale enough.

Test Your Understanding

17. a) Should products that fail to meet minimum requirements be dropped from consideration? b) In what two ways can solutions fail to be scalable?

Product Selection with Multicriteria Decision Making

Once a project is selected and initiated, the network staff must go through the traditional systems development life cycle to implement the project. Given that almost all readers know about the systems development life cycle, we will not discuss it in detail.

In software development projects, there usually is a **make-versus-buy decision**. Should the programming staff create the software itself, or should the company purchase the software? In networking projects, this first option rarely makes sense. User companies like banks and retail stores do not have the technical expertise to make their own switches and routers. Instead, they must *select* and *buy* these technologies. Consequently, in this book, we will look at the factors you need to understand in purchasing decisions that involve different technologies.

When making purchasing decisions, companies tend to use **multicriteria decision making**, which Figure 4-20 illustrates. In this approach, the company decides what product characteristics will be important in making the purchase. Things that are important in the purchasing decision are called **criteria**.

Of course, costs are important, both purchase costs and ongoing costs. However, other decision criteria are also important. In Figure 4-20, the criteria for the product are functionality, availability, cost, ease of management, and electrical efficiency.

		Product A		Product B	
Criterion	Criterion Weight (Max: 5)	Criterion Rating (Max: 10)	Criterion Score	Product Rating (Max: 10)	Criterion Score
Functionality	5	9	45	7	35
Availability	2	7	14	7	14
Cost	5	4	20	9	45
Ease of Management	4	8	32	6	24
Electrical Efficiency	1	9	9	8	8
Total Score			120		126

FIGURE 4-20 Multicriteria Decision Making in Purchase Decisions

Next to each criterion is the **criterion weight**. This weight gives the relative importance of each criterion compared to other criteria. Here, weights range from 1 to 5. Note that cost and functionality have the largest weights (5), emphasizing their importance.

For each product (there are only two in the figure), the evaluation team gives a **rating** for each decision criterion. In this example, the ratings range from 1 to 10, with higher values indicating higher value. More functionality is better, so higher numbers in ratings reflect greater functionality. In contrast, for cost, lower cost is better, so higher rated values must indicate lower cost.

After filling in the ratings on all criteria for all products, the network staff computes the **criterion score** for each product. To do this, the staff multiplies the criterion weight by the rating for that product in that criterion. It then totals the criterion scores into a **total score**.

In Figure 4-20, Product A has a total score of 120, while Product B has a total score of 126. Speaking simplistically, Product B appears to be a better choice. However, the two total scores are very close. Numbers should never drive out thinking. A closer look shows that Product A has very good functionality and ease of management, although its cost is high. Product B has poorer scores on functionality and ease of management. It may be possible to negotiate a lower price on Product A and redo the analysis.

Test Your Understanding

18. a) What is the make-versus-buy decision? b) For routers and switches, do firms usually make or buy? c) We are considering products A, B, and C. Our criteria are price, performance, and reliability with weights of 20 percent, 40 percent, and 40 percent, respectively. Product A's evaluation scores on these three criteria are 8, 6, and 6, respectively. For B, the values are 6, 8, and 8 , respectively. For C, they are 7, 7, and 7 , respectively. Present a multicriteria analysis of the decision problem in tabular form and showing all work. Interpret the table.

Cost Is Difficult to Measure

Systems Development Life Cycle Costs

- Hardware: Full price: base price and necessary components
- Software: Full price: base price and necessary components
- Labor costs: Networking staff and user costs
- Outsourcing development costs
- Total development investment

Systems Life Cycle (SLC) Costs

- System development life cycle versus system life cycle
- Total cost of ownership (TCO)
  - Total cost over entire life cycle
  - SDLC costs plus carrier costs
    - Carrier pricing is complex and difficult to analyze
    - Must deal with leases

FIGURE 4-21 Cost (Study Figure)

Cost

Although cost is only one factor in product selection, it is often a critical factor. If you are a mobile phone user, you know how difficult it is to figure your total cost for a month. Figuring the cost of a network alternative is equally complex.

SYSTEMS DEVELOPMENT LIFE CYCLE (SDLC) COSTS When you begin a project, you need to consider the cost of the system during its development life cycle.

Hardware Costs Consider what happens when you buy a personal computer. You first have to take hardware into account. When you look at the price of a computer, this may not include a display, and it usually does not include a printer. The **base price** is the price before adding components that will be needed in actual practice. In contrast, the **full price** of the hardware is the price of a complete working system. The distinction between base price and full price is also applicable to network hardware, including the switches and routers we will see later in this term.

Software Costs After your first computer purchase, you realized that the software can be almost as expensive as the hardware. You have to consider the software you will need very carefully and understand the cost of that software. Individual software products, furthermore, often have misleading base prices that do not include all necessary components. Network product software decisions are similarly complex.

Labor Costs in Development Although hardware and software costs are complex and difficult to measure, these problems pale beside the problems involved in estimating labor costs in development. Planning, procurement, installation, configuration, testing, programming, and other labor costs can easily exceed hardware and software costs.

User costs should also be considered as an aspect of labor cost because the time that users spend on the system's development during requirements definition and later development states is substantial. This time is not free to the company, any more than network staff time is free.

Outsourcing Development Costs If the company outsources some or all of the development costs, then outsourcing costs need to be considered in the overall picture.

Total Development Investment To evaluate potential projects, the networking staff must forecast the total development investment—the total of hardware, software, labor, and outsourcing costs during development. These expenditures truly are investments that should pay off over the life of the project.

SYSTEMS LIFE CYCLE (SLC) COSTS As noted earlier in this chapter, the systems development life cycle is only part of the overall systems life cycle, which lasts from conception to termination. It is important to consider **systems life cycle costs**, which are costs over a system's entire life, not just during the systems development period. The cost of a system over its entire life cycle is called the **total cost of ownership (TCO)**.

Operating and management costs usually are very important over the system's life cycle. When making equipment and software purchases, it is important to consider how much labor is involved in operating and managing the equipment and software. These costs must be considered very carefully in product selection.

In particular, a new leading-edge technology may give fantastic performance, but leading-edge technologies tend to be immature and tend to create far higher support costs than established technologies.

One new factor in systems life cycle analysis is carrier costs. If you must deal with a communications carrier to carry your signals from one corporate site to another, then you also have to consider carrier pricing. This is rarely easy to do, and it is even harder to compare the prices of alternative carriers offering roughly the same service because of the wording in their contracts. In addition, you usually have to sign equipment leases or service agreements that lock you in for various periods of time, sometimes up to several years.

Test Your Understanding

19. a) What period of a network's life does the SDLC cover? b) Why are hardware and software base prices often misleading? c) List the four categories of SDLC costs. d) Why must user costs in development be considered?
20. a) Distinguish between the systems development life cycle and the systems life cycle. b) What is the total cost of ownership (TCO)? c) Why should operating and management costs be considered in addition to hardware, software, and transmission costs in purchasing decisions? d) What additional cost factor comes into SLC costs, compared to SDLC costs?

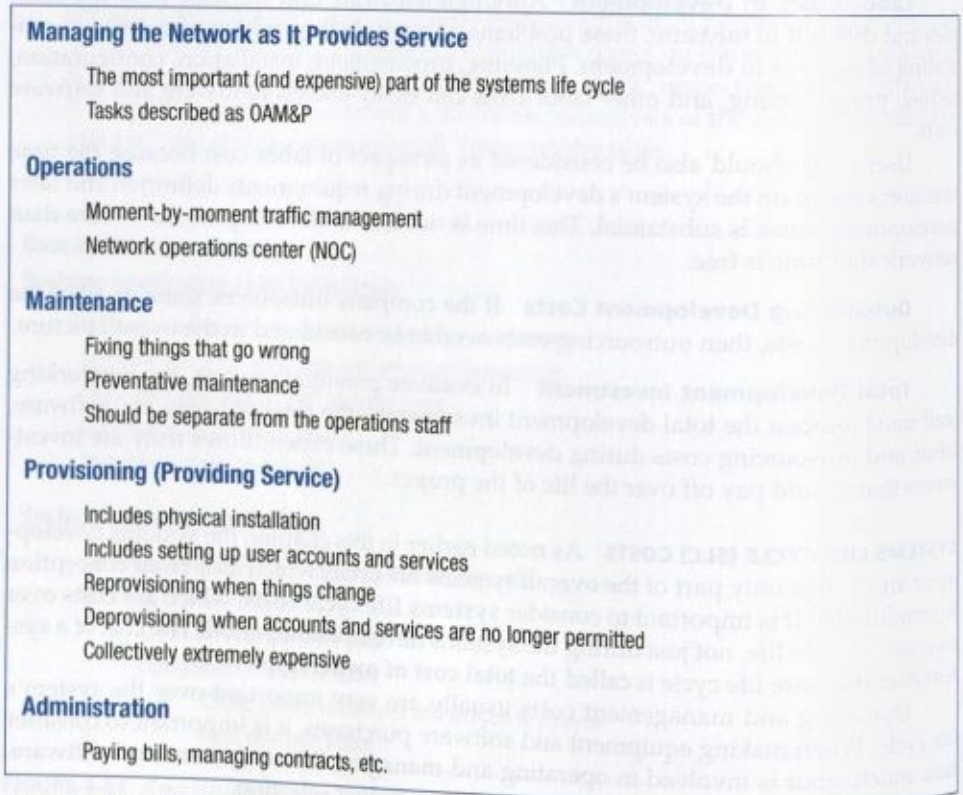

FIGURE 4-22 Ongoing Management (OAM&P) (Study Figure)

OPERATIONAL MANAGEMENT

In the networking systems life cycle, a great deal of the work takes place after the development finishes. This requires networking professionals to be able to do operational management as well as development.

OAM&P

After a network component is in place, it probably will be used for many years. During its **operational** life (its life after development), there will be substantial labor costs. We will classify these costs in a way that telecommunications carriers have traditionally done—in terms of **operations, administration, maintenance, and provisioning (OAM&P)**

OPERATIONS You probably have seen pictures of **network operations centers (NOCs)** for major telecommunications carriers. These are large rooms with dozens of monitors showing the conditions of various parts of the network. Most corporations also have network operations centers. These corporate NOCs are smaller, usually having only about a half dozen monitors. NOCs manage the network on a moment-by-moment basis.

MAINTENANCE You have undoubtedly seen telephone company maintenance trucks driving on their way to downed transmission lines, broken transformers, or other trouble spots. In addition to fixing equipment failures, telephone companies do preventative maintenance to prevent future failures.

In the same way, companies often have to fix their internal corporate network switches and other physical components. They also have to handle software problems. Although the network operations center can fix some problems remotely, most firms have separate NOC and maintenance staffs. The NOC staff usually is heavily occupied with the moment-by-moment operation of the network, so it makes sense to have other networking professionals focus on maintenance.

PROVISIONING If you get cable television service, the cable company has to **provision** your residence, that is, set up service. This includes physical setup (running the coaxial cable into your home). It also involves setting up your account on the company's computers. The cable company also has to reprovision customers when they change their service by adding channels, dropping optional services, or switching pricing plans.

Within a corporate network, provisioning may involve the installation of additional switches, routers, and transmission lines to serve new users. In networks, every time a new user joins the firm, the company has to provision service for that user. In fact, provisioning has to be done for every user account on every server and access point on the network.

Furthermore, once a user is provisioned for a particular resource, he or she may have to be **reprovisioned** if his or her authorizations change—say, if he or she is upgraded from read-only data access to full read/write access. The user also has to be reprovisioned if he or she changes jobs within a firm, joins project teams, or does many other things. Users also have to be **deprovisioned** when they leave project teams or leave the company entirely. Contractors and other outside organizations also have to be provisioned, reprovisioned, and deprovisioned when they start to work, change the way they work, or stop working with a company. Collectively, provisioning is extremely expensive.

ADMINISTRATION Operations, maintenance, and provisioning involve real-time work to keep the network running. In contrast, administrative work is dominated by such

mundane tasks as paying bills to vendors and telephone companies, managing proposals and contracts, doing network budgeting, comparing network budgets to actual costs, and doing other dull but necessary tasks.

Test Your Understanding

21. a) For what is OAM&P an abbreviation in ongoing management? b) Distinguish between operations and maintenance. c) What is provisioning? d) When may reprovisioning be necessary? e) When may deprovisioning be necessary? f) Into which of the four categories would you classify the task of comparing the inventory of parts with the inventory list on the computer?

Network Management Software

Given the complexity of networks, network managers need to turn to **network management software** to support much of their work. Many of these are **network visibility** tools, which help managers comprehend what is going on in their networks.

PING The oldest network visibility tool is the basic ping command available in all operating systems. If a network is having problems, a network administrator can simply ping a wide range of IP addresses in the company. By analyzing which hosts and routers respond or do not respond, then drawing the unreachable devices on a map, the administrator is likely to be able to see a pattern that indicates the root cause of the problem. Of course, manually pinging a wide range of IP addresses could take a prohibitive amount of time. Fortunately, there are many programs that ping a range of IP addresses and portray the results.

THE SIMPLE NETWORK MANAGEMENT PROTOCOL (SNMP) Ping can tell you if a host is available. It can also tell you the latency in reaching that host. For remote device management, most network operation centers use more powerful network visualization products based on the **simple network management protocol (SNMP)**, which is illustrated in Figure 4-23. In the NOC, there is a computer that runs a program called the **manager**. This manager manages a large number of **managed devices**, such as switches, routers, servers, and PCs.

Actually, the manager does not talk directly with the managed devices. Rather, each managed device has an **agent**, which is hardware, software, or both. The manager talks to the agent, which in response talks to the managed device. To give an analogy, recording stars have agents who negotiate contracts with studios and performance events. Agents provide a similar service for devices.

The network operations center constantly collects data from the managed devices using SNMP **Get** commands. It places these data in a **management information base (MIB)**. Data in the MIB allows the NOC managers to understand the traffic flowing through the network. This can include failure points, links that are approaching their capacity, or unusual traffic patterns that may indicate attacks on the network.

In addition, the manager can send **Set** commands to the switches and other devices within the network. Set commands can reroute traffic around failed equipment or transmission links, reroute traffic around points of congestion, or turn off expensive transmission links during periods when less expensive links can carry the traffic adequately.

Normally, the manager sends a command and the agent responds. However, if the agent senses a problem, it can send a **trap** command on its own initiative. The trap

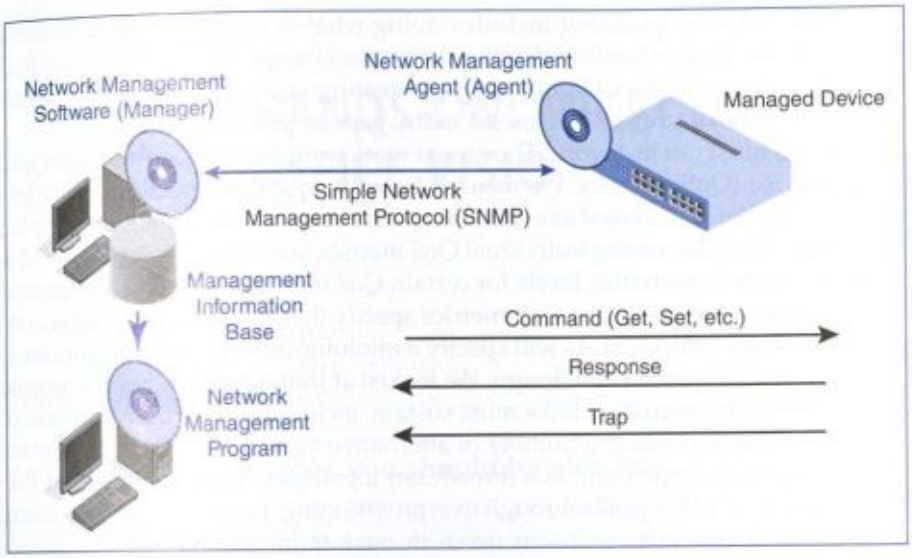

FIGURE 4-23 Network Management Software

command gives details of the problem. There is one more program in the figure—a **network visualization program**. This program takes results from the MIB and interprets the data to display results in maps, finds root causes for problems, and does other tasks. Note that this functionality is *not* included in the simple network management protocol. SNMP simply collects the data in a way that network visualization programs can use. This lack of specification allows network visualization program vendors to innovate without being constrained by standards. What do network visualization programs do?

AUTOMATION Many other network management chores can be automated to reduce the amount of work that network managers need to spend on minutia. For example, many routers are given a standard corporate configuration when they are installed. Doing this manually can take an hour or more per router. However, it may be possible to create a standard configuration, store it, and simply download it onto new routers. In addition, if corporate standard configurations change or a patch must be installed on all routers, it may be possible simply to "push out" these changes to all routers.

Test Your Understanding

22. a) List the main elements in SNMP. b) Does the manager communicate directly with the managed device? Explain. c) Distinguish between Get and Set commands. d) Where does the manager store the information it receives from Get commands? e) What kinds of messages can agents initiate?

CONCLUSION

Synopsis

This is the last of four introductory chapters. This chapter looked at network management. It began with a discussion of basic concepts, including the need to focus on the systems life cycle, the need to be efficient, and the strategic network planning process.

Strategic network planning includes doing what-is analysis (understanding the current situation), understanding driving forces for change, identifying gaps that will appear if the current system is not changed, creating strategies for closing gaps, and selecting a portfolio of projects to close as many gaps as possible.

Networks must not just work. They must work well. Networks must meet goals for quality-of-service (QoS) metrics. We focused heavily on speed, including how to write speeds properly. We also looked at availability, error rates, latency, jitter, and application response time. After discussing individual QoS metrics, we looked at service level agreements (SLAs), which guarantee levels for certain QoS metrics for a certain percentage of time. Many find it confusing that QoS metrics specify that service will be no worse than certain values. For example, SLAs will specify a minimum speed, not a maximum speed.

There was a long section on design. We looked at traffic analysis, which assesses the traffic that various transmission links must sustain, including redundancy in case of link failures. We looked in detail at a number of alternative topologies, including hierarchical topologies, mesh topologies, and bus (broadcast) topologies. We also looked at the handling of momentary traffic peaks through overprovisioning, priority, and QoS guarantees. Finally, we looked at reducing capacity needs through traffic shaping and compression.

When considering alternatives, it is important to evaluate them systematically. We discussed how to do this with multicriteria decision making. We looked in some detail at estimating costs realistically.

The chapter ended with a discussion of operational management. We looked at the four main traditional elements of network management: operations, administration, maintenance, and provisioning. We looked at the simple network management protocol (SNMP) and saw how important it is in collecting information needed for management and even for doing remote changes on network devices.

END-OF-CHAPTER QUESTIONS

Thought Questions

1. Assume that an average SNMP response message is 100 bytes long. Assume that a manager sends 40 SNMP Get commands each second. a) What percentage of a 100 Mbps LAN link's capacity would the resulting response traffic represent? b) What percentage of a 128 kbps WAN link would the response messages represent? c) What can you conclude from your answers to this question?

2. The telephone network has long boasted that it has the "five nines" (99.999 percent availability). a) How much downtime is this per year? Express downtime in days, hours, minutes, and so on as appropriate. b) How much downtime is there per year with 99 percent availability?

Perspective Questions

1. What was the most surprising thing you learned in this chapter?

2. What was the most difficult part of this chapter for you?

4a HANDS-ON: MICROSOFT OFFICE VISIO

LEARNING OBJECTIVES

By the end of this chapter, you should be able to:

- Create a simple Visio diagram.

WHAT IS VISIO?

Microsoft Office Visio is a drawing program. The professional version has special symbols for drawing network diagrams. Visio is widely used by network professionals to visualize networks they are designing.

USING VISIO

Visio is part of the Microsoft Office family. Installing Visio is like installing any other Office product.

Figure 4a-1 shows how to start a Visio drawing. Of course, this begins by selecting File and then New. In the figure, Network has been selected for the type of drawing. Basic Network Diagram has been selected.

As Figure 4a-2 shows, this brings up a window with a canvas on which you can drag shapes. In the figure, the shape of a generic server has been dragged on to the screen. As you can see, many other network diagramming shapes can be dragged onto the screen.

After you have added the devices you need, it is time to begin showing how they are connected. As Figure 4a-3 shows, there is a connector icon at the top of the screen. Select the connector tool. Then drag between the two icons to connect them. After you have connected them, try dragging one of the connected devices. You will see that the connectors move with them.

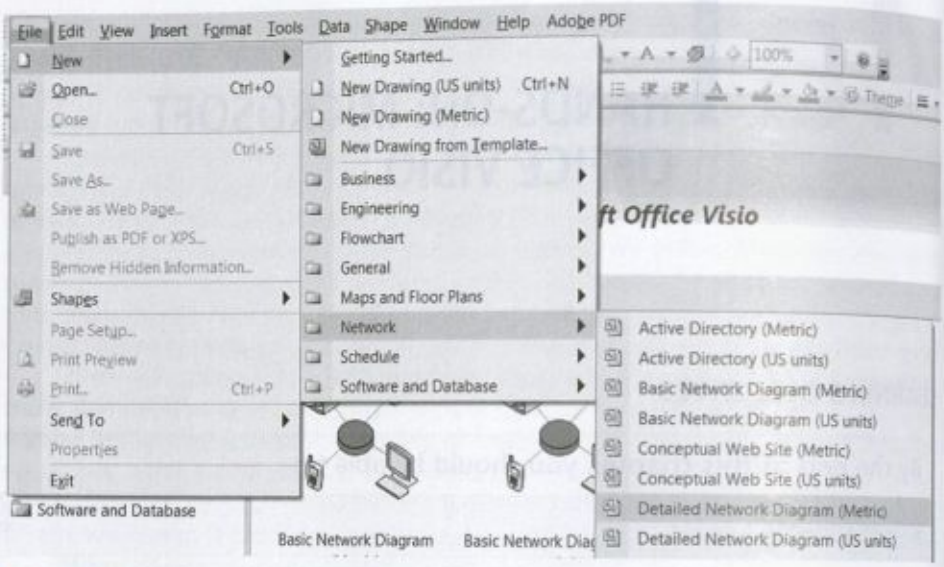

FIGURE 4a-1 Starting a Visio Drawing

Source: Screenshot © 2012 Microsoft Corporation. Used with permission from Microsoft.

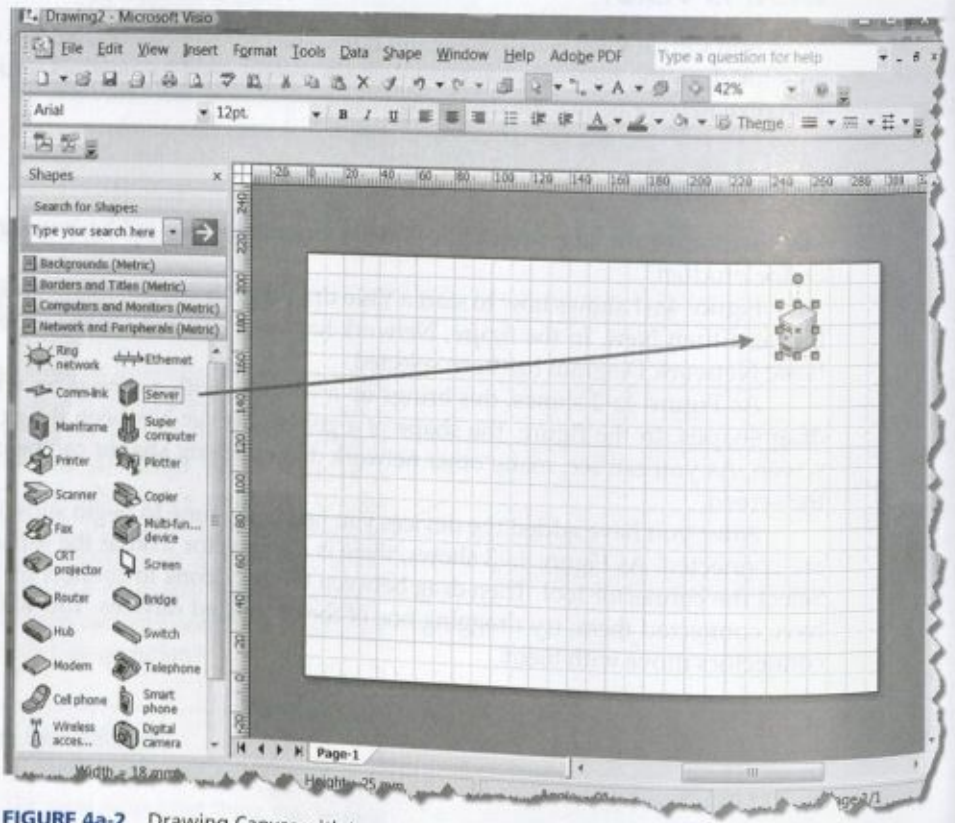

FIGURE 4a-2 Drawing Canvas with Icon Being Dragged

Source: Screenshot © 2012 Microsoft Corporation. Used with permission from Microsoft.

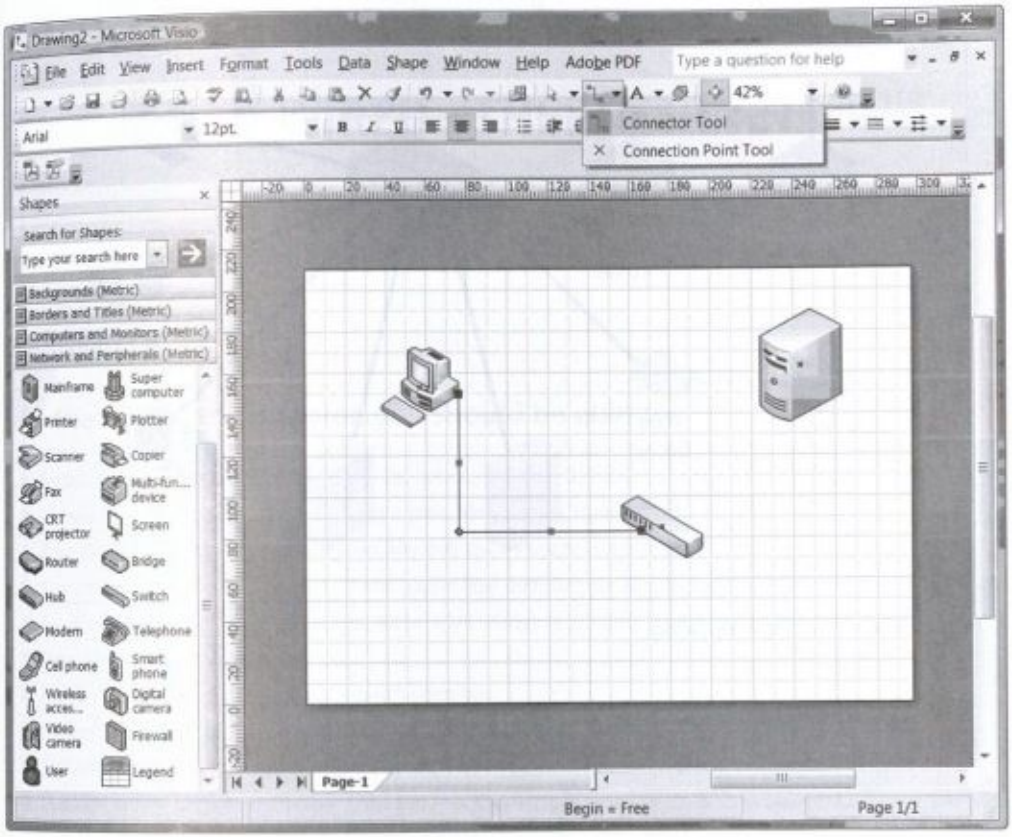

FIGURE 4a-3 Adding Connections

Source: Screenshot © 2012 Microsoft Corporation. Used with permission from Microsoft.

Not shown on the figure, you can double-click on an icon. This adds text below the icon. Visio is not fussy about preventing lines from overlapping text. Overall, Visio diagrams are easy to create but not extremely pretty.

Exercise

In Microsoft Office Visio, create something like the drawing in Figure 4a-4. The drawing has a print server. A print server is a device that allows several users in an office to share a printer. A print server plugs into a printer via a USB port. It also plugs into a switch via a UTP cord.

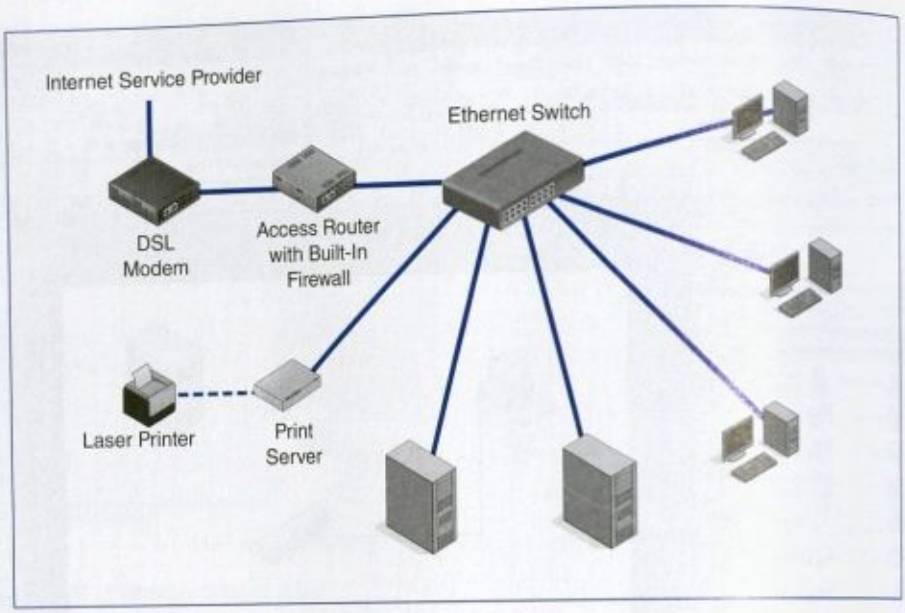

FIGURE 4a-4 Sample Drawing

5 WIRED ETHERNET LANs

LEARNING OBJECTIVES

By the end of this chapter, you should be able to:

- Explain the service and economic differences between LANs and WANs.
- Discuss the concept of Ethernet LANs and explain how they are standardized.
- Describe digital and binary signaling and why they reduce transmission errors.
- Explain the technologies of 4-pair UTP and optical fiber and compare and contrast their relative strengths and weaknesses.
- Be able to design a physical network based on a knowledge of Ethernet standards, including link aggregation.
- Describe the Ethernet frame in detail.
- Explain basic Ethernet data link layer switch operation.
- Explain why the Rapid Spanning Tree Protocol is necessary and how it functions.
- Explain virtual LANs, priority, manageability, and POE in Ethernet LANs.
- Describe security threats to Ethernet LANs and how they are addressed by the 802.1X standard.

INTRODUCTION

In the 19th century, scientists believed that radio waves traveled through an unseen transmission medium called the ether. In the early part of the 20th century, scientists disproved the theory. In the 1970s, a young researcher named Robert Metcalf visited the University of Hawai'i and saw the packet radio experiment being conducted there. Afterward, he designed a wired network technology, which he playfully called **Ethernet**. Today, few are aware of the irony in the name. *Ethernet* has become so common a term that almost nobody thinks about it. Ethernet is simply the way that wired LANs work today. And yes, the fact that it works with wires instead of radio is another piece of irony.

LANs and WANs

ON AND OFF THE CUSTOMER PREMISES There is a fundamental distinction in networking between local area networks and wide area networks. Figure 5-1 compares the two. Most fundamentally, **local area networks (LANs)** operate entirely on the **customer premises**—an office suite, office building, or other property owned by the corporation. A LAN in a corporate headquarters building may connect

Characteristic	Local Area Network (LAN)	Wide Area Network (WAN)
Location	Located entirely on customer's premises	Must carry transmissions beyond customer's premises
Consequence of Location	Owning company operates the LAN	User must contract with a carrier that has rights of way to carry wires between premises
Technology and Service Consequence of Corporate versus Carrier Ownership	Owner can use any technology and service options it wishes	Customer is limited to technologies and service options offered by available carriers
Labor Consequences of Corporate versus Carrier Ownership	Owner must do all operation and maintenance work	Operational and maintenance work is done by the carrier
Economics	Transmission distances are short, so the cost per bit carried is low	Transmission distances are long, so the cost per bit carried is high
Speed Consequences of Economics	Very high speeds are affordable	Customers are content with lower speeds
Design Consequences of Economics	Optimization of transmission capacity is not pressing	Optimization of transmission capacity is critical

FIGURE 5-1 LANs versus WANs

hundreds of hosts throughout the building. (A network you build to connect a few PCs and other devices in your home, apartment, or dorm room is also a LAN.) In this chapter, we will look at wired LANs. In Chapters 6 and 7, you will learn about wireless LANs.

In turn, **wide area networks (WANs)** operate outside the customer premises. While LANs exist within a company's site, WANs connect different sites within an organization or transmit data between organizations. A company might want to receive Internet by running its own wiring through the city to an Internet access point, but it does not have the legal right to lay wires outside its premises. (Imagine what your neighbors would say if you ran wires through their yards.) For transmission beyond its premises, a company must use a transmission **carrier** that has received government **rights of way** to lay wires in public locations or to send radio signals beyond the customer premises. We will look at WANs in Chapter 10, after we have looked at TCP/IP transmission in Chapters 8 and 9.

Local area networks (LANs) exist within a company's site, while wide area networks (WANs) connect different sites within an organization or transmit data between organizations.

SERVICE IMPLICATIONS The owner of a LAN can choose any technology and service options it wishes. In addition, it can implement the LAN any way it wishes. On the downside, it also must allocate labor to operate and maintain the network.

In contrast, as we will see in Chapter 10, there are only a few carriers in most communities. In addition, each carrier typically offers only a few technologies and service options. For international networking, countries vary widely in the WAN technologies and services their carriers offer. This makes international corporate network integration difficult. On the positive side, the carrier operates and maintains the WANs, freeing companies of the need to staff the network. Of course, the carrier charges its customers to pay for these services.

ECONOMICS Another fundamental difference between LANs and WANs stems from economics. You know that if you place a long-distance call, it will cost more than a local call. An international call will cost even more. As distance increases, the price of transmission increases. The cost per bit transmitted therefore is higher in WANs than in LANs.

You know from basic economics that as unit price increases, fewer units are demanded. Or, in normal English, when the price of an item increases, people buy less of it. Consequently, companies tend to purchase lower-speed WAN links than LAN links. Typically, LANs bring 100 Mbps to 1 Gbps of unshared capacity to each desktop. WAN speeds more typically range from 1 Mbps to about 50 Mbps, and these speeds are typically shared by multiple users.

Typically, LANs bring 100 Mbps to 1 Gbps to each desktop. WAN speeds more typically vary from 1 Mbps to about 50 Mbps, and these speeds are typically shared by multiple users.

As a consequence of the higher cost of WAN transmission, companies spend more time optimizing their expensive WAN traffic than their relatively inexpensive LAN traffic. For example, companies may be somewhat tolerant of looking at YouTube videos on LANs, but they almost always clamp down on this type of information on their WAN links. We will see other ways in which companies work to lower their WAN costs in Chapter 10.

Test Your Understanding

1. a) Distinguish between LANs and WANs. b) What are rights of way? c) What are carriers? d) Why do you have more flexibility with LAN service than with WAN service? e) What is the advantage of using carriers?
2. Why are typical WAN speeds slower than typical LAN speeds? Give a clear and complete argument.

Ethernet

The dominant standard for wired LANs today is Ethernet. Although Ethernet was created in the 1970s, it did not become economical until the 1990s. Even then, there were other wired LAN technologies. Over time, however, Ethernet's adequate performance and superior economics completely won the market. Now that Ethernet is the focus of wired LAN development efforts, it continues to grow in speed and sophistication.

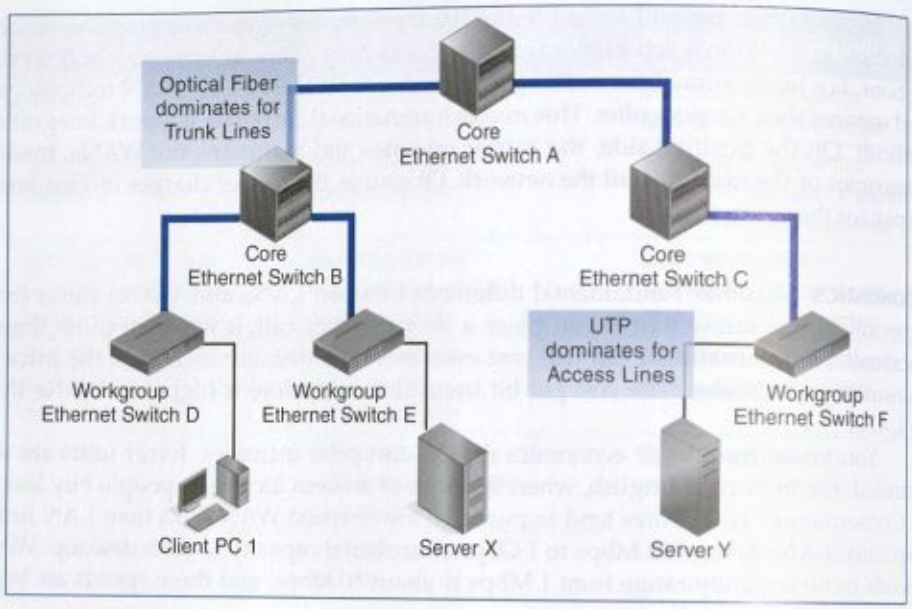

FIGURE 5-2 An Ethernet Network

Figure 5-2 illustrates a simple Ethernet network. At its heart is a collection of Ethernet switches. The switches that connect hosts to the network are called **workgroup switches**. Switches that connect switches to other switches are **core switches**.

Switches that connect hosts to the network are workgroup switches. Switches that connect switches to other switches are core switches.

To connect hosts to switches and switches to other switches, there is a transmission link. As we will see later in this chapter, there are two technologies for transmission links. One is unshielded twisted pair (UTP) copper wire, which we saw briefly in Chapter 1. The other is optical fiber. UTP carries electrical signals over copper wire pairs. Optical fiber carries light signals through very thin glass tubes. Transmission links that connect

FIGURE 5-3 Ethernet Workgroup Switch

Characteristic	Unshielded Twisted Pair	Optical Fiber
Medium	Copper wire	Glass
Signal	Electrical	Light
Maximum Distance in LANs	Usually 100 meters	Usually 200 to 500 meters
Speed	About the same	About the same
Cost	Lower	Higher

FIGURE 5-4 Optical Fiber and UTP

hosts to workgroup switches are called **access links**. Transmission links that connect switches to other switches are called **trunk links**.

Transmission links that connect hosts to workgroup switches are called access links. Transmission links that connect switches to other switches are called trunk links.

Figure 5-4 shows that the main benefit of optical fiber is distance span. While UTP in buildings is normally limited to 100 meters, optical fiber can span distances of 200 to 500 meters. Except at the very highest standardized speeds, which few corporations use, their speeds are comparable. Up to about 10 Gbps today, corporations can use either fiber or UTP. The penalty for greater distance span using fiber is greater cost. Fiber is more expensive to lay than UTP.

Switches in the network core must carry the frames of many conversations, so they must have high processing speeds. Workgroup switches, in contrast, only carry the conversations of the hosts they serve directly. They can operate much more slowly and still give adequate service. Slowness per se is not a virtue, but the low cost that comes with lower operating speeds is a definite virtue.

Test Your Understanding

3. a) Distinguish between the two types of Ethernet switches in terms of what they connect. b) Distinguish between the two types of Ethernet transmission link technologies in terms of what they connect. c) Why must core switches have more processing power than workgroup switches?

Ethernet Standards Development

Figure 5-5 shows that most LAN and MAN standards are developed by the 802 LAN/MAN standards committee of the IEEE Standards Association. A **MAN** is a **metropolitan area network**; it is a type of WAN limited to a large city and its surrounding communities. Distances are shorter for MANs than for national or international WANs. Constantly MAN prices per bit transmitted are lower than those of WANs with greater geographical scope. Consequently, MAN speeds are higher, although still less than LAN speeds.

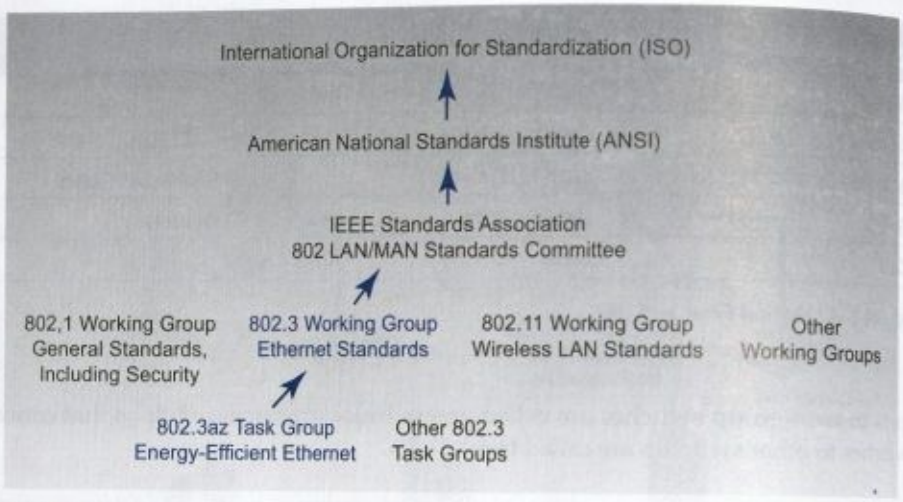

FIGURE 5-5 Ethernet Standards Development

The work of the 802 Committee is done in **working groups**. The 802.3 Working Group creates Ethernet standards, so the terms *Ethernet* and *802.3* are interchangeable. The 802.1 Working Group develops standards used in multiple working groups, for instance, security standards. We will look at the 802.1X security standard in this chapter. The 802.11 Working Group creates the wireless LAN standards, which we will see in Chapters 6 and 7.

The terms Ethernet and 802.3 are interchangeable.

More correctly, the 802.3 Working Group assigns subsets of its members to more specific **task groups**. For instance, the 802.3az Task Group was charged with creating standards for more energy-efficient Ethernet. After a task group completes its assigned work, it passes its results to the 802.3 Working Group, which accepts, rejects, or modifies the work of the task group.

After the 802.3 Working Group ratifies a new standard, it must pass through the 802 Committee and the IEEE Standards Association. From there, it must pass through the **American National Standards Institute (ANSI)**. Finally, because physical and data link layer standards are OSI standards, the standard must pass through ISO. In practice, however, as soon as a standard comes out of the 802.3 Working Group, vendors begin to build products using the standard. Other acceptances are automatic. Sometimes, vendors often begin building products based on draft standards. This sometimes leads to trouble later.

Test Your Understanding

4. a) What is a MAN? b) What standards association creates most LAN standards? c) What is the name of its committee for developing LAN standards? d) What 802 working group creates Ethernet standards? e) What working group is likely to develop security standards to be used by multiple LAN/WAN technologies? f) When do vendors begin to develop products based on new 802.3 standards?

Physical and Data Link Layer Operation

Ethernet networks are single switched networks. Single networks use standards at the physical layer and the data link layer. In the next section, we will look at Ethernet physical layer standards. In the following section, we will look at data link layer standards. After that, we will look at some advanced aspects of Ethernet operation, security, and management.

Single networks use standards at the physical layer and the data link layer.

On hosts, Ethernet physical and data link layer processes are handled in hardware. The circuitry that implements Ethernet is called a **network interface card (NIC)**. It received this name when it was a separate printed circuit board. Today, it is built into the computer's main printed circuit board. Perhaps we should call it the network interface circuit, but the old name has caught on.

Test Your Understanding

5. a) At what layers do single networks require standards? b) Is Ethernet processing executed in hardware or software? c) What circuit implements both of the physical and data link layer processes in Ethernet?

ETHERNET PHYSICAL LAYER STANDARDS

Physical layer standards govern connectors and transmission media. They also govern signaling. We will look at signaling first because it introduces concepts you will need when you look at UTP and optical fiber transmission media.

Test Your Understanding

6. What three things do physical layer standards govern?

Signaling

BITS AND SIGNALS A frame is a long series of 1s and 0s. To transmit the frame over a physical medium, the sender must convert the 1s and 0s into physical signals. These signals will **propagate** (travel) down the transmission link to the device at the other end of the physical link.

BINARY AND DIGITAL SIGNALING Figure 5-6 illustrates two popular types of signaling, binary and digital signaling. **Binary signaling** has two **states** (conditions), which may be two voltage levels or light being turned on or off. One state represents a 0. The other state represents a 1. In the figure, a 0 is represented as a high voltage, and a 1 is represented as a low voltage. In optical signaling, a 1 might represent light being turned on, while a 0 might represent light being turned off.

In binary signaling, there are *two* possible states. This makes sense because "bi" is Greek for two. The figure also shows **digital signaling**, in which there are a *few* states (2, 4, 8, etc.). How many "is few?" In some systems, there can be 32 or even 256 states, but the number of states is usually much lower. Also, because each state represents a

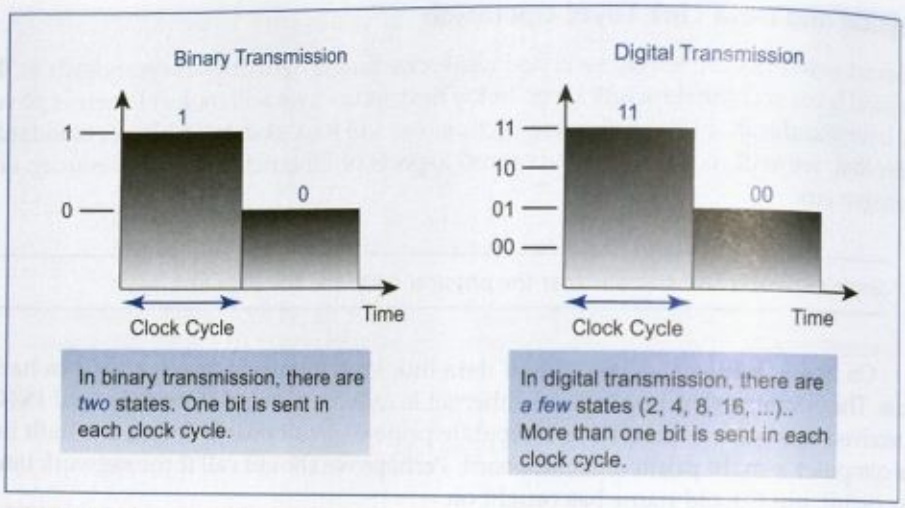

FIGURE 5-6 Binary and Digital Transmission

certain number of bits, the number of states is always a multiple of two—four, eight, sixteen, and so forth.

***Note:** In binary signaling, there are two possible states. In digital signaling, there are a few possible states (2, 4, 8, etc.).*

Having more than two states adds to the complexity and therefore the cost of signaling. However, Figure 5-6 shows that if you have multiple states, you can send multiple bits in a single clock cycle. With two states, you can only represent a single one or a zero. With four states, however, the lowest state might represent 00, the next lowest state might represent 01, the next 10, and the highest 11. With four states, then, you can send two bits at a time.

If "bi" means two, where does "digital" come from? It comes from the fact that we call our 10 fingers *digits*. In fact, some early computer systems operated on Base 10 arithmetic, the same arithmetic that we 10-fingered people use. Very quickly, however, the advantages of building computers and transmission systems that used two or a multiple of two states brought about binary and digital computation and later signaling.

We have talked about binary and digital transmission systems as if they were different. Actually, binary transmission is a subset of digital transmission. In binary transmission, *few* means two. Although binary transmission is the most common form of digital transmission and deserves its own name, all transmission in a typical network can properly be called digital.

Binary transmission is a type of digital signaling. Not all digital signaling, however, is binary signaling.

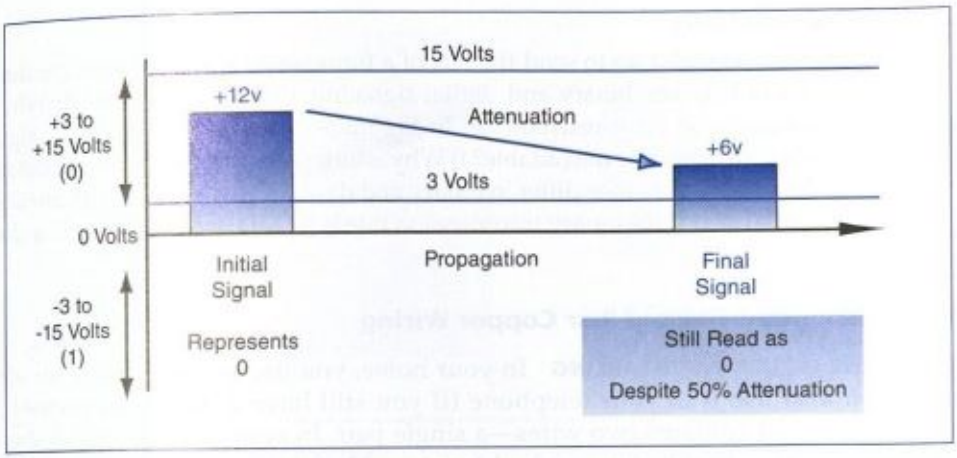

FIGURE 5-7 Error Resistance in Binary and Digital Signaling

ERROR RESISTANCE Why use digital transmission? The answer is that digital transmission is fairly resistant to transmission impairments. In Figure 5-7, a 0 is represented by a signal between 3 and 12 volts, while a 1 is represented by a signal between –3 and –15 volts. This is the signaling scheme used by the serial ports found on older computers. The signaling schemes on newer interfaces are too complex to describe simply.

Suppose that the sender transmits a 12 volt signal. This is clearly a 0. However, as the signal propagates, it will suffer some impairment. For instance, the signal might attenuate to 6 volts. This is 50 percent attenuation, which is a substantial loss. However, the receiver will still correctly record the signal as a 0 because 6 volts is between 3 and 15 volts. The attenuation does not cause an error in reading the signal. This is why binary transmission is error-resistant.

However, think about what happens to error resistance as the number of states increases. Even with four states, a much smaller propagation effect might cause a 11 to be misinterpreted as a 10. In general, as the number of states grows, error resistance declines proportionally. Consequently, there is a strong tendency to use binary signaling in practice. In examples in this book, we will always use binary signaling.

CLOCK CYCLES When a device transmits, it holds the signal constant for a brief period of time called the **clock cycle**. The receiver can read the signal at any time within the clock cycle and read it correctly. In addition, if the sender wishes to send 1111, it will transmit two highest signals in a row. The receiver can tell that this is two signals instead of a single longer signal because the transmission will be two clock cycles long.

To transmit more bits per second, the sender uses either more states or briefer clock cycles. The latter is much easier in practice. Suppose that you are transmitting in binary and the clock cycle is 1/1000th of a second. This means that you can transmit a thousand bits per second. To transmit a gigabit per second with binary signaling, each clock cycle needs to be *one-billionth* of a second long. The limiting factor on transmission speed today is the ability of sending and receiving devices to work properly over every decreasing clock cycle times.

Test Your Understanding

7. a) What must a sender do to send the bits of a frame over a transmission medium? b) Distinguish between binary and digital signaling. c) What is a state? d) Why is binary transmission error-resistant? e) In Figure 5-7, how much could the signal attenuate before it became unreadable? f) Why is binary transmission error-resistant? g) How does error resistance differ in binary and digital signaling? h) Why are clock cycles necessary? i) If the binary transmission rate is 50 Mbps, how long will a clock cycle be?

4-Pair Unshielded Twisted Pair Copper Wiring

THE ORIGINS OF UTP DATA CABLING In your home, you use copper wiring for electricity. You also use it in your telephone (if you still have a wired telephone). In both cases, a cord contains two wires—a single pair. In contrast, Figure 5-8 shows that type of wiring that businesses had long used before data networking existed. The figure shows that business telephone wiring used four copper wire pairs in each cord. The two wires of each pair are twisted around each other several times per inch. Consequently, this type of wiring is called **4-pair unshielded twisted pair** wiring. Today, we use **4-pair UTP** wiring to carry data in Ethernet. Typically it is just called **UTP**.[1]

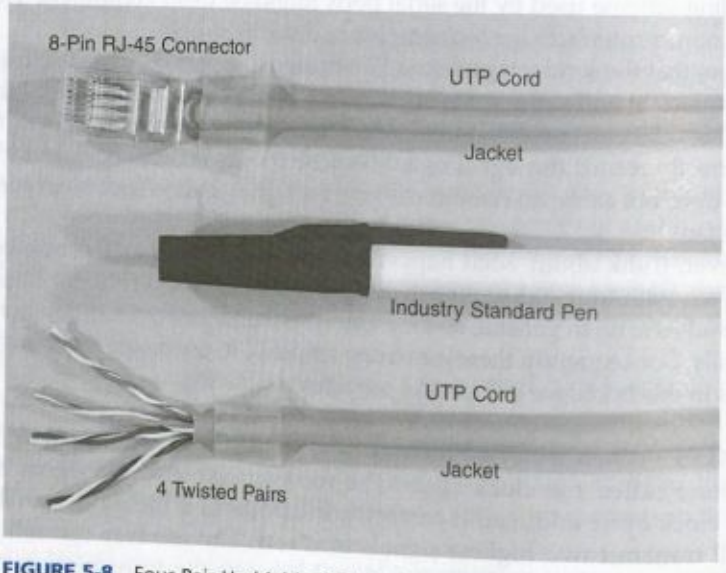

FIGURE 5-8 Four-Pair Unshielded Twisted Pair Wiring

Source: Courtesy of Raymond R. Panko

[1]OK, but what about the *unshielded* in the name? This is a hang-over from the early days of twisted pair copper wiring. In some early cases, metal shielding was placed around each wire pair, and more metal shielding was placed around the four shielded pairs. This provided protection from something we will see a little later, electromagnetic interference. However, shielded twisted pair wiring is thick and expensive. In addition, experience showed that it was rarely necessary for electromagnetic interference shielding. Consequently, the copper wiring available in stores is *unshielded* twisted pair wiring.

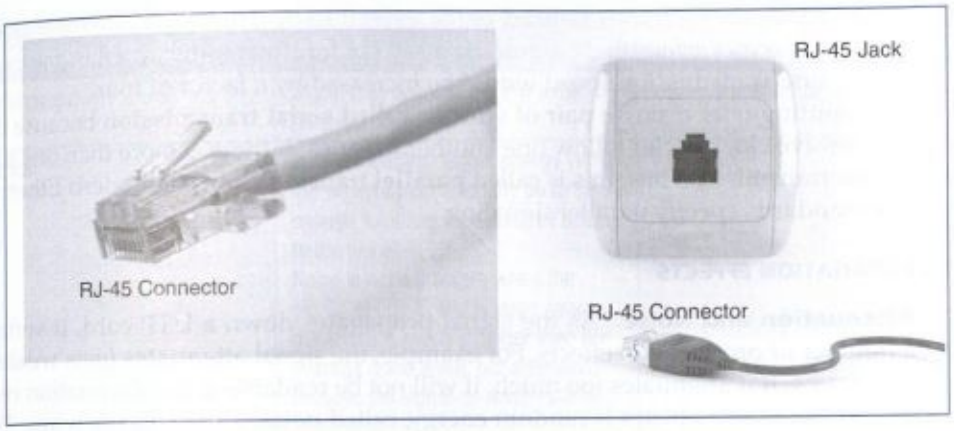

FIGURE 5-9 RJ-45 Connector and Jack

Source: © Rik Kirby/iStockphoto; © Talaj/iStockphoto

RJ-45 CONNECTORS AND JACKS Your home telephone cord terminates in a snap-in connector at each end. One connector snaps into your wall jack and the other snaps into your telephone. Figure 5-9 shows that 4-pair UTP uses a similar snap-in connector called the **RJ-45 connector**. It looks like your telephone connector, but it is a little wider because it has to terminate eight wires. It snaps into an **RJ-45 jack** in a host or a switch. The figure also shows the RJ-45 jack into which the RJ-45 connector snaps.

SERIAL AND PARALLEL TRANSMISSION Having four pairs of wires permits faster transmission speeds than having a single pair could provide. Figure 5-10 illustrates this fact. With a single pair of wires, the transmission speed would be limited to a particular value. For simplicity, assume that one bit would be transmitted per clock cycle. (This is

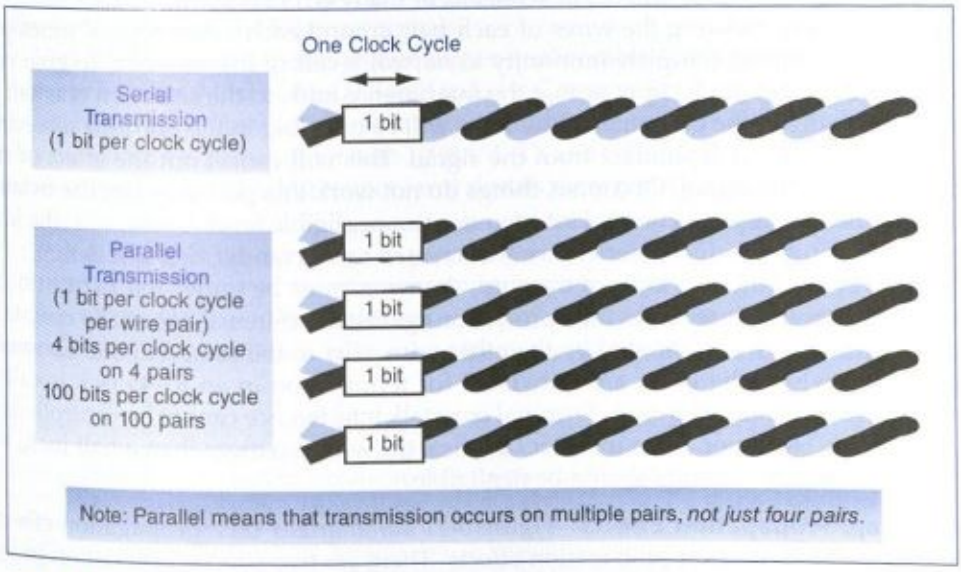

FIGURE 5-10 Serial and Parallel Transmission

binary transmission.) With four twisted pairs, even if each pair could only transmit a single bit in each clock cycle, there would be a total of four bits transmitted in the cycle by all four pairs. Transmission speed would be increased by a factor of four.

Transmitting over a single pair of wires is called **serial transmission** because the bits of successive clock cycles follow one another in series. If there is more than one pair carrying the transmission bits, this is called **parallel transmission**. All modern Ethernet signaling standards specify parallel signaling.

UTP PROPAGATION EFFECTS

Attenuation and Noise As the signal propagates down a UTP cord, it suffers from a number of propagation effects. For example, the signal **attenuates** (gets weaker) as it propagates. If it attenuates too much, it will not be readable at the destination end.

In addition, there always is random energy, called **noise**, in the copper wires. The noise adds to or subtracts from the signal. The noise is random, so it has occasional high spikes and low valleys. These will cause occasional misreads at the destination end.

The transmitted signal is much stronger than the average noise level. In technical terminology, there is a high **signal-to-noise ratio**. As the signal attenuates, it falls closer to the average noise level, and noise spikes and valleys large enough to cause errors become more likely.

With one exception, which we will see later, the 802.3 Working Group developed its signaling processes so that attenuation and noise will not be significant problems if cord lengths are limited to 100 meters. It is important to obey distance limitations because if attenuation and noise problems become significant, resultant problems can be maddeningly difficult to diagnose.

Interference UTP wire pairs are essentially long antennas. If there are air conditioners or other nearby devices that generate electromagnetic energy, the energy they generate may be picked up by each wire pair. It will be added to the signal. If this **interference** is large enough, the signal will not be readable, or many errors will occur.

Fortunately, twisting the wires of each pair around each other several times per inch provides almost complete immunity to normal levels of interference. To give you a simplified explanation, suppose that the interference adds to the signal on one half of a twist. The wire in the other half of the twist will be traveling in the opposite direction, so the interference will subtract from the signal. This will cancel out the effect of the interference on the signal. Of course, things do not work this perfectly, but the twisted nature of the wiring keeps normal interference to a negligible level. By the way, the idea of twisting wires to reduce interference was created by Alexander Graham Bell.

However, at the two ends of the cord, the wires must be untwisted to fit into the RJ-45 connectors. This removes their protection against interference, especially **crosstalk interference**, which is generated by the other wire pairs in the same bundle. Crosstalk interference where the wires are untwisted for termination in an RJ-45 jack is called **terminal crosstalk interference**. Terminal crosstalk interference cannot be controlled by wire twisting. However, if the installer untwists the wires no more than a half inch, the interference at the two ends should be negligible.

Recap: Propagation Effects Figure 5-11 summarizes UTP propagation effects. There are three important propagation effects. There are two simple installation procedures to make these propagation effects negligible.

Propagation Effect(s)	Impact	Installation Discipline
Attenuation	Signal may become too low to be received properly	Limit cord distance to 100 m
Noise	Random electromagnet energy within the wire may occasionally be large enough to produce an error by adding to the signal. Noise is not a problem when the signal-to-noise ratio is large (when the signal is much stronger than the average noise level), but attenuation reduces the signal-to-noise ratio on long cords.	
Terminal crosstalk interference	External interference by other wire pairs in the same bundle is crosstalk interference. Crosstalk interference at the two ends where the wires are untwisted to fit into RJ-45 jacks is terminal crosstalk interference. It can produce errors or make signals unreadable.	Limit the untwisting of the wires in each pair to 1.25 cm (0.5 in.) when placing them in an RJ-45 connector.

FIGURE 5-11 Recap: Propagation Effects

ETHERNET SIGNALING STANDARDS AND UTP QUALITY CATEGORIES Figure 5-12 shows that Ethernet has several signaling standards for UTP transmission. The **100BASE-TX** signaling standard has a transmission speed of 100 Mbps, while **1000BASE-T** has a transmission speed of 1 Gbps, and **10GBASE-T** has a transmission speed of 10 Gbps.[2] As you might expect, higher speeds require more expensive electronics in the device sending and receiving the Ethernet signals.

Higher transmission speeds require higher-quality cable. Cable quality is indicated by its **category** number. Most of the UTP sold today is **Category 5e** (the *e* stands for enhanced) and **Category** 6 UTP. They are almost always referred to simply as *Cat 5e* and *Cat 6*. Both quality categories are sufficient for both 100BASE-TX and 1000BASE-TX. Cat 6 cable is more expensive than Cat 5e cable but cannot carry signals more rapidly.

Cable quality is indicated by its category number.

For 10GBASE-T, which is still rare today, Category 6 wiring is sufficient for distances up to 55 meters. This is below the 100 meter maximum UTP cord runs typically

[2]In the three Ethernet signaling standards listed in Figure 5-12, the characters BASE appeared in the name. In fact, all current Ethernet standards have BASE in their names. This refers to the fact that all current standards use baseband signaling, in which the signal is injected directly into the wire. One early Ethernet standard used broadband transmission, in which the signal was sent in a radio channel. The original baseband signal was modulated onto that radio channel. There, it became the broadband signal. Broadband modulation and radios are expensive compared to baseband transmission, so broadband transmission was a dead end for future Ethernet standards. However, broadband transmission is not dead in LANs. The 802.11 wireless LANs we will see in Chapters 6 and 7 all use broadband transmission, although they do not mention this in their names.

Ethernet Signaling Standard	Transmission Speed	UTP Quality Category	Maximum Cord Length
100BASE-TX	100 Mbps	Category 5e, 6, or higher	100 meters
1000BASE-T	1 Gbps	Category 5e, 6, or higher	100 meters
10GBASE-T	10 Gbps	Category 6	55 meters
10GBASE-T	10 Gbps	Category 6A	100 meters

FIGURE 5-12 Ethernet Signaling Standards and UTP Quality Categories

possible in Ethernet. The Cat 6 wiring quality standard was supposed to be sufficient for 100 meter cord runs with 10GBASE-T signaling, so its failure to support 10 Gbps Ethernet with a length of 100 meters was a disappointment and an embarrassment. Consequently, **Category 6A** (advanced) wiring was created. Cat 6A wiring can carry 10GBASE-T signals a full 100 meters.

Test Your Understanding

8. a) What type of copper wiring is widely used in Ethernet? b) How many wires are there in a UTP cord? c) How many pairs? d) What type of connectors and jacks does 4-pair UTP use? e) What is the advantage of parallel transmission compared to serial transmission?
9. a) List the three main propagation effects that can impair a signal travelling through UTP wire. b) List the two ways in which these effects are controlled. c) Which types of propagation effects are controlled by which control method? d) Why is terminal crosstalk interference the main type of interference problem?
10. a) Of what wire characteristic is category a measure? b) What types of UTP wiring can carry signals 100 meters at 1 Gbps? c) What types of UTP wiring can carry signals in 10GBASE-T? d) Which can carry 10Gbps Ethernet 100 meters?

Optical Fiber

CORE AND CLADDING Figure 5-13 shows that optical fiber carries light signals through a thin strand of glass called the **core**. In fiber's simplest form, light is turned on for a 1 or off for a 0 during a clock cycle.

The figure also shows that the core is surrounded by a thin glass cylinder called the **cladding**. The cladding has a slightly lower index of refraction than the core. Consequently, when a light ray hits the boundary between the core and cladding, it is reflected back into the core with **perfect internal reflection**.[3] Consequently, there is very low attenuation in the light amplitude as the signal travels. Light signals can travel a very long way through LAN fiber—generally from 180 to 550 meters.

OPTICAL FIBER CORDS AND CONNECTORS Figure 5-14 shows an optical fiber cord. The cord has two **strands** for full duplex transmission, which is the ability to transmit in two directions simultaneously. Each strand carries the signal in one direction.

[3]If you remember your physics, this is Snell's Law.

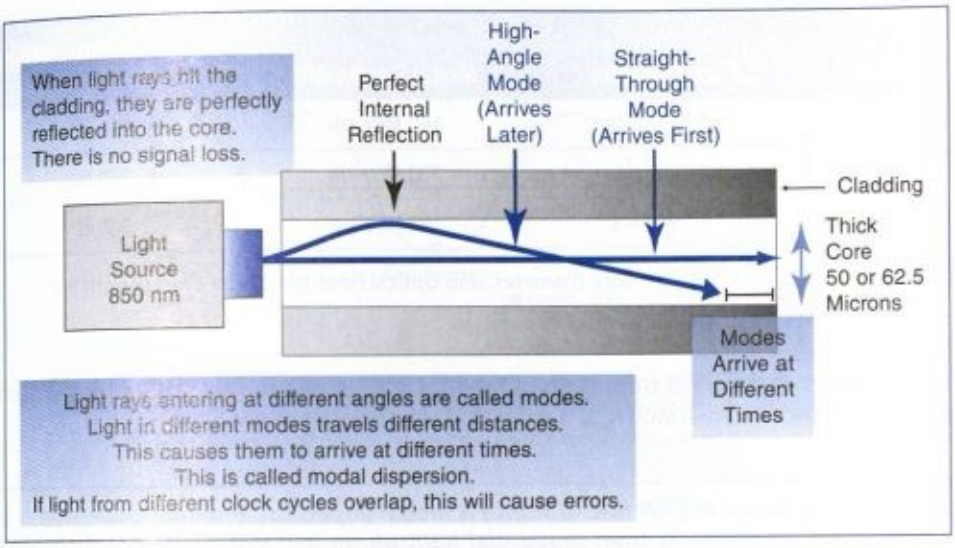

FIGURE 5-13 Optical Fiber Transmission

Note that the optical fiber cord does not terminate in RJ-45 connectors. UTP only has a single connector standard, but several optical fiber connectors and jacks have been standardized. The cord in Figure 5-14, in fact, has different connectors at each end. One end has a square **SC connector**. The other has a round **ST connector**. This cord would connect a core switch with **SC jacks** with a core switch with **ST jacks**. There is no problem mixing and matching fiber connectors. Optical fibers can be terminated with any type of standard connector.

MODAL DISPERSION AND MODAL BANDWIDTH The limiting factor in LAN fiber distance is modal dispersion. Light rays entering the fiber at different angles are called **modes**. In Figure 5-13, there are two modes. One travels straight down the center of the core. The other bounces repeatedly off the cladding back into the core. Although reflections do not lose noise energy, modes entering the core at higher angles will take longer to travel than the straight mode. This is called **modal dispersion**. If modal dispersion is

Optical Fiber Cord with Two Strands for Full Duplex Communication

FIGURE 5-14 Optical Fiber Cord and Connectors

Source: © Vitaly Shabalyn/iStockphoto

Wavelength	Core Diameter	Modal Bandwidth	Maximum Propagation Distance
850 nm	62.5 microns	160 MHz-km	220 m
850 nm	62.5 microns	200 MHz-km	270 m
850 nm	50 microns	500 MHz-km	500 m

FIGURE 5-15 Modal Bandwidth, Core Diameter, and Optical Fiber Maximum Cord Length for 1000BASE-SX

too large, parts of the light from sequential light pluses will overlap, making the signal unreadable. Modal dispersion sets a limit on LAN fiber distances.

The limiting factor in LAN fiber distance is modal dispersion. If modal dispersion is too large, parts of the light from sequential light pluses will overlap, making the signal unreadable.

To increase fiber cord length, it is necessary to use higher-quality fiber. Higher-quality LAN fiber has higher **modal bandwidth**, which is a measure of how well the fiber deals with modal dispersion. Modal bandwidth is expressed as **MHz-km**, with higher values being better. Figure 5-15 shows the relationship between modal bandwidth, core diameter, and maximum cord length in LAN fiber. The numbers are for the 1000BASE-SX, which has a signaling speed of 1 Gbps.

Note that quality in UTP is given by category number, and there are discrete categories (5e, Cat 6, etc.). In optical fiber, the situation is more fluid. Modal bandwidth is the measure of quality, and modal bandwidth varies over a wide range, giving the network designer more choices over quality and propagation distance.

Quality in UTP is expressed as a category number. Quality in LAN fiber is expressed as modal bandwidth (MHz-km).

To use Figure 5-15 in design, determine the required propagation distance for a cable run and select the appropriate core diameter and modal bandwidth. Figure 5-15 only shows three combinations. In practice, there are many, and selecting fiber for a **1000BASE-SX** trunk line requires going to catalogs from fiber vendors. We have focused on 1000BASE-SX because this is the dominant fiber standard today.

For 10 Gbps fiber runs, the situation is even more complex because while only the SX standard is popular for 1 Gbps transmission, there are multiple standards of 10 Gbps fiber signaling in Ethernet. They are collectively referred to as 10GBASE-x.

CORE DIAMETER AND LIGHT WAVELENGTH Figure 5-15 notes that modal bandwidth is not the only thing that affects a cord's maximum propagation distance. Another

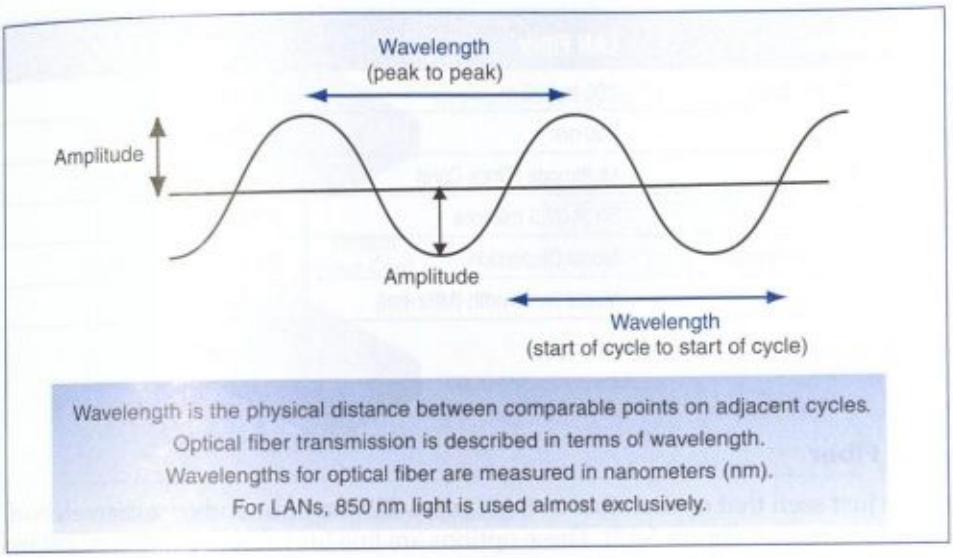

FIGURE 5-16 Light Wavelength

consideration is core **diameter**. In the United States, most companies have standardized on fiber with a core diameter of **62.5 microns**. In Europe, most selected **50 micron** fiber. A smaller diameter allows signals to travel farther. The main reason is that a larger core diameter allows modes to enter at higher angles. Consequently, there is more modal dispersion. In practice, the differences in propagation distances between 50 and 62.5 micron fiber are not great enough to cause corporations to switch standards.

In Figure 5-15, the light wavelength is listed 850 nanometers (nm). As Figure 5-16 shows, light of a single color is a cyclical electromagnetic wave. The distance between successive wave heights, troughs, starts, stops, and so forth is called the **wavelength**. The strength of the signal is called its **amplitude**. If you are familiar with ocean waves, the wavelength is the physical distance between successive waves. Amplitude is how hard the ocean wave will hit you.

No other light wavelength is listed in the table in Figure 5-15. Wavelengths of 1,310 and 1,550 nm give longer propagation distances, but they cost more to generate. For LAN distances, 850 nm light transmitters and receivers are perfectly adequate. In the 1000BASE-SX standard, the *S* stands for short wavelength (800 nm).

Test Your Understanding

11. a) Does the signal travel through the optical fiber core, cladding, or both? b) Why can signals travel very far through optical fiber? c) Why does an optical fiber cord have two strands? d) What is the ability to transmit in both directions simultaneously called? e) Why does modal dispersion happen? f) When will modal dispersion be a problem? g) What is the measure of optical fiber quality? h) In what units is modal bandwidth expressed? i) If you use 1000BASE-SX fiber, what modal bandwidth do you need to transmit a signal 250 meters? j) Will light travel farther in 50 micron fiber or 62.5 micron fiber? k) Of what is wavelength a measure?

Characteristic	LAN Fiber	Carrier WAN Fiber
Required Distance Span	200 to 300 m	1 to 40 m
Light Wavelength	850 nm	1,310 or 1,550 nm
Type of Fiber	Multimode (Thick Core)	Single-Mode (Thin Core)
Core Diameter	50 or 62.5 microns	8.3 microns
Primary Distance Limitation	Modal Dispersion	Absorptive Attenuation
Quality Metric	Modal Bandwidth (MHz-km)	Not Applicable

FIGURE 5-17 LAN Fiber versus Carrier Fiber

Carrier Fiber

We have just seen that optical fiber in LANs uses 850 nm light and core diameters of 50 or 62.5 microns (see Figure 5-17). These options are fine for LAN propagation distances, and they give reasonable cost. Therefore, we have called this type of fiber LAN fiber. Fiber with a core diameter of 50 or 62.5 microns is called **multimode fiber** because, as we saw in Figure 5-13, light modes can enter at various angles.

Telecommunications carriers, in contrast, need much greater propagation distances—10 km to 40 km or more. This requires them to use expensive 1,310 nm or even 1,550 nm light signaling. They also need to use fiber with very tiny cores. A typical diameter for carrier fiber is 8.2 microns instead of 50 or 62.5 microns. Fiber with such a thin core diameter is called **single-mode fiber** because only a single mode—the one traveling straight through the core—can propagate through the core. This completely eliminates modal dispersion, which is the main distance limiter for multimode fiber. This does not mean that signals travel forever in single-mode fiber. The signal still attenuates because it is slightly absorbed by the glass as it propagates. Beyond 10 km or so, the signal becomes too attenuated to be readable. **Absorptive attenuation**, then, is the main distance limiter for carrier fiber.

Test Your Understanding

12. a) Comparing LAN and WAN fiber, what are distance limits? b) What light wavelengths are used? c) What are the two diameters for multimode fiber? d) What is the diameter of single-mode fiber? e) What are the principle distance limiting factors for LAN and carrier fiber? f) Is modal bandwidth a quality measure for LAN fiber, carrier fiber, or both?

Link Aggregation (Bonding)

Ethernet transmission capacity usually increases by a factor of 10. What should you do if you only need somewhat more speed than a certain standard specifies? For instance, suppose that you have gigabit Ethernet switches and need to connect them at 1.5 Gbps?

Figure 5-18 illustrates that a company can use two or more trunk lines to connect a single pair of switches. The IEEE calls this **link aggregation**. Networking professionals also call this **bonding**.

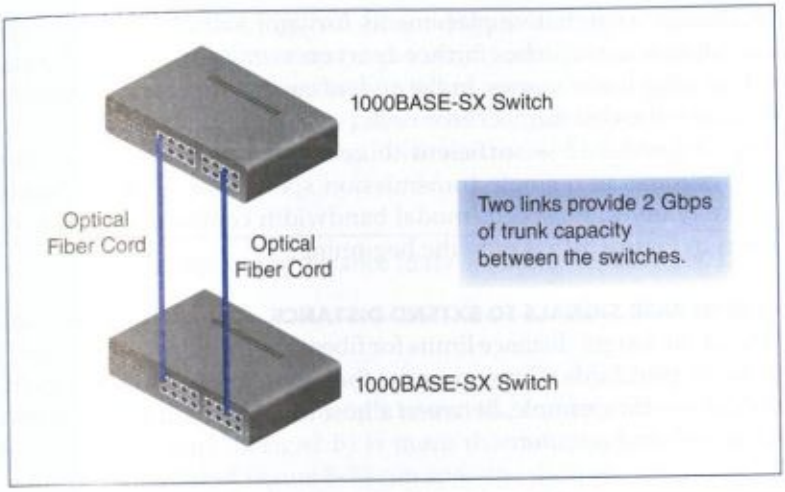

FIGURE 5-18 Link Aggregation (Bonding)

Link aggregation allows you to increase trunk speed incrementally, by a factor of two or three, instead of by a factor of ten. This incremental growth uses existing ports and usually is inexpensive compared to purchasing new faster switches.

However, after two or three aggregated links, the company should compare the cost of link aggregation with the cost of a tenfold increase in capacity by moving up to the next Ethernet speed. Going to a much faster trunk line will also give more room for growth.

Test Your Understanding

13. a) What is link aggregation? b) What is it also called? c) If you need to connect two 1000BASE-SX switches at 2.5 Gbps, what are your options? d) Why may link aggregation be more desirable than installing a single faster link? e) Why may link aggregation not be desirable if you will need several aggregated links to meet capacity requirements?

Ethernet Physical Layer Standards and Network Design

USING FIGURE 5-12 AND FIGURE 5-15 IN NETWORK DESIGN Note that if you know the speed you need (100 Mbps, 1 Gbps, etc.) and if you know what distance you need to span, the information in Figure 5-12 and Figure 5-15 will show you what type of transmission link you can use. Because link aggregation is available on all core switches, you have even more choices.

For instance, suppose that you need a speed of 2.5 Gbps between two switches that are 130 meters apart. This is over 100 meters, so you could not use UTP. You would need optical fiber to span this distance.

For speed, 1 Gbps would not be sufficient, and 10 Gbps might be expensive. Your best choice probably would be three bonded 1000BASE-SX links, although you would consider the cost of moving up to a 10 Gbps fiber standard.

Alternatively, if you are designing a network from scratch, say for a new facility, the options presented in Figure 5-12 for UTP and in Figure 5-15 for optical fiber will

allow you to consider alternative placements for your switches. With longer physical links, you can place your switches farther apart on average, reducing the total number of switches. This might save money. In the end, of course, you have to consider multiple options and crunch the cost numbers for each.

Although Figure 5-12 is sufficient to consider maximum Ethernet distances, Figure 5-15 only looks at a single transmission speed and a few modal bandwidth levels. There are many more speed/modal bandwidth combinations. For optical fiber, the information in Figure 5-15 is only the beginning.

SWITCHES REGENERATE SIGNALS TO EXTEND DISTANCE The normal 100-meter Ethernet limit for UTP and the longer distance limits for fiber shown in Figure 5-12 and Figure 5-15 are physical layer standards. Consequently, they only apply to connections *between a single pair of devices*—for example, between a host and a switch, between two switches, or between a switch and a router.

The 100-meter Ethernet limit for UTP and the longer distance limits for fiber shown in Figure 5-12 and Figure 5-15 only apply to physical links between pairs of devices, not to end-to-end data links between hosts across multiple switches.

What should you do if a longer distance separates the source host and the destination host? Figure 5-19 shows a data link with two intermediate switches. In addition to the two 100-meter maximum length UTP access links, there is a 220-meter maximum length 1000BASE-SX optical fiber link (using 62.5/125 micron 160 MHz-km modal bandwidth fiber) between the two switches. This setup can support a data link with a maximum span of 420 meters.

Each switch along the way **regenerates** the signal. If the signal sent by the source host begins as a 1, it is likely to be distorted before it reaches the first switch. The first switch recognizes it as a 1 and generates a clean new 1 signal to send to the second switch. The second switch regenerates the 1 as well.

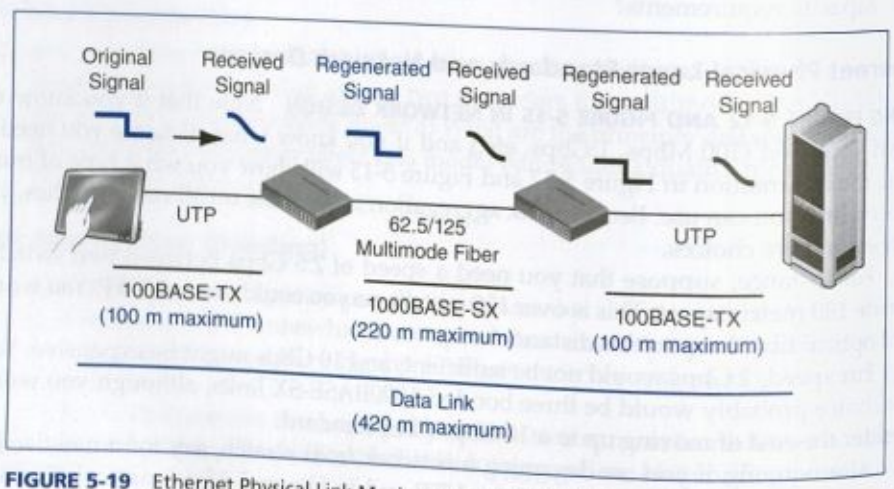

FIGURE 5-19 Ethernet Physical Link Maximums and Unlimited Data Link Distances

The key point again is that Figure 5-12 and Figure 5-15 show maximum distances between *pairs* of devices, not maximum end-to-end transmission distances. To deliver frames over long distances, intermediate switches regenerate the signal. There is no maximum end-to-end distance between pairs of hosts in an Ethernet network. In layer terminology, there are maximum physical link distances, but this does not equate to maximum data link distances.

There is no maximum end-to-end distance (data link distance) between pairs of hosts in an Ethernet network.

Test Your Understanding

14. a) What steps would you go through to use the information in Figure 5-12 and Figure 5-15 in network design? b) If more than one type of Ethernet standard shown in Figure 5-12 and Figure 5-15 can span the distance you need, what would determine which one you choose? c) In Figure 5-12 and Figure 5-15, is the maximum distance the maximum distance for a single physical link or for the data link between two hosts across multiple switches? d) At what layer or layers is the 802.3 100BASE-TX standard defined—physical, data link, or internet? e) How does regeneration allow a firm to create LANs that span very long distances? f) If you need to span 300 meters by using 1000BASE-SX, what options do you have? (Include the possibility of using an intermediate switch.) g) How would you decide which option to choose?

THE ETHERNET FRAME

So far, we have been looking at Ethernet physical layer standards. Now, we will look at Ethernet data link layer standards, beginning with frame organization. The Ethernet frame is called the Ethernet **media access control (MAC)** frame.[4]

The Ethernet Frame's Organization

Figure 5-20 shows the Ethernet MAC layer frame, which we saw briefly in Chapter 2. We will now look at the Ethernet frame in more depth. Recall that an *octet* is a byte.

Preamble and Start of Frame Delimiter Fields

Before a play in American football, the quarterback calls out something like "Hut one, hut two, hut three, hike!" This cadence synchronizes all of the offensive players.

In the Ethernet MAC frame, the **preamble field** (7 octets) and the **start of frame delimiter field** (1 octet) synchronize the receiver's clock to the sender's clock. These fields have a strong rhythm of alternating 1s and 0s. The last bit in this sequence is a 1 instead of the expected 0, to signal that the synchronization is finished.

[4]The 802 Committee divided the data link layer into two parts. The lower part was the media access control (MAC) sublayer. This sublayer is different for each technology (802.3, 802.11, etc.). The higher sublayer is the logical link control (LLC) sublayer. The LLC layer does not have planning or management implications, so we will not consider it. It is, however, the reason for the LLC subheader that we will see in the MAC layer frame's data field.

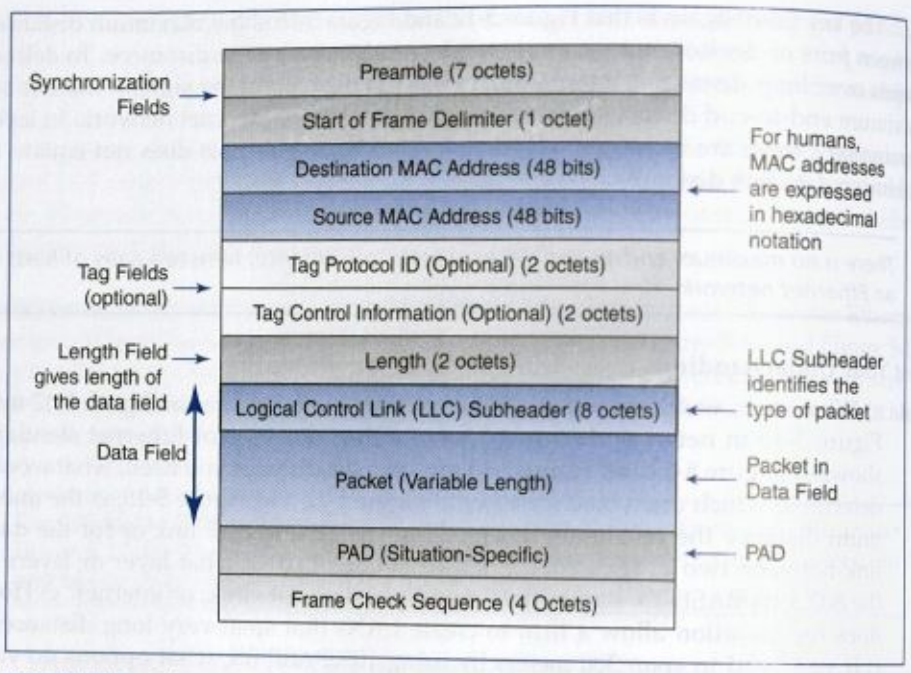

FIGURE 5-20 Ethernet Media Access Control (MAC) Frame Organization

Source and Destination Address Fields

HEX NOTATION We saw in Chapter 2 that the source and destination Ethernet address fields are 48 bits long and that while computers work with this raw 48-bit form, humans normally express these addresses in Base 16 **hexadecimal (hex) notation**. To convert a 48-bit Ethernet address into hex notation, follow these three steps:

- First, divide the 48 bits into twelve 4-bit units, which computer scientists call nibbles.
- Second, convert each nibble into a hexadecimal symbol, using Figure 5-21.
- Third, write the symbols as six pairs with a dash between each pair—for instance, B2-CC-66-0D-5E-BA. (Each pair represents 1 octet.)

To convert a hex address back to binary, change each symbol back to its 8-bit pattern. For example, if the first hex pair is E2, E is 1110, and 2 is 0010. So E2 is equivalent to the octet 11100010. In this conversion, you must keep the two leading zeros in 0010.

MAC ADDRESSES Ethernet addresses exist at the MAC layer, so Ethernet addresses are called **MAC addresses**. They are also called **physical addresses** because physical devices (NICs) implement Ethernet at the physical, MAC, and LLC layers.

Tag Fields

We will look at the two tag fields later in this chapter, when we look at virtual LANs (VLANs). These tag fields are optional and are only used under certain circumstances.

4 Bits	Decimal (Base 10)	Hexadecimal (Base 16)
0000	0	0 hex
0001	1	1 hex
0010	2	2 hex
0011	3	3 hex
0100	4	4 hex
0101	5	5 hex
0110	6	6 hex
0111	7	7 hex

4 Bits	Decimal (Base 10)	Hexadecimal (Base 16)
1000	8	8 hex
1001	9	9 hex
1010	10	A hex
1011	11	B hex
1100	12	C hex
1101	13	D hex
1110	14	E hex
1111	15	F hex

FIGURE 5-21 Hexadecimal Notation

Note: Divide a 48-bit Ethernet address into 12 four-bit "nibbles." (1010, 0001, etc.)
Convert each group of 4 bits into a hex symbol. (A, 1, etc.)
Combine two hex symbols into pairs and place a dash between pairs (A, 1, etc.)
The finished hex expression: A1-36-CD-7B-DF-01 hex.

Length Field

The **length** field contains a binary number that gives the *length of the data field* (not of the entire frame) in octets. The maximum length of the data field is 1,500 octets. There is no minimum length for the data field. However, we will see that if the data field is less than 46 octets long, a PAD field will be added.

The Data Field

The **data field** contains two subfields: the LLC subheader and the packet that the frame is delivering.[5]

LLC SUBHEADER The **logical link control layer (LLC) subheader** is 8 octets. The purpose of the LLC subheader is to describe the type of packet contained in the data field. For instance, if the LLC subheader ends with the code 08-00 hex (Base 16), then the data field contains an IPv4 packet.[6] Ethernet frames can also carry other types of packets. To give another example, the code 86DD (hex) indicates the presence of an IPv6 packet.

[5]Why does the data field have two parts? The answer is that the data field of the MAC layer frame actually is an encapsulated LLC layer frame, which has a header (the LLC subheader) and a data field consisting of the packet being carried in the LLC frame. However, to avoid damaging neurons, it is best simply to think of the MAC layer data field as having two parts.

[6]The LLC subheader has several fields. In the SNAP version of LLC, which is almost always used, the first three octets are always AA-AA-03 hex. The next three octets are almost always 00-00-00 hex. The final two octets constitute the Ethertype field, which specifies the kind of packet in the data field. Common hexadecimal Ethertype values are 0800 (IP), 8137 (IPX), 809B (AppleTalk), 80D5 (SNA services), and 86DD (IP version 6).

THE PACKET The data field also contains the packet that the MAC layer frame is delivering. The packet usually is far longer than all other fields combined.

PAD Field

The **PAD field** is unusual because it does not always exist. Although there is no minimum length for 802.3 MAC layer frame data fields, if the data field is less than 46 octets long, then the sender must add a PAD field so that the total length of the data field and the PAD field is exactly 46 octets long. For instance, if the data field is 26 octets long, the sender will add a 20-octet PAD field. If the data field is 46 octets long or longer, the sender will not add a PAD field.

There is no minimum length for the data field, but if the data field is less than 46 octets long, then a PAD field must be added to bring the total length of the data and pad fields to 46 octets.

How does the receiving NIC know what part of the data field plus the PAD is the data field? Recall that the length field gives the *length of the data field*. Consequently, after reading the LLC header, the remaining number of octets indicated in the length field must be the data field. Everything else beyond the data field that is needed to get the data field and PAD to 46 octets is the PAD field. The receiving NIC ignores the contents of the PAD.

For example, suppose that the length field is 40 octets. This means that the data field is 40 octets. The LLC subheader is 8 octets, so the packet length is 32 octets. The added PAD is 6 octets, because 6 octets must be added to 40 octets to make a total of 46 octets.

Frame Check Sequence Field

The last field in the Ethernet frame is the **Frame Check Sequence Field**, which permits error detection. This is a 4-octet field. The sender does a calculation based on other bits in the frame and places the 32-bit result in the Frame Check Sequence Field. The receiver redoes the calculation and compares its result with the contents of the Frame Check Sequence Field. If the two are different, there is an error in the frame. If there is an error, the receiver simply discards the frame. There is no retransmission of damaged frames.

Test Your Understanding

15. a) What is the purpose of the preamble and start of frame delimiter fields? b) Why are Ethernet addresses called MAC addresses or physical addresses? c) What are the steps in converting 48-bit MAC addresses into hex notation? d) Convert 11000010 to hex. e) Convert 7F hex to binary. f) The length field gives the length of what? g) If the length field is 1020, what is the length of the packet in the data field? h) What are the two components of the Ethernet data field? i) What is the purpose of the LLC subheader? j) What type of packet is usually carried in the data field? k) What is the maximum length of the data field? l) Who adds the PAD field—the sender or the receiver? m) Is there a minimum length for the data field? n) If the data field is 40 octets long, how long a PAD field must the sender add? o) If the data field is 400 octets long, how long a PAD field must the sender add? p) What is the purpose of the Frame Check Sequence Field? q) What happens if the receiver detects an error in a frame?

BASIC ETHERNET DATA LINK LAYER SWITCH OPERATION

In this section, we will discuss the basic data link layer operation of Ethernet switches. This is also governed by the 802.3 MAC layer standard. In the section after this one, we will discuss other aspects of Ethernet switching that a firm may or may not use.

Frame Forwarding

Figure 5-22 shows an Ethernet LAN with three switches. Larger Ethernet LANs have dozens of switches, but the operation of individual switches is the same whether there are a few switches or many. Each individual switch makes a decision about which port to use to send the frame back out to the next switch.

In the figure, Host A1 wishes to send a frame to Host E5. This frame must go to Switch 1, then Switch 2, and then to Switch 3. Switch 3 will send the frame to Host E5.

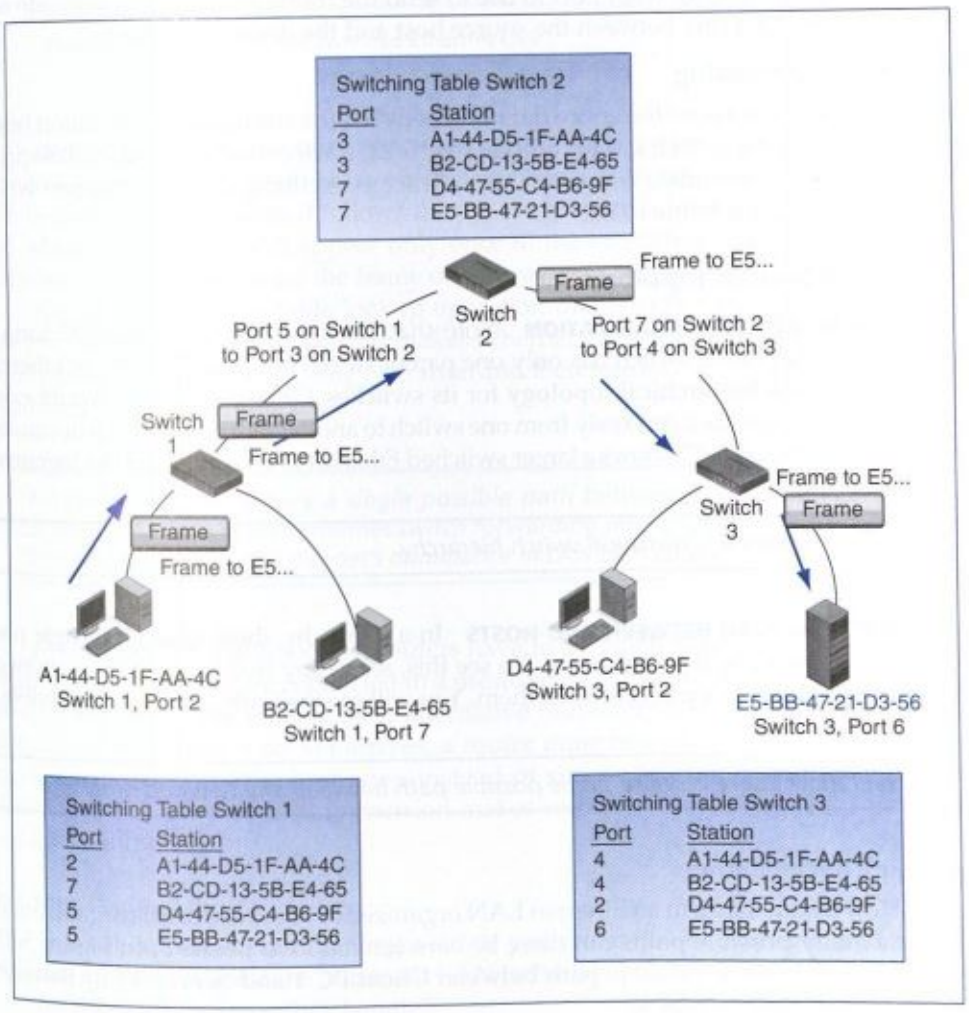

FIGURE 5-22 Multi-Switch Operation

To begin this process, Host A1 puts E5 (later octets dropped for brevity) in the destination address field of the frame. It sends the frame to Switch 1, through Port 2.

- Switch 1 looks up the address E5 in its switching table. It sees that E5 is associated with Port 5, so it sends the frame out Port 5. This is a very simple process, so it requires little processing power. This means that Ethernet switches are inexpensive for the volume of traffic they carry.
- The frame going out Port 5 on Switch 1 goes into Port 3 on Switch 2. Switch 2 now looks up the address E5 in its switching table. This address is associated with Port 7, so Switch 2 sends the frame out Port 7.
- The frame arrives at Switch 3 through Port 4. Switch 3 now looks up the address E5 in the switching table. This time, the address is associated with Port 6. Switch 3 sends the frame out Port 6. This takes it to the destination Host E5.

Note that each switch only knows the information in its switching table. More specifically, it only knows what port to use to send the frame back out. Switches do not know the entire data link between the source host and the destination host.

Test Your Understanding

16. a) Do switches know the entire data link between the source and destination host? b) What does a switch know? c) In Figure 5-22, trace everything that will happen when Host E5 sends a frame to D4. d) Trace everything that will happen when Host E5 sends a frame to B2.

Hierarchical Switch Topology

HIERARCHICAL SWITCH ORGANIZATION Note that the switches in Figure 5-22 form a **hierarchy**, in which each switch has only one parent switch above it. In fact, the Ethernet standard *requires* a **hierarchical topology** for its switches. Otherwise, loops would exist, causing frames to circulate endlessly from one switch to another around the loop or causing other problems. Figure 5-23 shows a larger switched Ethernet LAN organized in a hierarchy.

Ethernet requires a hierarchical switch hierarchy.

SINGLE POSSIBLE PATH BETWEEN END HOSTS In a hierarchy, there is only a single possible path between any two end hosts. (To see this, select any two hosts at the bottom of the hierarchy and trace a path between them. You will see that only one path is possible.)

In a hierarchy, there is only a single possible path between any two end hosts.

Test Your Understanding

17. a) How are switches in an Ethernet LAN organized? b) Because of this organization, how many possible paths can there be between any two hosts? c) In Figure 5-23, what is the single possible path between Client PC 1 and Server Y? d) Between Client PC 1 and Server X?

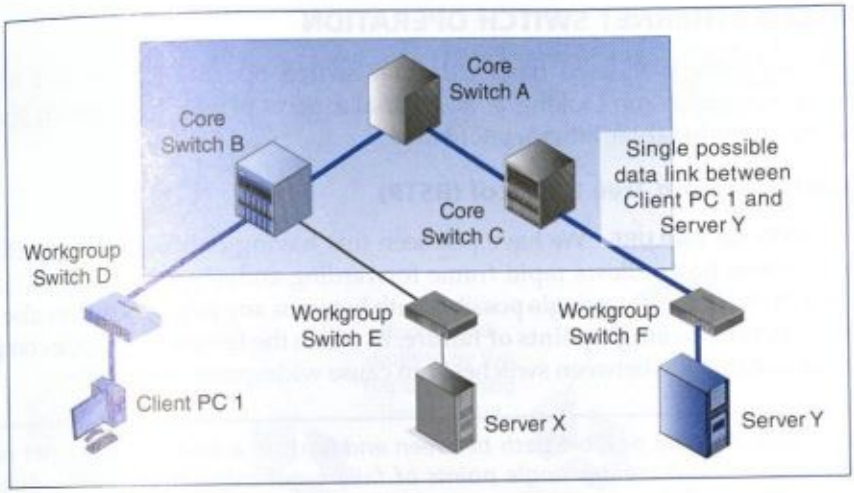

FIGURE 5-23 Hierarchical Switched Ethernet LAN

ONLY ONE POSSIBLE PATH: LOW SWITCHING COST We have just seen that a hierarchy allows only one possible path between any two hosts. If there is only a single possible path between any two hosts, it follows that, in every switch along the path, the destination address in a frame will appear only once in the switching table—for the specific outgoing port needed to send the frame on its way.

This allows a simple table lookup operation that is very fast and therefore costs little per frame handled. This is what makes Ethernet switches inexpensive. As noted in the introduction, simple switching operation and therefore low cost has led to Ethernet's dominance in LAN technology.

The fact that there is only a single possible path between any two end hosts in an Ethernet hierarchy makes Ethernet switch forwarding simple and therefore inexpensive. This low cost has led to Ethernet's dominance in LAN technology.

In Chapter 8, we will see that routers have to do much more work when a packet arrives because routers are connected in a mesh, so there are multiple alternative routes between any two hosts. Each of these alternative routes appears as a row in the routing table. Therefore, when a packet arrives, a router must first identify all possible routes (rows) and then select the best one—instead of simply finding a single match. This additional work per forwarding decision makes routers very expensive for the traffic load they handle.

Test Your Understanding

18. a) What is the benefit of having a single possible path? Explain in detail. b) Why has Ethernet become the dominant LAN technology? c) Why are routers expensive for the traffic volume they handle?