Supplementary Chapter 15

Wireless

What is essential is invisible to the eye
—Antoine de Saint-Exupéry, *The Little Prince*, 1943

Overview

Wireless networking has gained rapid popularity since its introduction in the late 1990s. Virtually every laptop computer sold since the early 2000s is capable of wireless networking. There are various kinds of wireless networks available today, each serving a specific user need. This chapter introduces the enabling legal provisions that make free wireless networking possible, along with the different kinds of wireless networks. At the end of this chapter, you should know:

- the business impact of wireless networking;
- the special features of frequency bands that are used for wireless networking;
- how wireless local-area networks work;
- the types of wireless local-area networks;
- how wireless personal-area networks work; and
- how wireless metropolitan-area networks work.

Introduction

Wireless networks are computer networks that use the ISM wireless frequency bands for signal transmission. Wireless networks have become enormously popular among both computer users and businesses since the early years of the twenty-first century. Many cities have experimented with citywide wireless networks to provide free or inexpensive Internet access to citizens. Reasons for the popularity of wireless networks include their convenience and ease of deployment. On battery-powered laptops, wireless networking allows users to compute and communicate without any power or network cords. Businesses like wireless networking because setting up a basic wireless network in a small office requires nothing more than an inexpensive wireless router. By comparison, wired networking requires cables to be drawn through ceilings, floors, and walls. Wireless networking is becoming so popular that many organizations are finding that more than half of the Ethernet ports in the organization are unused because users prefer wireless networks over wired networks.[1] Wireless networking may be one of those rare services loved by both businesses and employees.

Wireless networking introduces some important concerns and limitations that users and businesses should be aware of. The most visible concern is information security. Wired networks have wall outlets in specific locations that can only be reached by users with access to the building. By

[1] J. Cox, "Is It Time to Cut Back on Now-Idle Ethernet?" *Network World*, 26 (2009): 1.

Wi-Fi in stadiums

Wi-Fi is becoming the standard mechanism for supplementing cell phone capacity in dense areas. At Super Bowl 50 in 2016, the 10 terabyte Wi-Fi data transfer mark in a single game was crossed for the first time. Seventy thousand fans used more than 1,300 Wi-Fi access points and more than 1,200 Bluetooth beacons to send selfies and messages to friends around the world.

contrast, wireless signals spread out in all directions and can easily bleed outside the organization's boundaries. Without adequate security, malicious users can easily access the organization's computer network through an improperly secured wireless access point.

In a well-publicized example (as noted in Chapter 11), in 2006, the retail chain T. J. Maxx became the target of a hack when attackers were able to drive to the parking lot of a Marshalls store in Minnesota and sniff the passwords of store managers as they logged into the network. Because of other weaknesses in T. J. Maxx's network, these hackers were able to retrieve most of the credit card information stored on T. J. Maxx's computers. More than 45 million credit card records were stolen. The breach is estimated to have cost the company $250 million, including costs to settle lawsuits resulting from the breach. All this began with just an improperly secured wireless LAN at one of its stores.

One final point is the limitation of wireless networks. Wireless networks are generally slower and less reliable than wired networks. Most wireless networks share bandwidth with other applications, such as cordless telephones, and are affected by environmental conditions. Connection drop-offs are common with wireless networks. This is not a major concern for browsing, email, and other light applications. However, when continuity or speed of the connection is essential, wired networks are still greatly preferred over wireless networks.

Italian inventor Guglielmo Marconi started wireless communications in 1895 by sending a Morse message over a distance of a mile. Marconi was awarded the Nobel Prize in physics in 1909 for his contributions toward the development of wireless telegraphy.

ISM Frequency Bands

Free or inexpensive wireless networking is possible because of the existence of a very special category of wireless frequencies. Before we look at wireless technologies, it is useful to become aware of these enabling frequencies.

The special signal frequencies that enable wireless networking are called *ISM frequencies*. *ISM frequencies or ISM bands are radio frequencies that are available internationally for free use for industrial, scientific, and medical applications*. The terms *industrial* and *scientific* are interpreted very broadly, and ISM frequencies may be put to almost any use by anybody without permission from anyone or payments of license fees to anyone. These frequencies are therefore also called *unregulated frequencies*. Cordless phones, remote-controlled cars, microwave ovens, wireless keyboards, and mice are other applications that use ISM frequencies.

Bill: What did the Vikings use to communicate secretly?

Jill: I don't know.

Bill: The Norse code!

Source: *Boys' Life* magazine, 2013

Wireless frequencies have become big business for government. We know from the discussion in Chapter 2 on physical layers that for distinct separation at the receiving end, there must be only one sine wave at a specific frequency in any given location. Since there are many users who would like to use sine waves for wireless transmissions but only one user can transmit at any given frequency, some coordination and allocation is necessary to determine who can transmit at a specific frequency. In the United States, this coordination is done by the Federal Communications Commission (FCC). In the early days, the FCC did not charge fees for the privilege of using specific frequencies for transmission. Instead, frequency bands were allocated based on technological requirements. However, beginning in 1994, the FCC realized that operators of cell phones and other services would be willing to pay for access to specific frequencies. Accordingly, the FCC began spectrum auctions to allocate frequency bands for specific commercial services to the highest bidders offering these services.

Why are ISM frequencies available for free use when cell phone operators pay billions of dollars to use other frequencies? One reason is that regulators recognize the need for wireless frequencies for experimentation and amateur use. The specific frequencies that have been selected for ISM use are generally not very useful for commercial use. ISM frequencies typically have poor transmission properties and are unlikely to fetch meaningful prices at auctions. For example, signals at the 2.45 GHz band are strongly absorbed by water. Microwave ovens operate at this frequency, since almost 75% of food mass is made of water. By quickly transferring energy to the water in food, microwave ovens are able to heat and cook quickly and with very high efficiency. Similarly, water vapor in the atmosphere absorbs signals in the 2.45 GHz band, resulting in a very short range for these signals. ISM frequencies are also absorbed by walls and foliage. Commercial operators are unlikely to pay for signals that have poor transmission properties.

Percy Spencer, the inventor of microwave ovens[2]

Percy Spencer was a self-educated inventor who supported himself and his aunt from the age of seven. During his time in the US Navy, he became fascinated with wireless signals after learning about the wireless operators aboard the *Titanic*. While working at Raytheon on a radar project for the US Department of Defense, he noticed that the chocolate bar in his pocket melted when he got close to the radar equipment. Investigating this further led to the commercial development of the microwave oven.

Table 15-1. ISM frequency bands in the United States

ISM frequency	Bandwidth
6.78 MHz	± 15.0 kHz
13.56 MHz	± 7.0 kHz
27.12 MHz	± 163.0 kHz
40.68 MHz	± 20.0 kHz
915 MHz	± 13.0 MHz
2.45 GHz	± 50.0 MHz
5.8 GHz	± 75.0 MHz
6.525 GHz	± 600 MHz
24.125 GHz	± 125.0 MHz
61.25 GHz	± 250.0 MHz
122.5 GHz	± 500.0 MHz
245 GHz	± 1.0 GHz

The frequencies defined for ISM use in the United States are shown in Table 15-1. Fortunately, ISM applications make excellent use of these otherwise useless frequencies. You may recognize some of these frequencies. Remote controls for radio-controlled cars often use the 40.68 MHz band. Older cordless phones used the 915 MHz band, while most current cordless phones use the 2.45 GHz band. Most of the popular wireless LANs also use the 2.45 GHz band. The highest frequency ISM bands will become useful when electronic devices operating at these extremely high frequencies can be built at more affordable prices.

The 6.525 GHz band is the most recent addition to the ISM spectrum. It was approved in 2018 by the FCC to expand the availability of ISM bands for communication use, particularly to assist 5G services. Observe the unusually large bandwidth of the band compared to

[2] https://en.wikipedia.org/wiki/Percy_Spencer (accessed Feb. 2020).

other ISM bands. The success of wireless LANs motivated the FCC to carve out this bandwidth and facilitate high-speed wireless data services. This band was previously allocated to TV operators to transmit live footage from mobile trucks to studios. Wireless networks will require measures to minimize interference with these transmissions.

Wireless Network Categories

There are three primary categories of wireless computer networks. All these categories of wireless networks use ISM frequencies. The most familiar are wireless LANs, which go by names such as 802.11b and 802.11g. These networks have a range of about 100 feet, which is enough to cover an average suburban home or small office. The second category of wireless networks is Bluetooth, which is called a personal-area network. This technology is used for connectivity within about 10 feet, which is ideal for connecting peripheral devices such as wireless keyboards and cameras in the immediate vicinity of a computer. Finally, we have an emerging category of wireless networks called metropolitan area networks, which can provide coverage over a range of about 20 miles, which is enough to cover many metropolitan towns. In the rest of this chapter, we will look at the technologies and features of each of these categories of wireless networks.

Wireless Local-Area Networks (the 802.11 series)

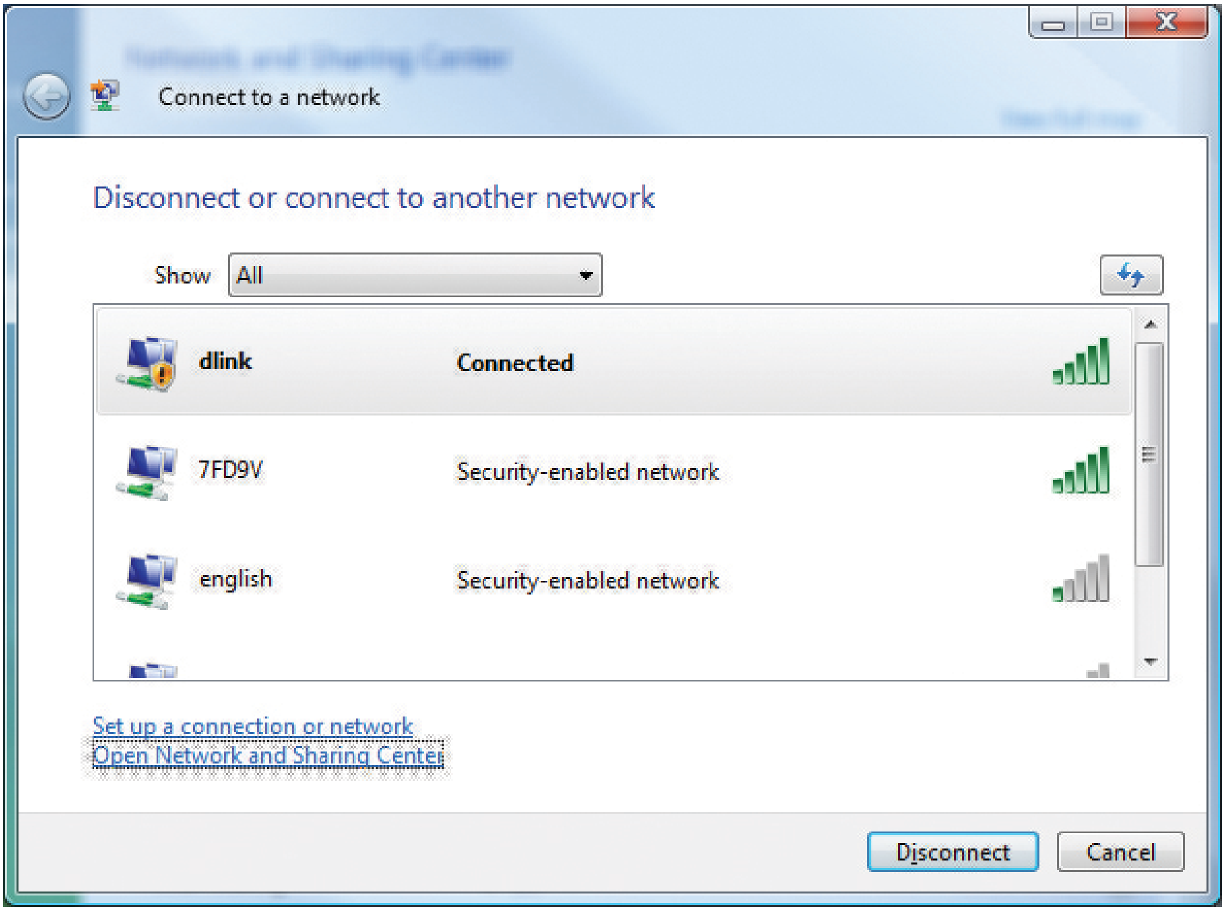
Wireless local-area networks are the most familiar of the three categories of wireless networks. Most college campuses now have blanket wireless LAN coverage, and many college students use laptops with built-in wireless LAN capability to access the Internet. The technologies used in wireless LANs are specified in the 802.11 series of IEEE standards such as 802.11b, 802.11g, 802.11a, and 802.11n. Wireless LANs are often known as Wi-Fi.

Technologically, wireless LANs share many similarities with Ethernet, which is a wired network and was discussed in detail in the context of the data-link layer. For example, the frame structure of wireless LANs is almost identical to the frame structure of Ethernet. Also, wireless LANs use the 48-bit MAC addresses discussed with Ethernet.

There are, however, some important differences between Ethernet and wireless LANs. The most important difference is that wireless LANs have no defined boundaries. Wall jacks define the end points for Ethernet. An Ethernet wall jack is hardwired to a specific port on a specific switch. As a result, a network administrator can control every aspect of the network traffic that flows through the wall jack and to the computer connected to the jack. When you connect to the network through a wall jack, you become part of a well-defined network. Typically one area of an office is served by one switch, and most users have no choice but to become a member of the Ethernet network that is closest to them.

On the other hand, wireless networks can overlap, and they often do. At home, if you open up your "connect to network" dialog (right-click on the wireless icon in your system tray → connect to a network), you are likely to see wireless networks from many of your neighbors, as shown in Figure 15-1. If any of these wireless networks is not security enabled, you can use it to connect to the Internet. At the airport, you are likely to see overlapping wireless networks from the airport operator, Starbucks, cell phone companies, and so on. Again, if any of these is not security enabled, you can use it to get Internet access.

Therefore, whereas geographical location uniquely defines network membership in Ethernet, it does not define network membership in wireless LANs. The technical implication is that whereas the signal strength of a wired connection always meets Ethernet standards, the signal strength, and hence the network experience of a wireless connection, cannot be specified. The network performance of a wireless connection depends upon the distance of the host from the access point. A user who is very far from an access point will get very weak signals. To best serve users at different distances, wireless LAN standards specify different signal-modulation schemes for



Connect to a network

Disconnect or connect to another network

Show: All

dlink	Connected	
7FD9V	Security-enabled network	
english	Security-enabled network	

Set up a connection or network
Open Network and Sharing Center
Disconnect
Cancel

Figure 15-1. Wireless networks can overlap

users at different distances from access points. Users who are close to an access point are served by faster-changing signals that can carry higher data rates but need strong signal strength for reliable detection. Users who are farther away from access points are served by signals that can only provide lower data rates but are easier to detect in the presence of noise. (You may be able to relate this to the discussion on signal detection, especially relating to signal reception in the presence of noise.)

Another difference between wireless LANs and wired LANs is that whereas wired networks are extremely reliable, wireless is an inherently unreliable medium. Wireless networks are hurt by adverse weather, humidity, temperature, and other environmental conditions. As a result, the boundaries of a wireless network are not stable and keep shifting as environmental conditions change. Also, wireless networks are unprotected from competing signals from other devices such as cordless phones, walkie-talkies, fluorescent lamps, car ignitions, and so forth. By contrast, Ethernet cables do not carry any signals besides data, thus providing excellent signal transmission properties.

Yet another difference concerns multiplexing. Ethernet does not use multiplexing because it uses all the available bandwidth in the medium to transmit signals. This is possible because Ethernet cables are not used for other applications. But wireless LANs share the bandwidth in the air with other users and have to send signals in specified signal bands. Therefore, wireless LANs use multiplexing. To use the available bandwidth efficiently, multiple channels have been defined within the 2.4 and 5.8 GHz bands. Since stations may be transmitting on any of these channels, wireless stations have to scan all the available channels to locate transmissions.

One last factor that makes wireless networks different from Ethernet is that whereas all stations on an Ethernet can hear every transmission, stations at two opposite ends of a wireless LAN may not be able to hear each other. As a result, collision detection may be unsuccessful in wireless LANs. Wireless LANs thus do not use CSMA/CD for medium-access control. Instead, wireless LANs use collision avoidance, and the medium-access control (MAC) mechanism used in wireless networks is called *carrier sense multiple access with collision avoidance* (*CSMA/CA*). What this means is that a waiting wireless station does not start transmitting immediately after a previous transmission ends. This is because the station knows that this is the time when other waiting stations are also likely to try to transmit and therefore the chances of a collision are highest at this time. Wireless stations wait for a certain time after a transmission ends before attempting to transmit data.

The primary implication of all these differences between wireless LANs and Ethernet is that wireless LANs require far greater error-detection capabilities than Ethernet. We will see that this is manifested in the physical layer of wireless LANs, which adds error protection over and above the CRC error detection introduced in Ethernet. This relates to the discussion with the physical layer about the impact of the transmission medium on data-communication technologies. The primary challenge in wireless networks is the increased noise level. The technology response is to increase error protection in wireless LANs compared to Ethernet.

Wi-Fi business models: Moja Internet service on Kenyan buses[3,4]

Egyptian bus service Swvl and Kenyan technology company BRCK are attempting to bring the Internet to the masses in Africa by installing Wi-Fi routers in public transport buses (*matatus*). The rugged Wi-Fi routers include a hard drive to store popular content locally (i.e., on the bus) and also provide Internet connectivity through SIM cards installed on the router. Users pay for usage and view the content while riding on the buses.

The Moja service aims to bring the Internet to the 70–80% of the population in Africa that cannot afford the Internet. The service will have to navigate the challenge of commercial viability. Currently, Facebook picks up the tab for Facebook and WhatsApp messages, and advertising subsidizes Internet access.

Wireless LAN Architecture

Like Ethernet, wireless LANs are a data-link-layer technology. Technology standards for both Ethernet and wireless LANs are defined by the 802 group at IEEE. The IEEE 802 group defines standards for local-area networks. As a result of this common origin, the frame structure of wireless LANs is almost identical to the frame structure of Ethernet. The differences between wireless and wired media discussed in the previous section are handled by differences in the physical layer. To account for the greater need for error detection in wireless media, the physical layer in wireless LANs adds header fields that help the receiver in error detection. We will see these fields later in this section.

To facilitate mobility, the designers of wireless LAN technologies planned wireless LANs in such a way that larger wireless LANs can be built from smaller wireless LANs. The smallest component unit of a wireless LAN is the area covered by a single access point. *A wireless access point is a device that allows wireless hosts to connect to a wired network using wireless LAN technologies such as*

[3] https://techcrunch.com/2019/07/30/startups-brck-and-swvl-partner-on-free-wifi-for-kenyan-ride-hail-buses/ (accessed Feb. 2020).

[4] https://allafrica.com/stories/201811110001.html (accessed Feb. 2020).

Wi-Fi. The area covered by an access point is called a basic service area (BSA). The basic service area and the access point covering that area are collectively known as a basic service set (BSS).

To create a larger network, such as a campus-wide wireless LAN, basic service sets can be connected to each other using any suitable networking technology. This connecting technology is called a *distribution system*. Thus a campus-wide wireless LAN consists of many basic service sets connected to each other through a distribution system. *The portal is the connection point where the entire wireless LAN is connected to the rest of the wired Internet*. This structure is shown in Figure 15-2. The larger campus-wide wireless LAN is called an extended service set. The 802.11 standard does not consider the distribution system to be a part of the extended service set because end users cannot directly connect to the distribution system for wireless access. End users have to use a basic service set for wireless access.

In Figure 15-2, when the first laptop (station 1) wants to send a message to the second laptop (station 4), it creates a MAC frame with the MAC address of the second laptop 2 as the destination MAC address, and sends the message to its access point (station 2). Station 2 will forward the message to station 3 over the distribution system, and finally, station 3 will send the message to the laptop at station 4.

The advantage of composing large wireless LANs from multiple basic service areas is that it facilitates mobility. The extended service set appears to end users as one large LAN. Users can move anywhere within an extended service set and still retain the same IP address and subnet membership. If wireless LANs were not designed as an extended service set, and each access point served as a router, each BSS would become an independent subnet. Each time a user moved from one access point to the next, he would connect to a different subnet. This would potentially

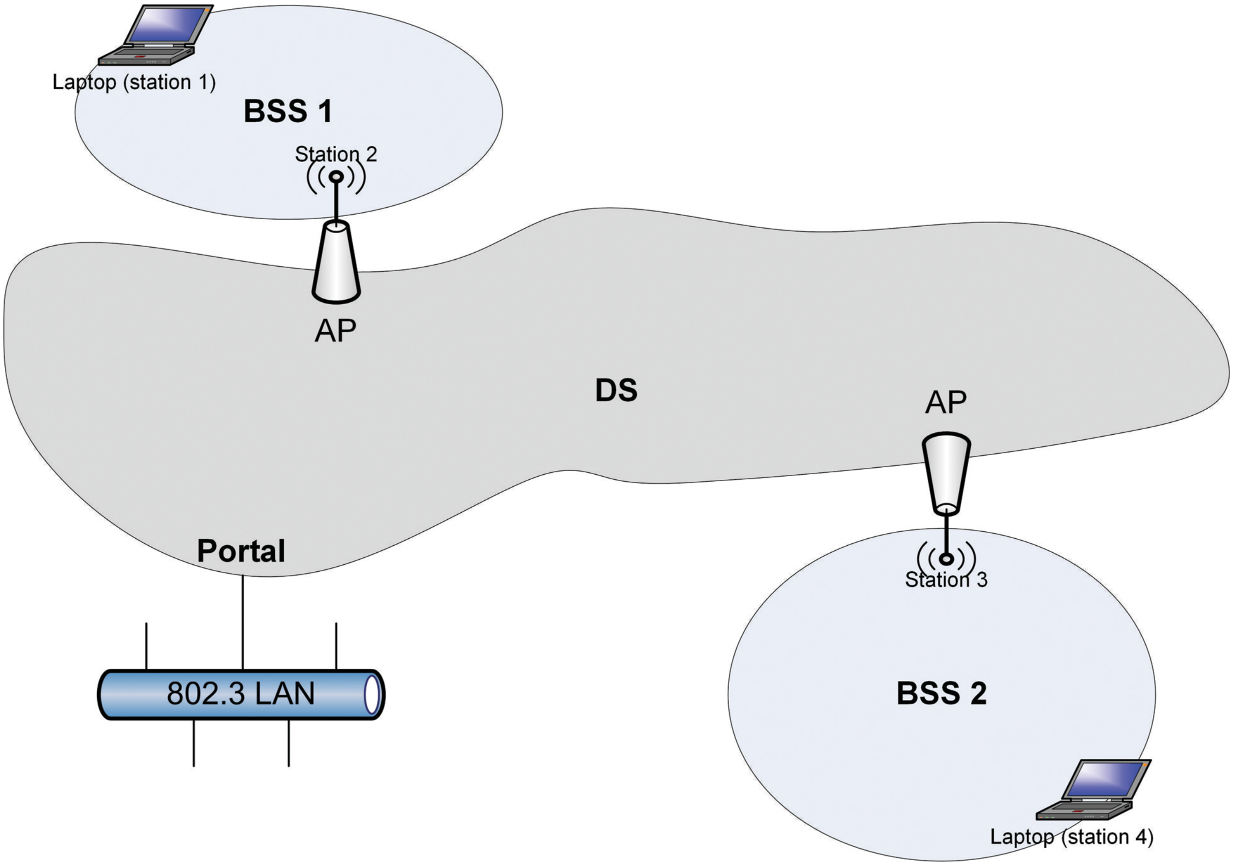


Figure 15-2. Structure of a campus-wide wireless LAN

give him a different IP address and gateway router address. This address reallocation would stop any ongoing transfers and could also potentially disturb the network connectivity of some applications. With the concept of an extended service set, when users move from one access point to another, there is no change to any network setting, and ongoing network transfers can continue without interruption.

Within an organization, basic service sets may be placed as appropriate to deliver the required coverage and reliability. For example, in high-traffic areas, basic service sets may overlap to provide redundancy and to share traffic. If areas requiring network coverage are far from each other, basic service sets may be organized as in Figure 15-2, where they are separated from each other.

On one extended service set, a host needs to get associated with one access point through which it will send and receive messages. The distribution system uses this association information to deliver messages for a host to the correct access point.

The 802.11 standard does not specify how the distribution system should send messages between access points. Any local-area network technology can be used for the purpose. It is common for network administrators to use Ethernet for the distribution system.

The portal acts as the gateway between the extended service set and the rest of the Internet. When a message is sent to a host that is not in the extended service set, the distribution system sends the message to the portal. The portal performs all necessary packet format changes required for the message to be transported on the neighboring network. For example, in Figure 15-2, the portal transforms the outgoing message from the wireless 802.11 frame format to the 802.3 Ethernet frame format.

Most likely, your home network is built using one wireless router. This router acts as the access point as well as the portal. Depending upon the technologies used by your Internet service provider, this router may transform packets from 802.11 format to 802.3 format, or from 802.11 to the WAN frame format used by the ISP.

802.11 Frame MAC-Layer Frame Format

The general MAC-layer frame format for wireless LANs is shown in Figure 15-3. You may note that it has many similarities with the Ethernet frame format. However, the wireless frame is more complex than Ethernet. Extra fields are necessary to identify the basic service set and access points, and to provide reliable transmission in the presence of noise. Also, most wireless LAN packets do not have all the fields in the general frame structure shown in Figure 15-3. The header fields in a captured packet are shown in Figure 15-4.

As in Ethernet, the wireless LAN frame format includes the source and destination MAC addresses, a frame check sequence, and data from the IP layer. These fields perform the same functions as in Ethernet—addressing, error detection, and data transfer. However, we see that wireless LANs also have some additional fields such as frame control, sequence control, QoS control, and duration/ID.

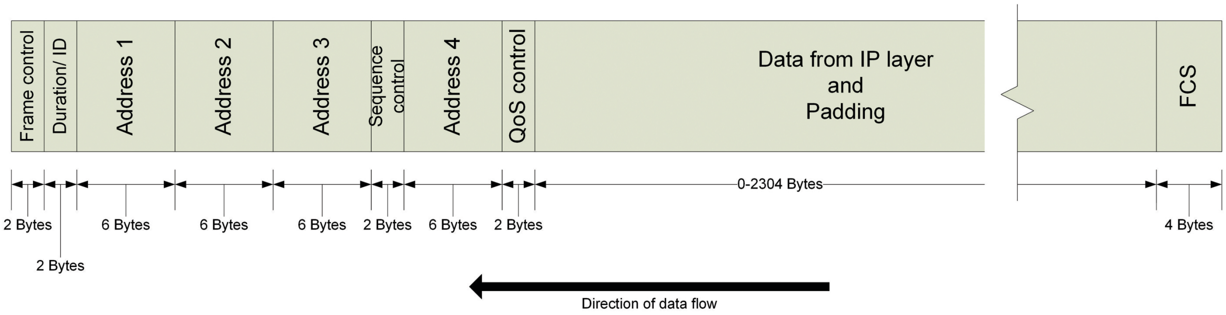
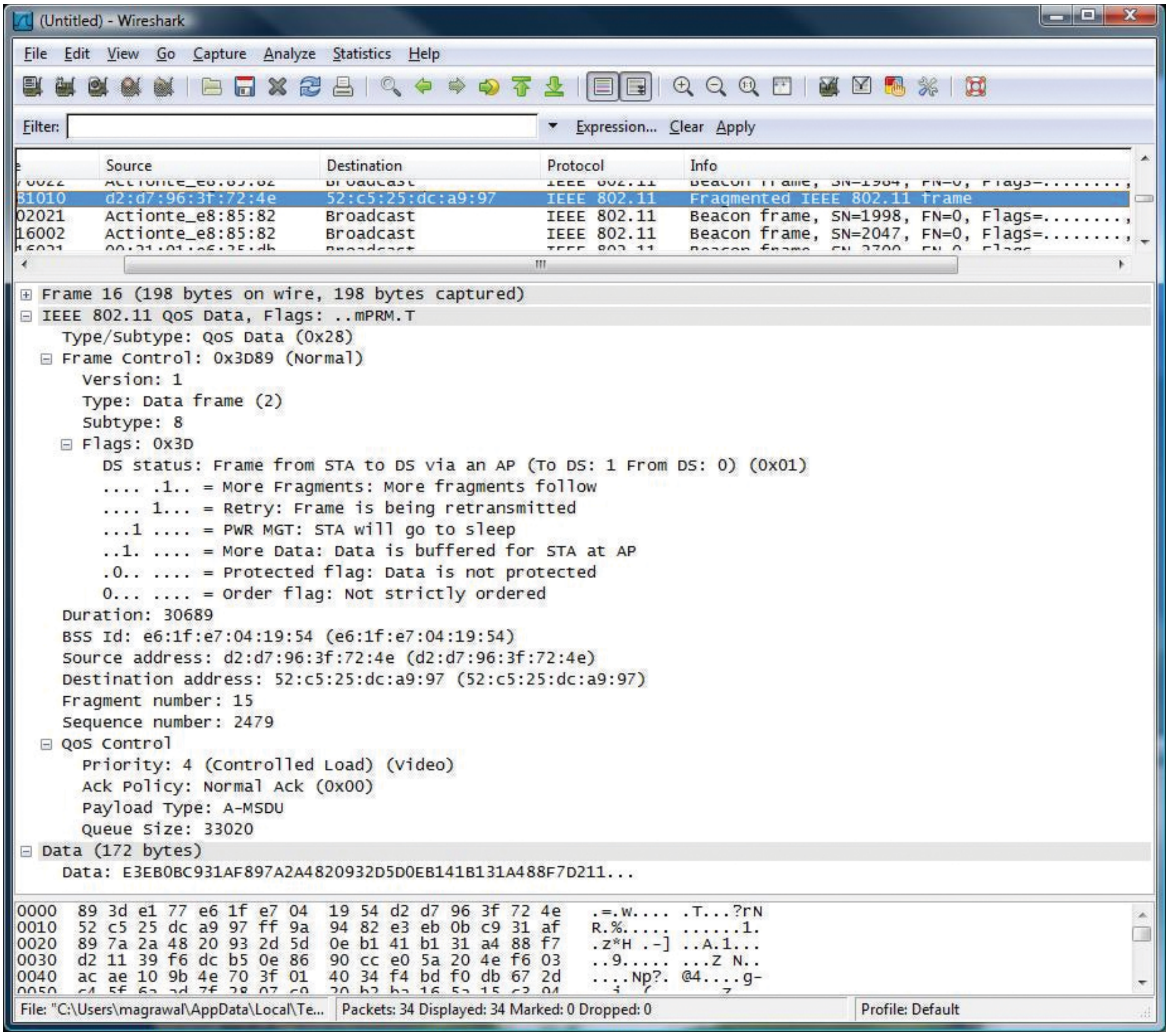


Figure 15-3. 802.11 frame format


(Untitled) - Wireshark
File Edit View Go Capture Analyze Statistics Help
Filter:
Expression... Clear Apply

```
         Source              Destination         Protocol       Info
31010    d2:d7:96:3f:72:4e   52:c5:25:dc:a9:97   IEEE 802.11    Fragmented IEEE 802.11 frame
02021    Actionte_e8:85:82   Broadcast           IEEE 802.11    Beacon frame, SN=1998, FN=0, Flags=........,
16002    Actionte_e8:85:82   Broadcast           IEEE 802.11    Beacon frame, SN=2047, FN=0, Flags=........,
```

```
⊞ Frame 16 (198 bytes on wire, 198 bytes captured)
⊟ IEEE 802.11 QoS Data, Flags: ..mPRM.T
    Type/Subtype: QoS Data (0x28)
  ⊟ Frame Control: 0x3D89 (Normal)
      Version: 1
      Type: Data frame (2)
      Subtype: 8
    ⊟ Flags: 0x3D
        DS status: Frame from STA to DS via an AP (To DS: 1 From DS: 0) (0x01)
        .... .1.. = More Fragments: More fragments follow
        .... 1... = Retry: Frame is being retransmitted
        ...1 .... = PWR MGT: STA will go to sleep
        ..1. .... = More Data: Data is buffered for STA at AP
        .0.. .... = Protected flag: Data is not protected
        0... .... = Order flag: Not strictly ordered
    Duration: 30689
    BSS Id: e6:1f:e7:04:19:54 (e6:1f:e7:04:19:54)
    Source address: d2:d7:96:3f:72:4e (d2:d7:96:3f:72:4e)
    Destination address: 52:c5:25:dc:a9:97 (52:c5:25:dc:a9:97)
    Fragment number: 15
    Sequence number: 2479
  ⊟ QoS Control
      Priority: 4 (Controlled Load) (Video)
      Ack Policy: Normal Ack (0x00)
      Payload Type: A-MSDU
      Queue Size: 33020
⊟ Data (172 bytes)
    Data: E3EB0BC931AF897A2A4820932D5D0EB141B131A488F7D211...
```

```
0000  89 3d e1 77 e6 1f e7 04  19 54 d2 d7 96 3f 72 4e   .=.w.... .T...?rN
0010  52 c5 25 dc a9 97 ff 9a  94 82 e3 eb 0b c9 31 af   R.%..... ......1.
0020  89 7a 2a 48 20 93 2d 5d  0e b1 41 b1 31 a4 88 f7   .z*H .-] ..A.1...
0030  d2 11 39 f6 dc b5 0e 86  90 cc e0 5a 20 4e f6 03   ..9..... ...Z N..
0040  ac ae 10 9b 4e 70 3f 01  40 34 f4 bd f0 db 67 2d   ....Np?. @4....g-
```

File: "C:\Users\magrawal\AppData\Local\Te... Packets: 34 Displayed: 34 Marked: 0 Dropped: 0 Profile: Default

Figure 15-4. Header fields in a captured wireless frame

There are also four possible address fields (recall that Ethernet frames only have two address fields). We also see that the preamble and SFD fields of the Ethernet MAC frame are missing in the wireless LAN frame. The missing fields are the simplest to understand. These are moved to the physical-layer header and retain their positions as the earliest fields of incoming frames. It is the additional MAC-header fields that are more complex to describe and to understand. These additional fields help in identifying the access points and in improving reliability. Their functions are described as follows:

- *Frame control.* This field describes the attributes of the frame. For example, Does the frame carry data, or does it report the status of the network? Is the frame going toward an AP? Is it being sent by an AP?
- *Duration/ID.* This field announces the expected amount of time required to transmit this frame. All listening stations will wait for this duration before attempting to send data.

- *Sending/receiving access-point addresses (address 1–address 4)*. Since wireless LAN packets need to pass through access points, the MAC addresses of the sending or receiving access points are added to the frame as required. Packets leaving an access point have the sending AP address, and packets sent to an AP have the receiving access point address.
- *QoS control*. This field specifies the desired type of service. The available types of service include best effort, voice, and video.

> Since 802.11 frames have different values in the address fields depending upon the type of packet, you might wonder how a receiver knows that a packet is addressed to it. To deal with this issue, the address to be used for address matching is always placed in the address 1 field. If a station finds that the value in the address 1 field is the same as its own MAC address, it knows that the packet is addressed to it.

802.11 Frame Physical-Layer Format

Recall that the physical layer in Ethernet added no header fields to the frame. It simply converted the frame to a signal. However, the wireless LAN physical layer does add fields to the frame header. The wireless LAN physical-layer header is shown in Figure 15-5. The primary function of the physical-layer header is to add error protection to the frame header. It also specifies the data rate being used in the transmission. Figure 15-4 shows the header fields in a captured wireless packet.[5]

Popular 802.11 Technologies

Three technologies are currently specified for wireless LANs—802.11a, 802.11b, and 802.11g. 802.11a and 802.11b were specified in 1999, and 802.11g was specified in 2003. A fourth technology, 802.11n, was recently specified to increase the speed and range of wireless networks.

802.11b and 802.11g operate in the 2.4 GHz ISM band. 802.11b is a simpler technology and can support a data rate of up to 11 Mbps. The signal modulation techniques used in 802.11b are defined in Chapters 14, 15, and 18 of the 802.11 standard.[6] 802.11b uses the direct sequence spread spectrum (DSSS) modulation technique. 802.11g adds the orthogonal frequency division multiplexing (OFDM) modulation technique to obtain higher data rates than 802.11b. 802.11g can support data rates up to 54.11 Mbps.

802.11a operates in the 5.5 GHz ISM band. It was the first 802.11 technology to use OFDM to achieve data rates of up to 54 Mbps. OFDM is defined in Chapter 17 of the 802.11 standard. 802.11a has many features, such as a greater number of channels, which should make it the wireless technology of choice. Unfortunately, the signals in the 5.5 GHz band do not travel as far as

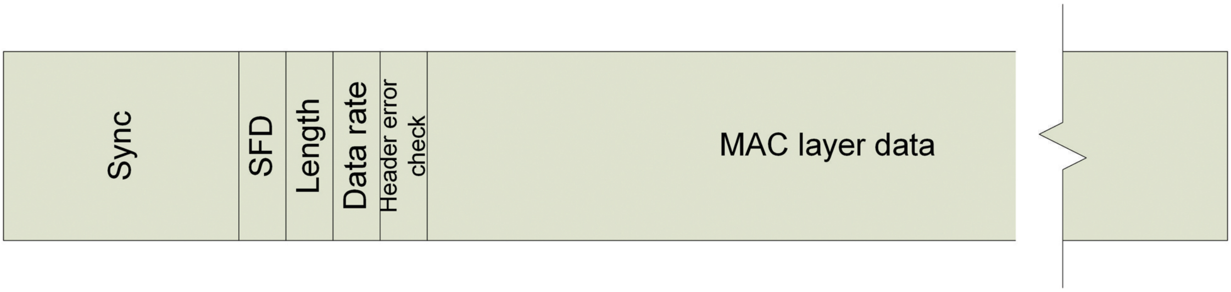


Figure 15-5. Wireless LAN physical-layer header

[5] At this point, it is highly recommended to check out the quick technical tutorial on 802.11 by Pablo Brenner, “A Technical Tutorial on the IEEE 802.11 Protocol,” http://www.sss-mag.com/pdf/802_11tut.pdf (accessed Dec. 2015).

[6] The 802.11 standard is accessible from the IEEE website.

signals at the same power in the 2.4 GHz band. As a result, each 802.11a access point covers a slightly smaller area than a 802.11b/g access point. 802.11b/g is thus economically more efficient than 802.11a for wireless coverage. Hence 802.11b/g technologies are more popular than 802.11a.

802.11n

The 802.11n standard was finalized in September 2009. The primary goal of 802.11n is to provide a data rate of up to 600 Mbps. 802.11n also aims to provide wider coverage than 802.11b/g, so that a single access point can cover more than twice the area covered by a 802.11b/g access point. 802.11n operates on both the 2.4 and 5 GHz bands and uses OFDM to maximize data rates.

The primary innovation of 802.11n is multiple-input, multiple-output signal transmission. 802.11n access points and receivers use multiple antennas. Each antenna pair creates an independent data channel. Each 802.11n transmission thus may be seen as multiple 802.11a/b/g transmissions.

Apart from the technical innovations introduced in 802.11n, the technology is also very interesting because it allows us to peek into the process by which new data communication technologies are developed and standardized. The IEEE website maintains a log of the standardization process for 802.11n.[7] According to the logs, four candidate technologies were proposed for 802.11n in November 2004. The TGn technology that will be used in 802.11n obtained the required level of support from the standards committee in March 2007. In late 2008–early 2009, the technical and editorial issues in the standard were being fixed. Since technology vendors already knew most details of the 802.11n technology, and only editorial changes were expected in the standards document, draft 802.11n products became available in the market in 2008, even before the final standards document had been published.

802.11ac

The 802.11ac standard was designed to improve upon 802.11n technologies to support the transmission of multiple channels of HD video streams in homes and support many users per access point in enterprises. The standard was finalized in January 2014. 802.11ac achieves these higher data rate design objectives primarily by making three refinements to 802.11n: (1) moving from the 2.4 GHz band to the 5 GHz band, (2) refining technologies such as MIMO introduced with 802.11n, and (3) capitalizing on the improved sensitivity of affordable electronics by packing in more information over the same bandwidth.

Moving from the 2.4 GHz band to the 5 GHz band is one of the primary drivers of improved 802.11ac performance. You may recall from the discussion on the physical layer that data rates are directly proportional to bandwidth. The 2.4Ghz band, used by 802.11n, includes frequencies in the range 2.4 GHz–2.5Ghz, for a total of 100 MHz. In conventional use, these are split into three non-overlapping channels of 22 MHz each. The 5 GHz band includes frequencies ranging from 5.15 GHz to 5.35 GHz, for a total of 200 MHz, or twice the bandwidth of the 2.4 GHz band. 802.11ac can split this band into two non-overlapping bands of 80 MHz each, providing almost four times the bandwidth per channel compared to 802.11n.

Multiple-input, multiple-output (MIMO) is a very recent development (in the 2000s) that allows transmitters to send multiple streams of data at the same frequency to the same receiver, but through different paths so that the conflicting signals never overlap. 802.11n allowed up to four MIMO paths, and 802.11ac allows up to eight MIMO paths. By doubling the number of channels, where such MIMO benefits are possible, 802.11ac can provide twice the overall data capacity as 802.11n.

Finally, 802.11ac tries to talk faster than 802.11n (256 QAM instead of 64 QAM), packing more information over a single transmission than 802.11n.

[7] http://grouper.ieee.org/groups/802/11/Reports/tgn_update.htm (accessed Feb. 2020).

The benefits of 802.11ac come with trade-offs. The total transmission power in ISM bands is regulated. When an 802.11ac transmitter uses the 80 MHz channel, the available power is distributed over four times the bandwidth compared to a 22 MHz channel. Reduced power means that the signal covers a shorter distance before it becomes indistinguishable from noise. Reduced power also means that the fast-talking technique (256 QAM) can only be used when the transmitter and receiver are in fairly close proximity (typically about 15 feet), without any obstacles. To visualize this, imagine talking to a friend. As you try talking faster, your friend will have greater difficulty keeping up, and the same level of noise will be more disruptive than when you were speaking slowly.

For these reasons, 802.11ac is most useful in high-density areas such as lecture halls and cafeterias, where users can have a clear line-of-sight to the 802.11ac access point. These locations were very challenging to 802.11a/b/g technologies anyway because the 2.4 GHz band used by these technologies only supported three slower non-overlapping channels in a given space. By supporting two high-speed, nonoverlapping channels in the same space, 802.11ac can allow network designers to serve high-density areas more conveniently.

Most data communication technologies go through this process of standardization. The standardization process begins with the recognition of a need. In the case of 802.11n and 802.11ac, it was the need for a high-speed, long-range wireless technology. A neutral standards-making body takes the lead in organizing an expert group to identify technical solutions that meet the need. In the case of 802.11, this body was the IEEE, which has led the standardization of all local-area network standards. Any interested persons from universities, technical companies, and even members of the public can become members of the group. The group develops proposals and votes on them until one technology solution receives overwhelming support. Finally, the technology is described in adequate detail in a standards document so that any interested vendor can use the standards document to create an implementation of the standard. The standardization process assures users that equipment they buy from one vendor will always be compatible with equipment sold by other vendors. In the case of 802.11, for example, you can buy a network interface card from Broadcom and rest assured that it will work with access points sold by Linksys.

It is also hoped that the formal standardization process will lead to the adoption of the best possible technical solutions to meet requirements.

DLNA

The ubiquity of wireless LANs has created a strong desire among users to exchange content among home devices, for example, to stream music from smartphones to a stereo speaker, watch home security-camera footage on smartphones, and so forth. The digital living network alliance (DLNA) is an industry consortium aiming to achieve this interoperability.

Personal Area Networks (The 802.15 Series)

The previous section described wireless local-area networks. These networks use an access point and provide high-speed connectivity to any hosts within a radius of about 100 feet. Hosts typically use wireless LANs for Internet access.

While wireless LANs are extremely useful, there are many connectivity applications where wireless access would be very useful but where Internet access is not necessary. An important example is replacing the short wires on desktops that are used for data transfer at low speeds, for example, to connect keyboards and mice to the desktop. This is where personal-area networks (PANs) come in. PANs, specified by the IEEE 802.15 standard, are designed to remove these wires that clutter desktops and make many devices cumbersome to use. *Personal-area networks are computer networks designed for data transmission among devices in close proximity, typically owned by the same individual.*

Personal-area networks like Bluetooth have been developed to provide communication over short distances, usually within 30 feet. This distance is sometimes called the personal operating space because people and devices within this range are usually in visual range. By limiting itself to this range, Bluetooth is designed to serve a small group of participating devices, usually carried by one person. Apart from keyboards and mice, other devices that use Bluetooth include cell phone headsets and digital cameras (to transfer pictures to computers).

The focus of Bluetooth is to develop extremely small, inexpensive, and low-power connectivity solutions. This makes it easy for electronics manufacturers to add Bluetooth capabilities to virtually any electronic device—even headsets and telephones—which are not traditionally considered computers. Since many of these devices are very small and can carry only a limited amount of battery power, power efficiency is an extremely important requirement for Bluetooth. Since signal transmission requires power and more power is needed to send signals to greater distances, power efficiency concerns are the reason why Bluetooth devices have very short communication ranges.

Bluetooth operates in the 2.4 GHz ISM band, the same as 802.11b/g wireless LANs. Bluetooth is designed to offer data rates of up to 1 Mbps. This is much slower than the 11/54 Mbps offered by 802.11 LANs. But 1 Mbps is adequate for applications such as keyboards and headsets that use Bluetooth. Bluetooth uses frequency-hopping spread spectrum (FHSS) modulation for signal transmission. This is shown in Figure 15-6. The sender and receiver communicate at a predefined sequence of frequencies. Any channel conflicts only disturb individual transmission blocks, and the bulk of the communication generally proceeds without interruption. This makes Bluetooth especially suitable for voice, since we generally don't notice brief interruptions.

Though Bluetooth has some similarities with wireless LANs, there are some important distinctions between the two, such as:

1. Wireless LANs are largely used by computing devices such as laptops. Bluetooth is designed to be used by any electronic device to communicate with any other Bluetooth-capable electronic device.
2. Wireless LANs are typically used to obtain Internet connectivity. Bluetooth is typically used to connect to other nearby devices—for example, a keyboard to a desktop or a headset to a cell phone. As a result, while a high data rate is a very important requirement for wireless LANs, it is less important for Bluetooth.

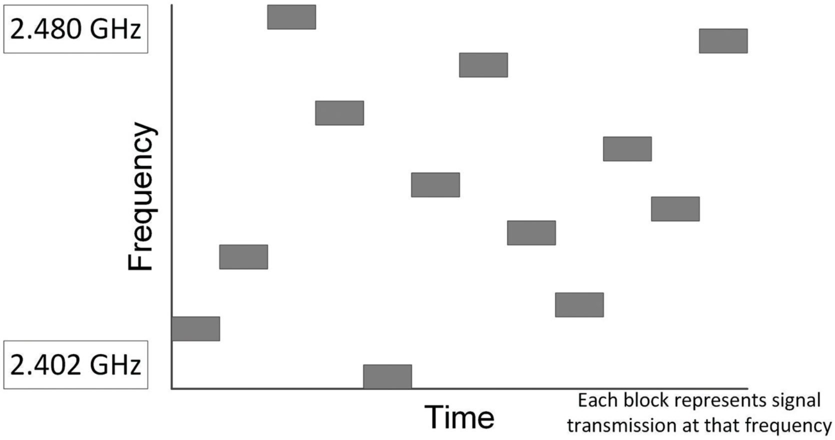


Figure 15-6. Bluetooth frequency hopping transmission

3. Wireless LANs require an infrastructure of access points. Bluetooth requires no such infrastructure. In fact, each Bluetooth device is capable of acting as a Bluetooth access point. Bluetooth devices automatically locate other Bluetooth devices in their vicinity.
4. Devices using wireless LANs are typically connected to power outlets, and there are no special power-efficiency concerns in wireless LANs. Bluetooth devices are almost always driven by battery power, and long battery life is an important concern for Bluetooth.
5. Finally, since Bluetooth is often used by devices that are relatively inexpensive, it is important for Bluetooth solutions to be extremely inexpensive, generally costing less than $10.

Swedish origins of the term *Bluetooth*

The development of Bluetooth was initiated in 1994 by the Swedish mobile phone company Ericsson, to help laptops make calls using cell phones. The rather unusual name comes from King Harald "Bluetooth" Blaatand II of Denmark (940–981 CE). The nickname came from the king's love for blueberries, which eventually stained his teeth. King Bluetooth unified Denmark and Norway during his reign. Ericsson hoped that the technology would similarly unite the communication and computing industries.[8]

Ericsson also created another well-known technology product in widespread use today—the MySQL database. The motivation was to create a small database engine to store contact information on cell phones.

Bluetooth Architecture

The basic unit of a Bluetooth network is the piconet. *A piconet is a collection of devices connected to each other using Bluetooth*. On the piconet, one master device connects with up to seven other active slave devices. A Bluetooth piconet serves a function similar to the basic service set (BSS) in 802.11 LANs. However, whereas an 802.11 BSS has a dedicated device called an access point that performs various management functions in the BSS, any device in a Bluetooth piconet can perform the management functions of a piconet. The device that performs this function is called the *master*. All other devices in the piconet are called *slaves*. The master provides a synchronization clock that helps all other devices in the piconet remain in sync with each other. Whereas a device may be a slave on multiple piconets at the same time, it can only be a master on one piconet at a time.

Many piconets may coexist in the same location. All the co-located piconets are called a *scatternet*. Think of a gathering of tech-savvy students in a classroom, with many students carrying Bluetooth-capable cell phones and music players. Each such student forms a piconet, and the entire classroom becomes a scatternet. Devices connected to two different piconets in a scatternet do not have to route packets between the piconets.

Piconets are the personal-area equivalent of the basic service set. However, there are some major differences between basic service sets and piconets. The basic service set has a fixed location defined by the geographic area covered by the signals from the access point. The piconet, on the other hand, has no defined location. The piconet exists wherever the Bluetooth devices go. For example, the Bluetooth devices in a car form a piconet. As the car hurtles down the highway, the piconet moves along with it. Also, whereas the basic service set can support tens or even hundreds of devices, a piconet can connect at most eight devices.

To enable interference-free communication within co-located piconets, Bluetooth has mechanisms that enable each piconet to operate on a different physical channel. Recall from Chapter 2 that only one signal may be transmitted at a given frequency at a given location. Since all Bluetooth

[8] Elias M. Awad, *Electronic Commerce: From Vision to Fulfillment*, 3rd ed. (Pearson, 2007).

transmissions are in the 2.4 GHz ISM band, Bluetooth needs to create mechanisms whereby multiple transmissions can occur at the same frequency.

Bluetooth creates multiple communication channels at the same frequency by enabling devices to transmit at different time slots. Though only one device may transmit at a given time in a given location at a specified frequency, different devices may transmit at different time slots on the same frequency. Stations keep hopping from frequency to frequency in a systematic manner within the 2.4 GHz ISM band. This is called *frequency hopping*. Devices in each piconet use a different hopping sequence, thereby reducing the chance for collision. Finally, to maintain confidentiality, devices in each piconet use a different access code and header encoding to ensure that even if their signals are received by devices in other piconets, the data is unreadable.

Bluetooth Frame Structure

The structure of Bluetooth frames is shown in Figure 15-7. The payload header is similar to the MAC header of 802.11 wireless LANs. The packet header is analogous to the 802.11 physical-layer header. The channel access code is unique to Bluetooth.

You may note that the Bluetooth frame has some fields such as flow control, sequence number, and channel access code that are absent in wireless LANs. These fields help Bluetooth devices operate within a scatternet without interfering with each other. These fields also help Bluetooth provide reliable signal transmission for voice applications.

Bluetooth Device Discovery

A very special capability of Bluetooth is device discovery. Two Bluetooth devices in close proximity to each other will automatically discover each other. If you have used a Bluetooth-enabled keyboard, you may have noticed this behavior when you bring the keyboard near your desktop. Your computer becomes aware of the presence of the keyboard and instantly pairs up with it. Device discovery makes Bluetooth extremely user-friendly and eliminates configuration-related problems for end users. The devices seem to become aware of each other as if by magic.

To enable device discovery, Bluetooth defines a special channel for inquiry requests and responses. Devices that are looking for nearby devices are called *inquiring devices*. Inquiring devices send out inquiry requests on the special inquiry channel. Devices willing to be found are called discoverable devices. Discoverable devices listen on the inquiry channel for inquiry requests and respond to these requests. Once the two devices become aware of each other, the inquiry procedure ends and the connection procedure begins.

Device discovery—LANs versus PANs

Device discovery is not needed in wireless LANs because, in most cases, user intervention is necessary to determine the LAN to connect to. There are also security issues associated with wireless LAN membership, as a result of which, network administrators like to have control over the users who have access to the LAN. However, once a laptop successfully joins a wireless LAN, most laptops offer to join the network in the future without user intervention. Therefore, subsequent wireless LAN connections do operate in a manner that is similar to the device discovery procedure.

Channel access code	**Packet header** Includes flow control, seq number, header error check (HEC)	**Payload header** Includes data length, transport link ID	**Payload** User data such as IP packet, possibly segmented by Bluetooth	**CRC**

Figure 15-7. Bluetooth frame structure

In the connection procedure, one of the devices must be willing to receive a connection request from the other device. This device is called the *connectable device*. The connecting device sends a connection request to the connectable device on a connection channel specified by the connectable device. According to the Bluetooth standard, the device initiating the connection becomes the master for the connection.

Device discovery in other contexts—the MH 370 disaster[9]

This principle of requiring a discoverable device sending out a specific signal whose properties are known in advance is common across most contexts. A notable incident where this procedure received worldwide attention was associated with the discovery of the flight data recorder of the doomed flight MH 370, which vanished without a trace on March 8, 2014. Immediately following the accident, search efforts focused on detecting the signals emanating from the flight data recorders (black boxes) at 37.5kHz. The black boxes are designed to send out this signal for 30 days at enough strength so they can be detected from a distance of up to a mile. Unfortunately, the devices could not be located within the appointed time, after which the batteries eventually would have died, ending the signal transmission.

If you think about it, you may notice that the Bluetooth device discovery and connection procedure is almost identical to the connection procedures used on social networks such as LinkedIn or Facebook. The websites act as the inquiry channel. Without websites such as Facebook or LinkedIn, you would not know where to search for your old friends. People willing to be found create profiles. People with profiles become discoverable. People searching for friends inquire of (search) the social-network site to see if their friend has a profile on the site. If the profile is found, the inquiry procedure is over.

For the connection procedure, you need to send a special connection-request message (friend request or invitation) to your friend. If the friend is connectable (responds favorably) and accepts your invitation, the two of you become connected.

NFC

Another technology for communication among proximate devices is called near-field communication (NFC). NFC is intended for communication among devices closer than 10 centimeters from each other. Example applications include smartphone payments and smartphone check-in.

WLAN and WPAN Coexistence

Wireless LANs and Bluetooth operate at the same ISM band (2.45 GHz). Therefore, there is a high possibility that the signals from the two technologies may interfere with each other. Since Bluetooth is the more recent of the two technologies, it is only natural that the designers of Bluetooth had the responsibility of ensuring that Bluetooth minimized interference with the existing wireless LAN technology. Therefore, the Bluetooth standard defines two mechanisms to minimize interference between 802.11 and 802.15.

The first of these two mechanisms is collaborative and occurs where Bluetooth and 802.11 communicate with each other. This is possible when both 802.11 and 802.15 are present on one device, such as a laptop with both 802.11 and 802.15 capability. In the collaborative mechanism, Bluetooth avoids transmission during an ongoing 802.11 transmission. Alternately, Bluetooth transmits signals on a different channel than the channel on which the ongoing 802.11 communication is taking place.

[9] https://en.wikipedia.org/wiki/Malaysia_Airlines_Flight_370 (accessed Feb. 2020).

The second mechanism is noncollaborative. The noncollaborative method is used when communication between 802.11 and 802.15 systems is not possible. For example, Bluetooth keyboards do not have 802.11 capability, and the Bluetooth system on the keyboard has no way to collaborate with 802.11. In the noncollaborative method, the 802.15 system senses the medium before transmitting. It tries to find a channel in the 2.45 GHz ISM band that is not very busy and transmits signals on that channel.

Bluetooth Categories

The early Bluetooth specification supported data rates of up to 1 Mbps. However, as Bluetooth grew in popularity, newer applications for the technology were identified, each with slightly different requirements. As a result, additional categories of Bluetooth have been defined as subcategories of Bluetooth. The traditional Bluetooth specification is now called 802.15.1.

The first additional Bluetooth category supports higher data rates. It is useful to be able to transfer pictures from digital cameras to computers without the need to take out the picture card or connect the camera to the computer using a wire. Since digital images can get very large (a compressed picture from a 4-megapixel camera is about 1.5 MB), high data rates are very useful for image transfers. Accordingly, the high-data-rate Bluetooth specification, 802.15.3, supports data transfer rates of up to 25 Mbps. This is accomplished by improving the efficiency at which the physical layer encodes data into signals, so that more data can be sent using the same bandwidth. Currently, higher data rates are becoming possible by integrating Wi-Fi with Bluetooth.

The second additional category of Bluetooth is for remote-control devices such as the remote controls for TVs, door openers, fans, lights, and so forth. These devices need very low data rates because, after all, the only data these remotes send is "ON" or "OFF" or "CHANNEL = 2." However, we expect the batteries in remote controls to work for at least a couple of years. A unique feature of remote controls is that they are idle most of the time, used only for a few seconds in a day to operate remote devices. In almost every case, it is also true that the remote control does not need to be a connectable device. Only the controlled device, which usually is connected to a power outlet, needs to be connectable. To meet the requirements of remote controls, the 802.15.4 standard supports very low data rates, up to 250 Kbps. But to achieve long battery life, 802.15.4 devices are not in a connectable state (are switched off) when they are not being used. As a result, they do not lose power by scanning the medium, listening for other devices that may be interested in connecting to them.

RFID (Radio Frequency Identification)

Thus far, we have seen how ISM bands are used for communication. In recent years, ISM bands are seeing another important use—asset tracking—using a technology called RFID. *RFID stands for radio-frequency identification and is the method to read ID information using wireless signals*. When used for identification, assets are tagged with RFID tags. Each tag contains a chip and an antenna. At any location where the asset needs to be identified, RFID readers transmit signals (energy) at ISM frequencies. The antenna in the tag draws power from the RFID reader's signal and uses this energy to activate the chip. The chip essentially contains an identifier, similar to the way a browser cookie contains an identifier (Figure 6-7). The activated chip also responds to the reader's request by transmitting the identifier back to the reader. The reader, in turn, can transmit the ID to other larger systems, to offer interesting and value-added services to users.

RFIDs are similar to bar code scanners, with the difference that whereas bar code scanners need to be in the line of sight of the code, an RFID tag can work as long as it is within range of the reader. This greatly simplifies logistics where large volumes of items are to be tracked, for example at warehouses. RFID is therefore increasingly popular in the shipping industry. Toll tags to automatically pay tolls are another example of an RFID system. The use of ISM bands enables anyone to deploy RFID wherever it is convenient without the complexities of licensing and approvals. US Passports have had RFID chips in the cover for several years now.

Airlines have begun to see benefits from RFID adoption. In 2018, the industry consortium, IATA (International Air Transport Association), resolved that all checked baggage would be tracked at four key points in its journey. This has greatly reduced lost baggage across the world. IATA is also likely to mandate that all baggage tags should have an RFID tag. When this is implemented, travelers will no longer need to add baggage labels to checked luggage since the RFID tags in the bags will offer their own identification. This can further expedite international travel.

Summary

Wireless networks enable mobility and have become extremely popular in homes and businesses. Wireless networks enable inexpensive Internet access in many homes and offices. Most wireless networking technologies use ISM frequency bands. Frequencies in these bands can be used without cost or licensing restrictions. To meet the requirements of different applications that benefit from wireless access, three different categories of wireless technologies have been defined.

The best-known wireless technology is the 802.11 wireless LAN technology, sometimes also called Wi-Fi. Wireless LANs use access points to provide high-speed Internet access within a range of about 300 feet from the access point. Multiple access points can be connected using a distribution system to provide wireless LAN coverage over an arbitrarily large area. There are three wireless LAN standards popular today: 802.11a, 802.11b, and 802.11g. 802.11 technologies can provide network connection speeds of up to 54 Mbps. The newest wireless LAN technology, 802.11n, is expected to provide data rates of up to 600 Mbps.

The second category of wireless networks is personal area networks, better known as Bluetooth. These are standardized by the IEEE as 802.15 networks. Bluetooth provides data rates of up to 1 Mbps within a radius of about 30 feet. Bluetooth helps eliminate wire clutter created by peripheral devices such as keyboards and mice. The primary design goal of Bluetooth is to provide adequate data connectivity while maximizing battery life and minimizing costs.

The last category of wireless networks is IEEE 802.16 wireless metropolitan area networks, also known as WiMAX. These networks can substitute for cable and DSL connections and provide high-speed connectivity to fixed receivers at large distances. Mobile functionality has recently been added to WiMAX. WiMAX is likely to be offered as a paid service in many metro areas in the coming years.

About the Colophon

The line in the colophon was uttered by the fox to the little prince in French aviator Antoine de Saint-Exupéry's most famous novella, *The Little Prince*. The novella is believed to have been inspired by the aviator's real-life experiences in the Sahara desert. It is one of the best-selling books ever—80 million copies—and has been translated into more than 180 languages. Though written and illustrated for children, the book makes many thoughtful observations about life. One of the best known of these is "On ne voit bien qu'avec le cœur. L'essentiel est invisible pour les yeux," which translates as "It is only with the heart that one can see rightly. What is essential is invisible to the eye."

Computers and networks have no heart. At the heart of their operations, though, are properties of the universe that are invisible to the eye. Computer networks do not need a visible medium to transport data. The properties required from nature to support signal transmission are invisible to the naked eye. The invisible outer space can transport data just as effectively as visible wired networks. In an earlier age, this ability of the universe to carry electronic signals was given a name—Ether.

Review Questions

1. What are *wireless networks*? Why are they useful?
2. Some cities took up projects to set up wireless LANs all over the city. Read about the project taken up by one such city. Was the project a success? Why or why not?
3. What are some of the concerns with using wireless networks?
4. What are *ISM frequency bands*? Why are they useful?
5. What are some differences between wired and wireless LANs? How do they impact the design of the wireless LAN header?
6. What is a *basic service set*? A *basic service area*?
7. What is an *access point*? What are some reasons why you would prefer access points to wireless routers when creating a wireless network in your organization?
8. What is a *distribution system* in wireless LANs?
9. What is an *extended service set*?
10. What is a *portal* in a wireless LAN?
11. What are some differences between the physical layers in wireless and wired LANs?
12. What are the common wireless LAN categories? What are the important differences between them?
13. What is *802.11n*? What are some likely advantages of 802.11n over the traditional wireless LANs? How does 802.11n obtain these advantages?
14. What are *personal area networks*? How are they different from LANs?
15. What are some important characteristics of Bluetooth?
16. What is a *piconet*? What are some differences between a piconet and a basic service set?
17. What are *master* and *slave* devices in a piconet?
18. What is a *scatternet*?
19. What are some advantages of having distinct physical channels in Bluetooth?
20. Why is device discovery useful in Bluetooth? How is device discovery accomplished? Why is device discovery not needed in wireless LANs?
21. Describe the mechanisms that have been defined for WLANs and WPANs to coexist at the same frequency bands without interfering with each other.
22. What are the different categories of Bluetooth? What are they used for?
23. What is RFID? What are its primary uses?
24. What are some ISM frequency bands used by RFID? (Note: This question aims to encourage you to look up this information online and become more aware of the technology.)
25. What are some important differences between RFID and barcodes for asset tracking?

EXAMPLE CASE: *The Oil Industry*

When gas prices rose rapidly in recent years, oil companies earned record profits during 2007–2008. When the economy slowed down in late 2008, oil companies experienced slumping demand for the first time in years. Improved supply-chain management using computer networks is helping oil companies deal with these boom-and-bust cycles and also to lower oil prices.

Integrated petroleum firms are some of the largest companies in the world. The industry had sales of $1.99 trillion in 2008, comparable to US government tax revenues of $2.54 trillion in 2008. The industry is very volatile, however, with estimated net sales in 2009 of $1.28 trillion—a drop of more than 35% compared to 2008, due to falling prices and reduced demand resulting from weaknesses in the economy. The industry also has very low profit margins, with net margins of only 8.1% in 2008. With political sensitivities and customer behavior limiting price increases, the way to improve profitability in the industry is to lower costs.

We saw in the case of Walmart that utilizing point-of-sale data to optimize distribution and manufacturing can eliminate inventory accumulation, reduce waste, and lower costs. In most industries, this requires information sharing between many companies that complete the supply chain. But the petroleum industry has a unique advantage. It is probably the only industry still dominated by vertically integrated firms. Vertical integration refers to a single firm controlling all aspects of a product's manufacture, from raw materials to distribution. The major oil companies such as Exxon-Mobil, British Petroleum, Shell, and Chevron own or control all factors of production starting from the oil fields where oil is drilled from the ground to the gas stations where drivers fill their cars.

Whereas retailers such as Walmart have to work out legal and technical barriers to protect their intellectual property from being stolen by business partners with whom they share information, vertically integrated oil companies have a unique advantage. With the right systems, they can share point-of-sale data from company-owned gas stations all the way up the supply chain to company-owned or company-controlled oil rigs and refineries, thereby optimizing drilling, distribution, and storage to lower costs. Publicly available information indicates that Chevron is a leader in its industry.

A large distribution company such as Chevron has a number of places where unnecessary costs can add up. For safety reasons, containers such as oil tankers and trucks do not deliver unless they can be emptied fully. Ships waiting for storage space to accept their crude can pay port charges as high as $30,000 per day. Delivery charges for a truckload of gas can be as high as $150. If a truck returns because the gas station is not empty enough to receive the entire load of fuel (retain), the company incurs an unnecessary expense. On the other hand, if the gas station is out of fuel when a customer arrives (run-out), Chevron could lose the customer and earn a bad reputation.

Chevron uses wireless and satellite networks in many parts of the company to manage its supply chain. Many of its gas stations are linked by a satellite network to a central dispatch center. These stations have electronic level monitors in the underground gas tanks to monitor fuel levels in real time. A wired network in the gas station transports the data to the satellite dish on top of the station from where the data is sent to the dispatch station. Using this real-time data, the dispatch station is able to optimally schedule fuel deliveries from terminals located on the outskirts of major metros, minimizing unsuccessful deliveries (retains) and stock-outs (run-outs). Chevron can also use the computer network to monitor fuel levels in the terminal tanks to schedule oil tanker deliveries so that tankers do not have to wait at ports to make deliveries.

The popular satellite communication technology for retail data-communication applications is called *VSAT*, or *very small aperture terminals*. This technology

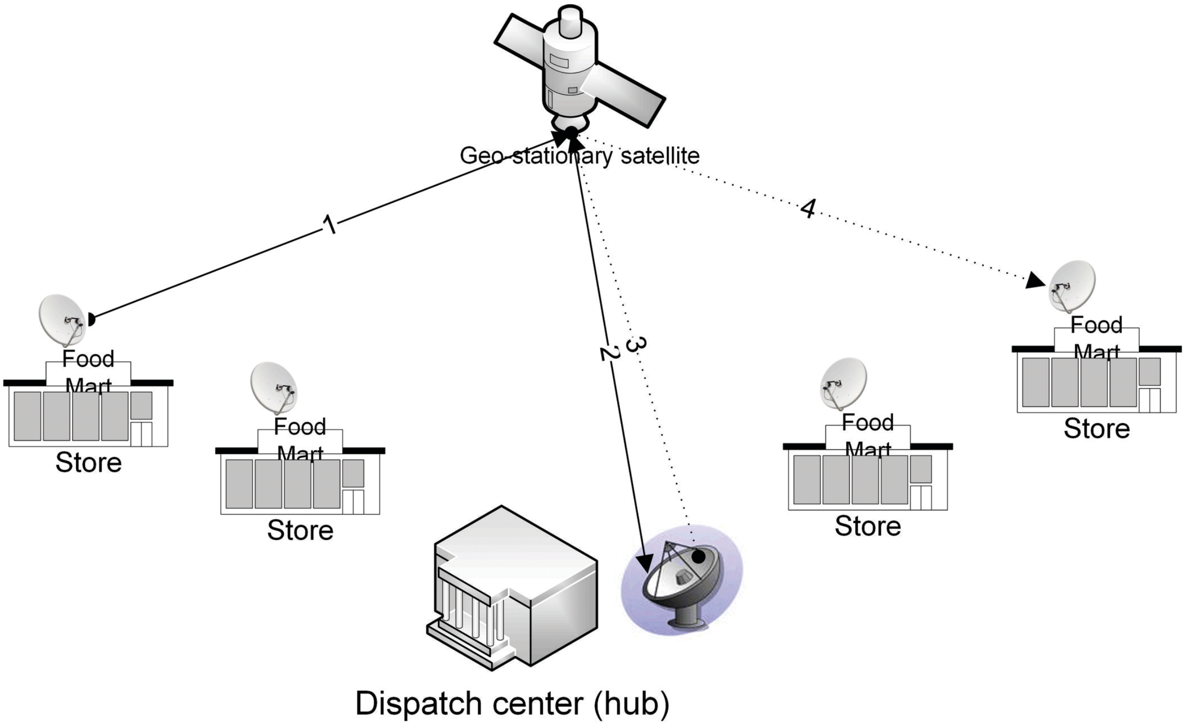


Figure 15-8. VSAT system operation

uses small-sized dish transmitter-receivers communicating with a central hub with a large dish transmitter-receiver through a geostationary satellite. The high-performance antenna at the hub improves the data-transmission capabilities of the overall network. The low-performance requirements from antennas at each retail outlet such as gas stations reduce the costs of the overall system. The data-transfer mechanism in a VSAT system is shown in Figure 15-9. All data exchange occurs through the hub. The geostationary satellite acts as a broadcast medium for the network. If a station wants to send data to another station in a VSAT system, the transmission has to be routed through the hub to gain adequate signal strength, making two hops through the satellite.

The satellite network has helped Chevron move toward being a demand-driven company where upstream activities are performed only in response to observed demand. The company has been moving in this direction since 1997. In 2000, demand-driven operations were improving profits by almost 15% in parts of the company where they were being applied.

Chevron and other oil companies also use computer networks in other parts of their businesses. By linking equipment to Internet-based networks, these companies are able to centralize all operational information to a central monitoring location. Improved monitoring helps companies respond quickly to problems, preventing fires and other hazards, and improving the uptime of pipelines and storage tanks. Industry experts believe that the benefit from these efforts is equivalent to adding 2% to 5% refining capacity.

Wireless networks are particularly useful to Chevron in its drilling operations. As new oilfields are becoming difficult to find, oil companies are focused on improving the productivity of existing oilfields. Installing wireless sensors on pumps and other equipment allows operators to access maintenance data on all equipment in a location directly from their trucks, significantly improving their productivity. Equipment defects that may have gone undetected for months can now be attended to in days.

At remote oil fields, Chevron has used a type of wireless network called a "mesh" network to monitor its oil wells. Unlike traditional 802.11 networks where dedicated devices act as base stations and routers, in mesh networks, each field device acts as both a sensor and a wireless router. By placing devices suitably close to each other, each device in a mesh network requires very little power because the radio signal from the device has to travel only a short distance to the nearby node. The devices can also be designed to transmit only when needed, further reducing power requirements and allowing sensor batteries to last for up to seven years. For many sensor-deployment projects, wiring costs can be up to 75% of the cost of the project. Wireless technologies can eliminate this huge cost.

Example Case Questions

1. What are the different kinds of wireless data communication technologies used in the case?
2. What is a *retain* in the context of supply chains?
3. What is a *run-out* in the context of supply chains?
4. What is a *mesh network* in the context of wireless sensor networks? What are its advantages and disadvantages? (Wikipedia is a solid resource.)
5. Why do companies with a nationwide footprint often use satellite-based data networks for data transmission instead of wired networks such as DSL?
6. A leading provider of satellite-based data-communication services is DirecPC. Visit the company's website and write a short (one-paragraph) report on the services offered by the company based on information provided at the website. Include information such as data rates, plan prices, and other information relevant to new subscribers.

References

1. Davis, A. "Job Losses Cut into U.S. Driving." *Wall Street Journal*, January 2, 2010, A3.
2. Malik, N. S. "Refiners Keep Tab with Real-Time Monitoring." *Wall Street Journal*, December 23, 2009, B2.
3. Mir, R. M. "Satellite Data Networks." http://www.cse.wustl.edu/~jain/cis788-97/ftp/satellite_data.pdf (accessed August 17, 2018)
4. Pister, K., and G. LaFramboise. "Wired Warriors," http://www.isa.org (accessed February 2020).
5. Value Line, Industry Report. "Integrated Petroleum Firms," http://www.valueline.com/Stocks/Industries/Industry_Overview__Petroleum_(Integrated).aspx (accessed August 17, 2018).
6. Worthen, B. "Drilling for Every Drop of Value." *CIO*, June 1, 2002.

HANDS-ON EXERCISE: *AirPCap Wireshark Captures*

You have used Wireshark to capture packets on your local computer. In this exercise, you will use two Wireshark captures on the wireless interface to visualize the operation of IEEE 802.11 networks. Both captures show the download of a web page over a wireless LAN. The network topology of the setup for the capture is shown in Figure 15-9.

The two captures are included in the readings for this chapter on the student resources website: (1) Wireshark HTML capture, 802.11-header only (called "802.11 only" in this exercise); and (2) Wireshark HTML capture, 802.11 header and radio header (called "radio header" in this exercise). These captures have been made using AirPCap, which includes the hardware and software required to capture information specific to the IEEE 802.11 protocol. You may need to look at online resources to answer some of these questions.

Hands-on Exercise Questions

Answer the following questions from the 802.11-only capture.

1. Look at the first frame. The information field for this packet states that this is a beacon frame. What is the role of a Beacon frame in IEEE 802.11? Why is this frame not necessary in IEEE 802.3 Ethernet?
2. Which device on the wireless LAN sends out the beacon frame? Based on this information, what is the MAC address of the wireless router (which is also used as an access point for this capture)?
3. Based on the above question and the information in the beacon frame header fields, what information serves as the basic service set (BSS) identifier?
4. The second packet in the capture is a probe request. What is the role of a probe-request frame in IEEE 802.11? Why is this frame not necessary in IEEE 802.3 Ethernet?
5. Which device(s) send(s) out probe requests?
6. What is the BSS ID of the destination in the probe request? What does this number signify?
7. How are frames identified as beacon frames or probe-request frames or data frames? (Hint: Look at the type/subtype field.)
8. Examine the MAC address fields in a few frames. What are the three MAC addresses included in all frames?
9. Recalling the wired Wireshark captures, there were only two MAC addresses in the Ethernet header: source and destination. Why is it necessary to include a third MAC address, the BSS ID, in 802.11 frames?

Now, answer the following questions from the radio-header capture:

1. Select any frame in the capture and expand all the subheaders of the radiotap header (e.g., present flags and flags). What is the channel frequency at which the frame was transmitted?
2. Briefly describe the channels used by 802.11b/g.
3. Why is channel 6 one of the recommended channels for transmitting 802.11 wireless LAN data?
4. Was the frame transmitted using FHSS (frequency hopping spread spectrum) or OFDM (orthogonal frequency division multiplexing)?

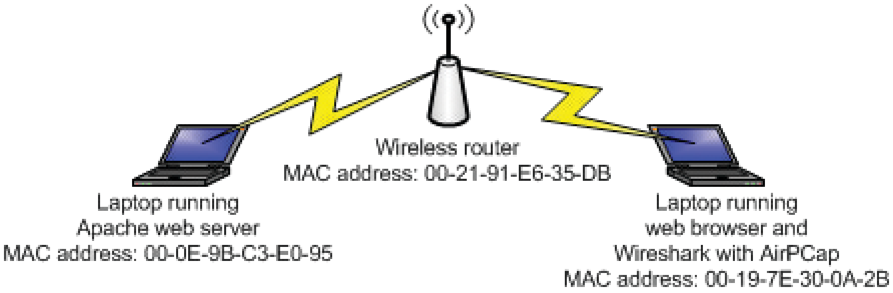


Figure 15-9. AirPCap capture topology

CRITICAL THINKING EXERCISE: *Ubiquitous Wi-Fi*

1. Technologies such as 802.11ac are bringing wireline speeds to Wi-Fi. Do you think Wi-Fi will eliminate the need for wired LANs within the next decade? Why or why not?

IT INFRASTRUCTURE DESIGN EXERCISE: *Add Wi-Fi*

The employees at the Amsterdam service center use laptops for work and need wireless coverage throughout the two floors of the building. The company has therefore decided to install a wireless LAN at this location. (In general, the company will also create an additional subnet for this wireless network, but you can ignore this detail for this class.) Answer the following questions:

1. What wireless technology would you recommend to create the wireless LAN—IEEE802.11a, IEEE802.11b, IEEE803.11g, or IEEE802.11n? Justify your choice.
2. Assume that both floors of the building have the same dimensions. Making typical assumptions about the needed workspace for each employee, what is the total area in the building that needs wireless coverage?
3. Given your technology choice and the area calculated previously, how many access points would you need to provide satisfactory coverage, where needed, on both floors?
4. Update your network diagram to include the portal for the wireless network at Amsterdam.

Supplementary Chapter 16

Phone Networks

Mr. Watson—come here—I want to see you

—Alexander Graham Bell (phone call, March 10, 1876)

Overview

The earlier chapters have focused on the technologies that enable the Internet. However, a large volume of information exchange occurs over a more humble technology—the telephone. Toward the end of the twentieth century, phone networks also served as the precursors of modern WAN networks. In recent years, the telephone has evolved with the introduction of cell phones, which add mobility to phone technology. Given the importance of technology for modern businesses, it is useful to know about the technologies that underlie land and cellular phone networks. At the end of this chapter, you should know about:

- the architecture and components of landline phone networks;
- the signals used in phone networks;
- DSL—an early technology developed by phone companies to offer high-speed Internet access;
- the architecture and components of cellular networks;
- the evolution of cell phone networks; and
- code division multiple access—the signaling scheme used in modern cell phone networks.

Introduction

Phone networks served as the early access mechanism for the Internet. In the 1990s, as Internet service providers such as AOL were perfecting their business models and technologies (recall the reference to AOL's growth), the phone network served as the medium over which most residents obtained their Internet connections. Only in the early years of the twenty-first century have carriers installed dedicated networks for Internet service. It is thus useful to have a high-level understanding of the architecture of phone networks.

The phone network is best seen as an analog information-transmission system. Signals are not digitized or packetized by landline phones. Instead, signals are simply allowed to flow over the wires in the same form in which they are generated at the source. In this chapter, we will look at the components of the phone system, and some significant events that helped the phone system evolve to its current form.

The data communication timeline shows that for more than 100 years (1840–1969), the only technologies available for long-distance information exchange were the telegraph and the telephone. Of the two, the telephone continues to be widely used globally. The basic technology underlying the phone system is quite simple. When a user dials a phone number, the switches in the phone network establish a dedicated circuit from the sender to the receiver so that signals can flow uninterrupted between the sender and the receiver. As long as the call is in progress, the network resources allocated to the call are unavailable to other callers. This mechanism of connecting users is called *circuit switching*. *Circuit switching is a process that, on demand, connects two or more communicating devices and permits the exclusive use of a data circuit between them until the connection is released.*

Why didn't the telephone pass eighth grade?

Because it wasn't a "Smart Phone."

Source: *Boys' Life* magazine, July 2012

Once the circuit is established, a microphone in the handset converts the talker's voice to electronic signals. These signals are carried over the phone network from the sender to the receiver. A speaker in the receiver handset converts these signals to sounds that are heard by the receiver.

Since the phone technology is so simple, particularly compared to the more complex packet-switched technologies used on the Internet, the phone network is sometimes also called plain old telephone service (POTS), with the term "plain old," implying that the phone is a plain vanilla kind of service. However, such descriptions should not lead us to underestimate the importance of phone service. Though the phone technology is quite simple, it is one of the most important technologies for businesses. Figure 16-1 shows the levels of adoption of landline phones in different parts of the world.[1] As seen in the figure, over the last decade, landline adoption has been relatively steady around the world. Assuming three inhabitants per family, there would be approximately 35 families for every 100 inhabitants (100/3) and with about 50 phones for every 35 families, the figure shows that on average, there is more than one phone line per family in the developed world.[2]

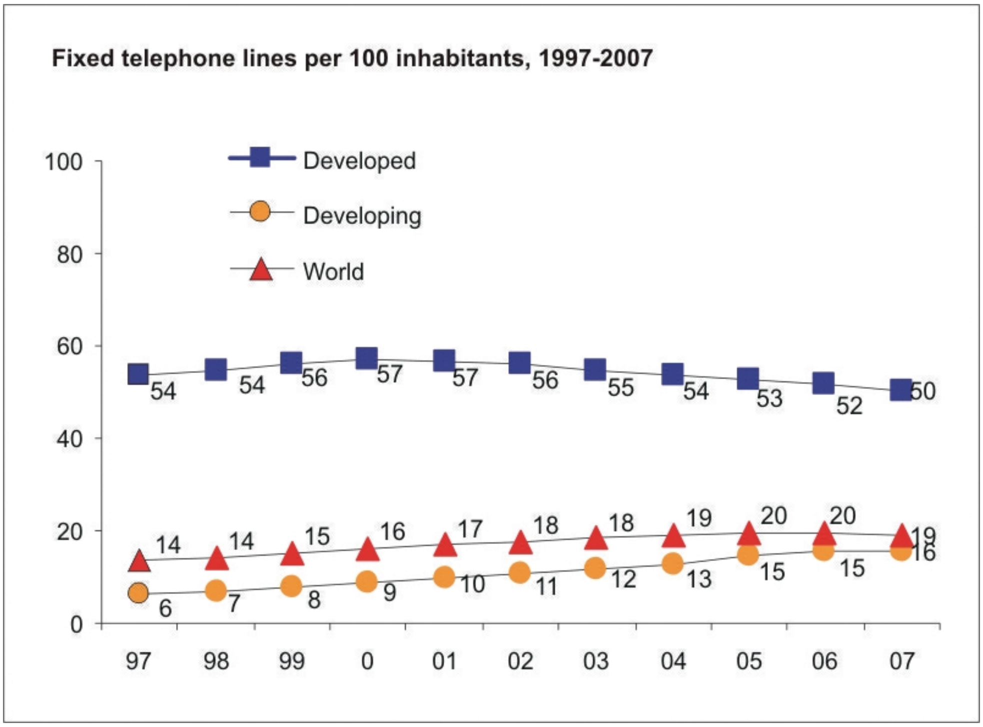


Figure 16-1. Landline adoption

[1] http://www.itu.int/ITU-D/ict/statistics/ict/ (accessed Feb. 2020).

[2] Since the statistics also include business phones, it is fair to say that most homes in the developed world have a landline phone connection.

Phone Network Components

The phone network is called the *public-switched telephone network* (*PSTN*). For basic understanding, the components of the network are shown in Figure 16-2. The PSTN terminates on the walls of end-user homes and offices. Customers maintain the phone networks within their own premises. Between the customer premises and the central office is a very visible and critical component of the phone network—the local loop. *The local loop is a circuit from the customer premises to the last switch of the phone company's network*. The local loop ends at the end office or central office of the phone company. *The end office, also called the central office, is the location where the phone company operates equipment that is responsible for providing the customer's dial tone*. The last switch of the PSTN is in the final office.

If a call is made to a user connected to the same end office, it is connected to the user by the end office. The user then hears the phone ring. If the call is being made to a user who is connected to a different end office, the call is forwarded to the appropriate end office for completion. Figure 16-2 shows this as a link to the interexchange carrier (IXC). *IXCs are networks that carry traffic between end offices*. When the remote end office connects the call, there is a completed circuit between the sender and receiver over which signals can flow.

The IXC link in Figure 16-2 has some special properties. This link is where phone networks get integrated with WAN technologies. It may be noted that whereas the local loop is dedicated to a single user, the IXC link is shared among all users of the end office. Since the end user can only make one phone call at a time, there can only be one signal at any given time on a local loop. However, at any given time, multiple customers of an end office may need connections to the IXC to get connected to users in other end offices. Hence, the phone network needs a multiplexing technology in the IXC link to enable multiple users to simultaneously share the common IXC link.

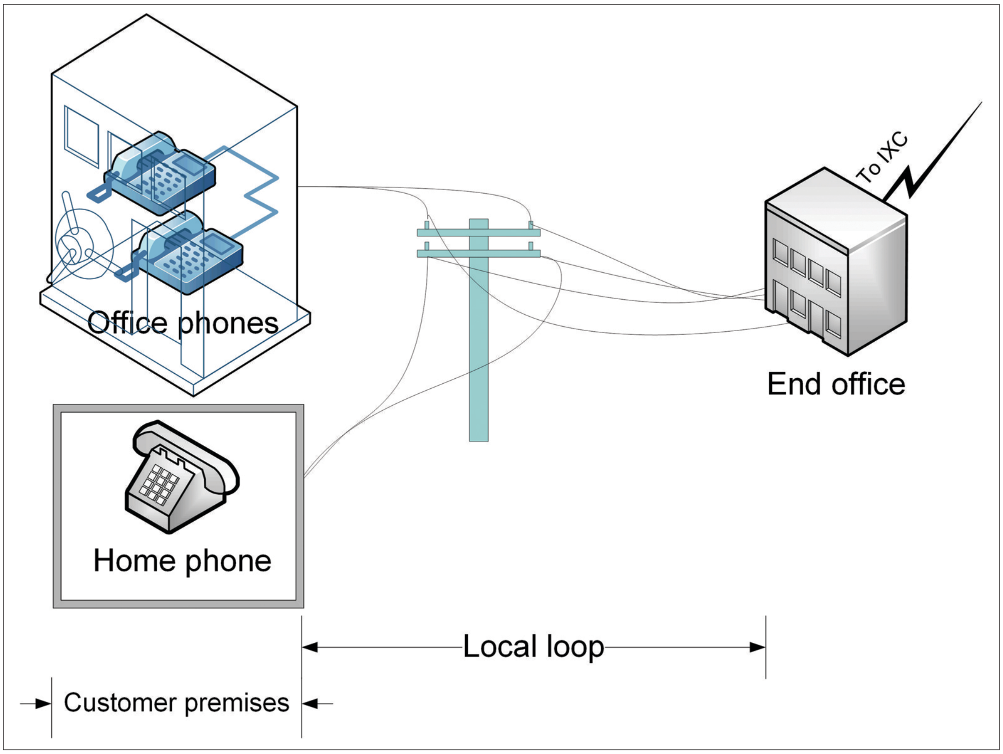


Figure 16-2. Phone network components

This can be accomplished using WAN technologies. Hence, the IXC link shown in Figure 16-2 is where WAN technologies such as ATM were first deployed for widespread use.[3]

Phone Signals

Voice is carried over the phone wires as signals. In this section, we will take a high-level look at the properties of phone signals. We have avoided getting into the details of signals in this book because the information can get very technical. However, some aspects of phone signals are useful to know because they define the capabilities of phone modems.

The phone system tries to reproduce the speaker's voice at the receiving end. This can be accomplished if all the tones in the speaker's voice can be captured and transmitted to the receiver. Tones are determined by signal frequencies. Frequencies are measured in hertz. *Hertz is a unit of frequency that is equivalent to one cycle per second*. Any operation that repeats once per second is said to operate at 1 Hertz. An operation that repeats twice per second is said to operate at 2 Hertz and so on.

Observations suggest that the human voice lies in the frequency range 80 Hz to about 1,100 Hz. Male voices typically lie at the lower end of this range, and female voices are at the higher end of this range. A system that can capture, transmit, and reproduce all the signals in the range 80 Hz–1,100 Hz should thus be able to reproduce voice with high fidelity.

The phone system has been designed to transmit signals in the 300 Hz–3,400 Hz range. Since 3,400 is greater than 1,100, the high end of the range captures all human voices at high frequencies. However, the cutoff at the low end, at 300 Hz, causes some loss of information. So, why does the phone system not capture the signals in the frequency range 80 Hz–300 Hz? Why are we willing to lose the information in these frequencies? The reason to eliminate lower-frequency signals is that the power line transmits power at 60 Hz. Filtering away signals below 300 Hz eliminates the strong hum that is likely to be created in the phone receiver by the power line and its harmonics. Transmitting signals within the frequency band of 300 Hz–3,400 Hz has been found to be adequate to convey the human voice over the telephone. While this does not create a high-fidelity (hi-fi) reproduction of the sender's voice, the sender's voice is clearly recognizable over the phone network.

The phone system only needs to transmit a signal within a relatively narrow bandwidth of about 3 kHz (3,400 Hz – 300 Hz = 3,100 Hz). The transmission of phone signals does not require a very high quality of wire in the local loop. As a result, the phone network was built using relatively inexpensive copper cables (inferior in quality even to Cat 3 cables) in the local loop. This kept costs low and served the phone network well throughout the twentieth century. By comparison, Ethernet signals require thousands of times the bandwidth for signal transmission. When customers began to demand broadband connections in the last decade of the twentieth century, the poor signal-carrying properties of Cat 3 phone cables in the local loop became a stumbling block for phone companies. Serving high data rates requires the transmission of high-bandwidth signals, something the local loop could not do. To overcome this limitation, phone companies could either invest in a new local loop or figure out a way to do the best they could with the existing local loop. Phone companies thus developed DSL—a technology that allowed high data-rate signals to be carried over phone lines for short distances.

[3] The following infographic shows how the evolution of the phone network gained widespread coverage along with the introduction of Google Voice: "A Modern History of Human Communication," http://visual.ly/modern-history-human-communication (accessed Dec. 2015).

Legal Developments

The evolution of the phone network in the United States has been influenced not just by competitive forces but also by legal action taken by the government.[4] Alexander Graham Bell had created AT&T to commercialize his invention of the telephone. Over the years, AT&T invested in creating a nationwide telephone network. Many residents were located in remote areas. Even though it was very expensive to connect these users to the telephone network, AT&T provided these customers with phone service at the same rates as urban customers. To facilitate these investments, in the early years of the twentieth century, the US government allowed AT&T to operate as a regulated monopoly. The monopoly status guaranteed the company suitable returns on its investments in developing a national network that provided phone service to all residents in the United States. This was a lot like the way most states allow gas and electric utilities to operate as regulated monopolies. The utilities agree to invest in providing services to all residents, however inaccessible they might be. In return, the state protects them from competition (monopoly) and allows them to earn reasonable (regulated) profits.

Early misevaluations of telephony technology

In 1876, Western Union, the largest American telegraph company, refused to buy Graham Bell's patent for $100,000, arguing that people are not savvy enough to handle a phone: "Bell expects that the public will use his instrument without the aid of trained operators. Any telegraph engineer will at once see the fallacy of this plan. The public simply cannot be trusted to handle technical communications equipment."

A group of British experts thought somewhat differently: "The telephone may be appropriate for our American cousins, but not here, because we have an adequate supply of messenger boys."[5]

Note: These anecdotes were written so beautifully in the original, they have not been paraphrased.

1984—Competition in Long-Distance Phone Service and the Modified Final Judgment

By the mid-1970s, technological innovations had led to the emergence of a number of competitors providing long-distance phone service. There was no longer any need to provide monopoly protection to AT&T in long-distance phone service. To level the playing field for competitors in the long-distance phone business, the US government filed an antitrust lawsuit against AT&T in 1974. The lawsuit ended in 1982 when AT&T and the US government agreed on the terms of a solution to enable competition in long-distance phone service. As part of the agreement, seven regional phone companies that offered local phone service were divested (separated) out of the old AT&T. AT&T continued in business as a long-distance phone company, competing with other firms such as MCI. The seven local phone companies were called Regional Bell Operating Companies (RBOCs). These RBOCs were granted regulated monopoly status in providing local phone service, as the government believed that this was necessary to ensure continued investments in operating the local loop. The RBOCs were prohibited from providing long-distance service. This new structure was put in place effective January 1, 1984. Figure 16-3 shows a map of the operating areas of the seven RBOCs in the United States after the breakup of AT&T.

The judgment that led to the divestiture of the seven RBOCs in 1984 is sometimes called the modified final judgment. The name relates to a judgment in 1956 from an earlier antitrust lawsuit

[4] For a very informative graphic that shows the evolution of the phone network in the United States, from Graham Bell to the contemporary cell phone providers, please see the graphic, "A Tangled Family Tree," *Wall Street Journal*, Mar. 29, 2011, http://www.wsj.com/articles/SB10001424052748704471904576229250860034510 (accessed Feb. 2016).

[5] Gerd Gigerenzer, *Risk Savvy: How to Make Good Decisions* (Penguin, 2015).

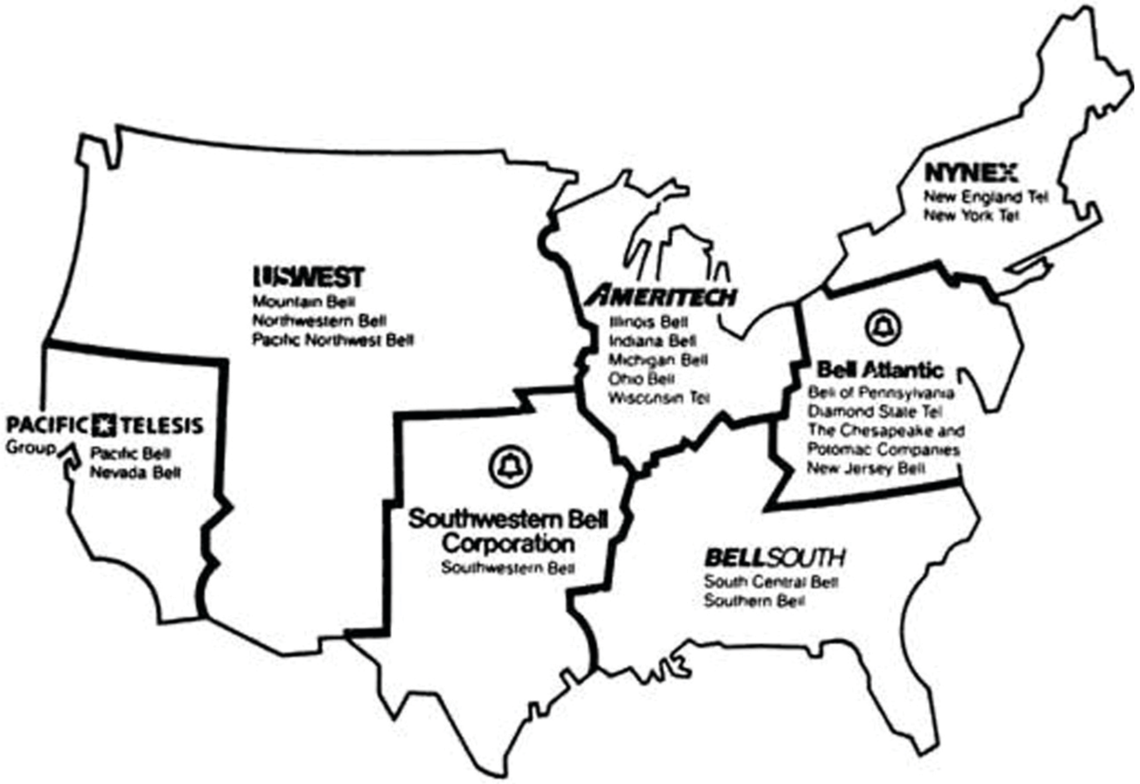


Figure 16-3. Map showing operating areas of the seven RBOCs in 1984

filed against AT&T. As a result of the 1956 judgment, AT&T agreed to restrict its activities to running the national telephone system and performing other government work. The 1956 judgment is referred to in the industry as the final judgment. The decision in 1984 modified the final judgment from 1956 and is thus called the "modified final judgment" in the telecom industry.

1996—Competition in Local and Long-Distance Phone Service: Telecommunications Act

There were rapid developments in the telecom industry following the modified final judgment. Competitors entered into the long-distance telephony market. New technologies were beginning to be deployed for Internet service. Competitive markets reduced the need for the government to impose price controls, but prior regulatory barriers had to be removed to allow free entry of competitors for providing current and emerging telecommunications services. To respond to the changes in the industry from these developments, in 1996 the US Congress passed the Telecommunications Act. The website of the Federal Communications Commission states that "the Telecommunications Act of 1996 was the first major overhaul of telecommunications law in almost 62 years. The goal of this new law was to let anyone enter any communications business—to let any communications business compete in any market against any other." Specifically, for our purposes, the major provision of the act was that it introduced competition in both local and long-distance phone service. Thus, whereas a judicial decision introduced competition in long-distance phone service in 1984, the legislative process introduced competition in both local and long-distance phone service in 1996 through a new law, the Telecommunications Act of 1996.

The Telecommunications Act of 1996 had other features that affected the phone industry. In a very novel provision, in order to create competition in local phone service, the act required RBOCs or other local phone companies to provide access to their networks at reasonable rates to competitors who wanted to provide local phone service. This provision created two classes of local phone companies—the incumbents and the competitors. The incumbents (the RBOCs) were called the incumbents local exchange carriers (ILECs). The competitors were called the competitive local exchange carriers (CLECs). You may remember regularly receiving solicitations from startup phone companies offering phone services at very competitive rates in the late 1990s and early 2000s. These startups were the CLECs. The law also allowed the RBOCs to offer long-distance phone service.

Digital Subscriber Line (DSL)

We saw earlier that when customers began to demand broadband Internet access, the limitations of the phone network immediately became apparent. The local loop in the phone network was designed to carry narrowband signals in the range 300 Hz–3,400 Hz. Broadband signals need cables capable of carrying much higher bandwidths than what the local loop was designed to carry.

So, when users began to demand broadband connections, the phone companies were at a disadvantage compared to cable companies. The coaxial cable used by cable TV companies is capable of carrying signals over a very high bandwidth. Cable companies were therefore capable of offering broadband Internet access to customers using their existing network infrastructure. Cable companies began to offer service packages that bundled cable TV, phone, and broadband Internet access. If replacing the cables in the local loop is extremely expensive, how were phone companies going to compete with cable companies once broadband Internet access became important to customers?

To respond to this business need, phone companies developed the digital subscriber line (DSL) technology. *DSL is a technology that provides full-duplex service on a single, twisted, metallic pair of phone wires at a rate sufficient to support basic high-speed data service*. Phone companies realized that though Cat3 cables could not carry high-bandwidth signals over long distances, they could carry these signals over short distances—say up to three to five miles. From a business perspective, this was very useful. In densely populated areas (e.g., places with many apartment complexes), phone companies could create end offices in a central location and offer DSL services to as many nearby customers as possible. By suitably locating end offices, DSL could allow phone companies to offer broadband Internet access to a large number of customers.

When the phone line carries DSL signals, the signals are transmitted in the frequency ranges shown in Figure 16-4. The phone signals are carried as before in the 300 Hz–3,400 Hz range. The upstream (upload) DSL signals are carried in the frequency band 25.875 kHz–138 kHz. The downstream (download) signals are carried in the frequency band 138 kHz–1,104 kHz. For reference, Figure 16-4 also shows the frequencies used by the phone channel (the narrow column at the extreme left, identified as "phone channel" in the figure).

We see from Figure 16-4 that the upstream bandwidth is about 112 kHz (138 – 25.875), while the downstream bandwidth is about 966 kHz (1,104 – 138). Thus, the downstream bandwidth is more than eight times the upstream bandwidth. Why this asymmetry? Is there something special about the downstream signal that it requires such a high bandwidth compared to the upstream bandwidth?

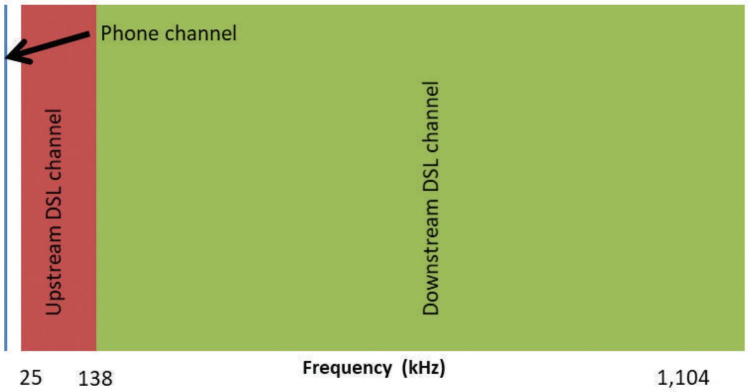


Figure 16-4. Phone and DSL signal frequencies

The large downstream bandwidth has to do with typical customer behavior. Recall the relationship between bandwidth and data rates. For the same signal and noise levels, the data rate of the channel increases in direct proportion to the bandwidth of the channel. Most end users download huge quantities of data but upload very little data. Downloads include video, web pages, and other Internet content. The most common data uploaded by end users is email. The total bandwidth of the upstream and downstream channels is fixed (1,104 – 25 = 1079 kHz). Since most end users care most for high-speed downloads and not so much for high-speed uploads, most Internet service providers engineer their systems to provide acceptable upload speeds and the highest download speeds possible.

You may note a slight gap in Figure 16-4 between the phone channel and upstream DSL channel. This gap is deliberate. End users of DSL service install a DSL splitter that separates out the phone and DSL signals on the cable. The gap between phone and DSL channels helps the DSL splitter separate out the phone and DSL signals.

As an analogy for how Cat3 cables can carry high bandwidth signals over a short distance, but not long distances, imagine driving a Corvette over a dirt road. You will be able to reach high speeds over short distances. But if you try to maintain the high speed over long distances, you will end up with a sprained back, a damaged vehicle, or both. A Cat3 cable is like a rough road for signals. You can carry low-speed signals for long distances, but high-speed signals can only be carried for short distances.

Cell Phones

Cellular telephony is a mobile communications system. It uses a combination of radio transmission and conventional telephone switching to permit mobile users within a specified area to access full-duplex telephone service.

> *Why was the cell phone wearing glasses?*
>
> *Because it lost its contacts.*
>
> Source: *Boys' Life* magazine, March 2012

The rapid adoption of cellular telephony is one of the most important developments in telecommunication technology in the first decade of the twenty-first century. Figure 16-5 shows how cell phones have rapidly become popular in most regions of the world. There is almost one cell phone for every individual in the developed world. In developing countries, cell phone adoption seems to be rising at an even faster rate than in developed countries, and there is now approximately one cell phone per family in the developing world (approximately 45 cell phones for every 35 families).

To enable cellular telephony, large geographical areas are segmented into many smaller areas. Each small area is called a cell. Each cell has its own radio transmitters and receivers and a single controller interconnected with the public-switched telephone network.

Since one cell phone tower can provide phone coverage to a wide area, and laying landlines to each home can get very expensive, building a cell phone infrastructure can sometimes actually be cheaper than building a landline infrastructure. Thus cellular telephony can help developing countries build nationwide phone networks at lower costs than is possible with conventional landline telephony.

Cell-Phone Technology Evolution

Commercial cell phone networks have been available in different parts of the world from the early 1980s. Since then, there has been a steady evolutionary improvement in the technology. There have also been two clearly identifiable revolutionary improvements in cell phone technologies. Each revolutionary improvement in cell phone technology is labeled as a generation in

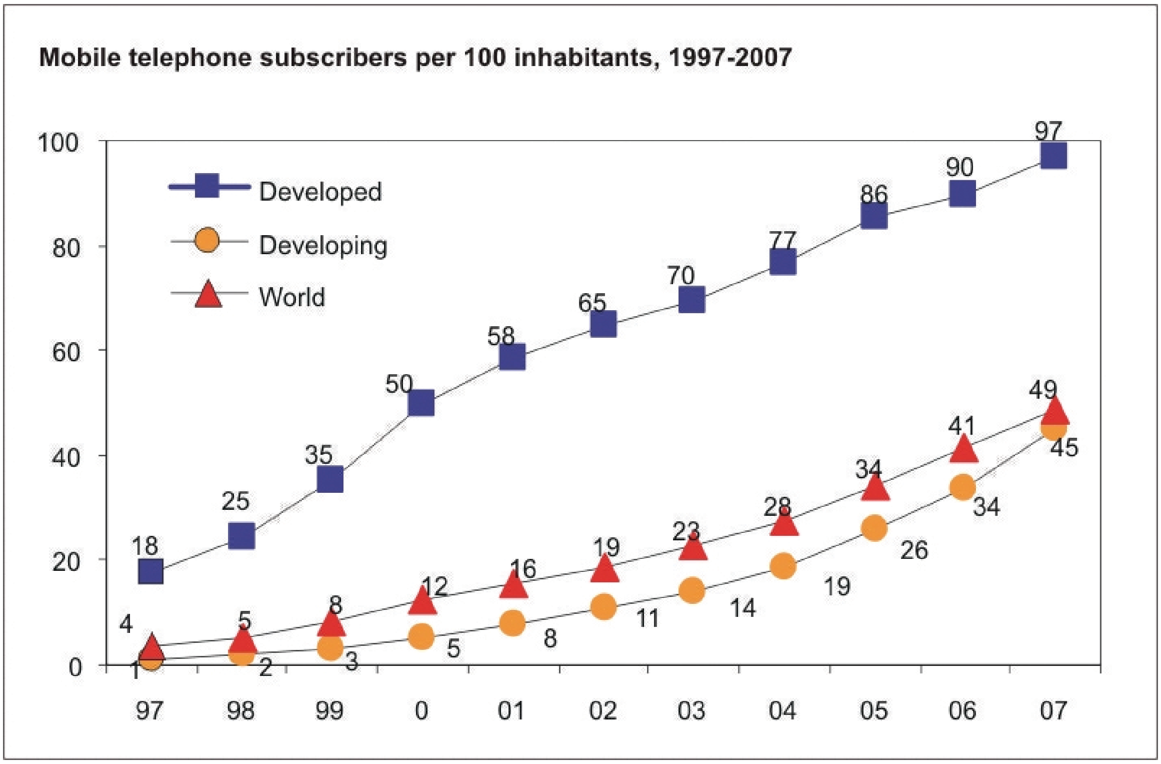


Figure 16-5. Cell-phone adoption

the telecommunications industry. Figure 16-6 shows the key features of each generation of cell phones.[6]

The earliest cell phone networks were built starting around 1980 and are now called first-generation cell phone networks. These phones used analog signals, using frequency modulation to transmit the speaker's voice over the allocated carrier frequency. This is a very simple technology that had been used for more than 50 years for wireless communication in applications such as car radios. Though the phones using this technology were heavy and unwieldy, did not support data transmission, and used wireless frequencies very inefficiently, the first generation set the stage for the future development of the technologies and markets. First-generation cellular phone networks were popular throughout the 1980s.

In the first revolutionary change, digital signals replaced analog signals to carry voice. Networks and equipment using this technology are called *second-generation cell phone networks*. The primary advantage of using digital signals is that using data-compression techniques, it is possible to send multiple digital signals using the same bandwidth used by one analog signal. Thus, second-generation networks made more efficient utilization of bandwidth compared to the first-generation cell phone networks. Second-generation networks began to be deployed just after 1990. Since the second generation used digital signals, second-generation networks also added packet-data service, enabling cell phones to be used as data modems at data rates of approximately 15 Kbps.

The current cell phone networks are called third-generation cell phone networks. Specifications for the third generation of cell phone networks were defined under the leadership of the International Telecommunications Union (ITU) in 2000. The revolutionary feature of third-generation

[6] M. Agrawal, K. Chari, and R. Sankar, "Demystifying Wireless Technologies: Navigating through the Wireless Technology Maze," *Communications of the AIS*, 12 (2003): 166–82. (Excerpt used with permission from the Association for Information Systems, Atlanta, GA. All rights reserved.)

cell phone networks compared to second-generation networks is much higher data rates. Both second- and third-generation cell phone networks use digital signals and both support voice and data. Thus, the distinction between second-generation and third-generation cell phone networks is not as distinct as the distinction between the first and second generation of cell phones. However, as the ITU states, the third-generation cell phone networks "raised the bar." Third-generation networks offer performance levels significantly in excess of those obtainable from second-generation (2G) cell phone networks. In particular, minimum data speeds for various environments are defined for third-generation cell phone networks.[7]

In general, third-generation cell phone networks are designed to offer data rates that are sufficient to simultaneously support voice and high-speed data communication. Third-generation (3G) networks are expected to offer minimum speeds of 2 Mbps for stationary or walking users, and 348 Kbps in a moving vehicle. By comparison, second-generation systems only provided data rates in the range of 9.6–28.8 Kbps. Third-generation networks help provide a desktop-like network experience on smartphones such as the iPhone.

4G LTE and 5G

Most users now use 4G networks that are sometimes marketed as LTE (long-term evolution) networks. The 4G standard specifies peak downlink data rates of 300 Mbps and peak uplink data rates of 75 Mbps. Widespread deployment of 4G networks enabled several new services such as car hailing (Uber and Lyft) and mobile multimedia.

5G further extends cellular network capabilities by supporting data rates of up to 10 Gbps and network latencies (delays) of less than 1 msec.[8] 5G is designed with the Internet of Things in mind—particularly, an urban cyberinfrastructure that supports autonomous vehicles. If cars are to drive themselves, they need to be able to communicate with other cars in their vicinity, as well as servers on the ground, and make rapid decisions to stop, change lanes, accelerate, and so on to avoid accidents. Camera and radar detection is inadequate because vehicles do not always have visibility to oncoming traffic—for example, at intersections, blind turns, and so forth. In these locations, vehicles will need to communicate with other vehicles and ground-based systems to develop a map of local traffic. The latency guarantees ensure that autonomous vehicles will have real-time traffic information to make these decisions safely.

Figure 16-6 summarizes the evolution of cell phone networks. There were two early cell phone services—TACS and AMPS. All current cell phone networks have evolved from these services. The figure also shows the important evolutionary technologies that were introduced between generations. For example, before 2G was deployed, some operators were already experimenting with a technology called the *intermediate system*. This technology introduced digital signals and TDM over FDM, which later became the accepted standard for 2G.

Similarly, before high-data-rate 3G networks were deployed, cell phone carriers introduced higher data rates than were available from 2G, though less than proposed 3G speeds. The first of these networks was called General Packet Radio Service (GPRS). These networks offered data rates of about 128 Kbps (compared to about 14 Kbps in 2G). The 128 Kbps GPRS is popularly called 2.5G. GPRS evolved to EDGE (Enhanced Data Rates for Global Evolution), offering data rates of 384 Kbps. EDGE is popularly called 2.75G.

One interesting technological feature of 3G cell phone networks is that all 3G networks are built using a form of multiplexing that we have not seen before. This is called *code division multiple*

[7] https://www.itu.int/ITU-D/tech/FORMER_PAGE_IMT2000/DocumentsIMT2000/What_really_3G.pdf (accessed Feb. 2020).

[8] https://www.thalesgroup.com/en/markets/digital-identity-and-security/mobile/inspired/5G (accessed Feb. 2020).

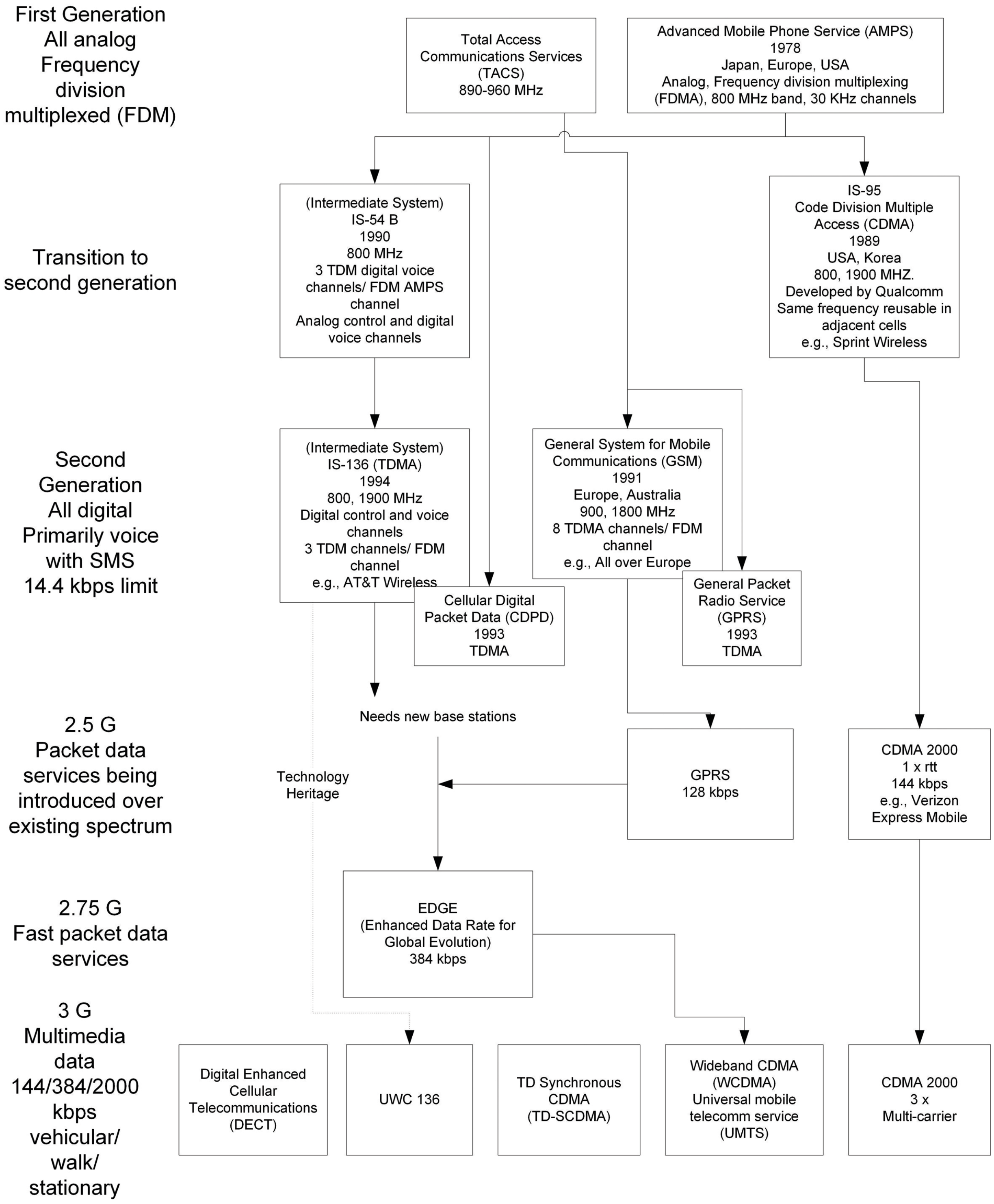


Figure 16-6. Cell-phone technology evolution

access, or *CDMA*. Supporting a large number of wireless users, with each user transmitting at very high data rates as required by the 3G standard while using the limited wireless bandwidth that is available, requires multiplexing technology that uses bandwidth very efficiently. Fortunately, CDMA is such a technology. CDMA is discussed later in this chapter.

Cell-Phone System Architecture

The architecture of the cell phone system is shown in Figure 16-7. The service area of the mobile-phone network is divided into small areas called *cells*. Cells can be of different sizes, but it is quite common for cells to be approximately two to four miles in diameter. Each cell is served by a base station that houses antennas and other electronic equipment to send and receive signals from end-user devices (cell phones) within the cell. Base stations in an area are connected to a mobile-telephone switching office (MTSO). The MTSO connects the cell phone network in its area to the PSTN, or the landline phone system, through a connection to a nearby phone exchange. This way, phone calls can be seamlessly connected between landline and cellular phones.

Tower locations

Though Figure 16-7 shows towers in the middle of each cell, in practice, towers are generally at the cell intersections, with each tower carrying multiple antennas and each antenna serving one neighboring cell.

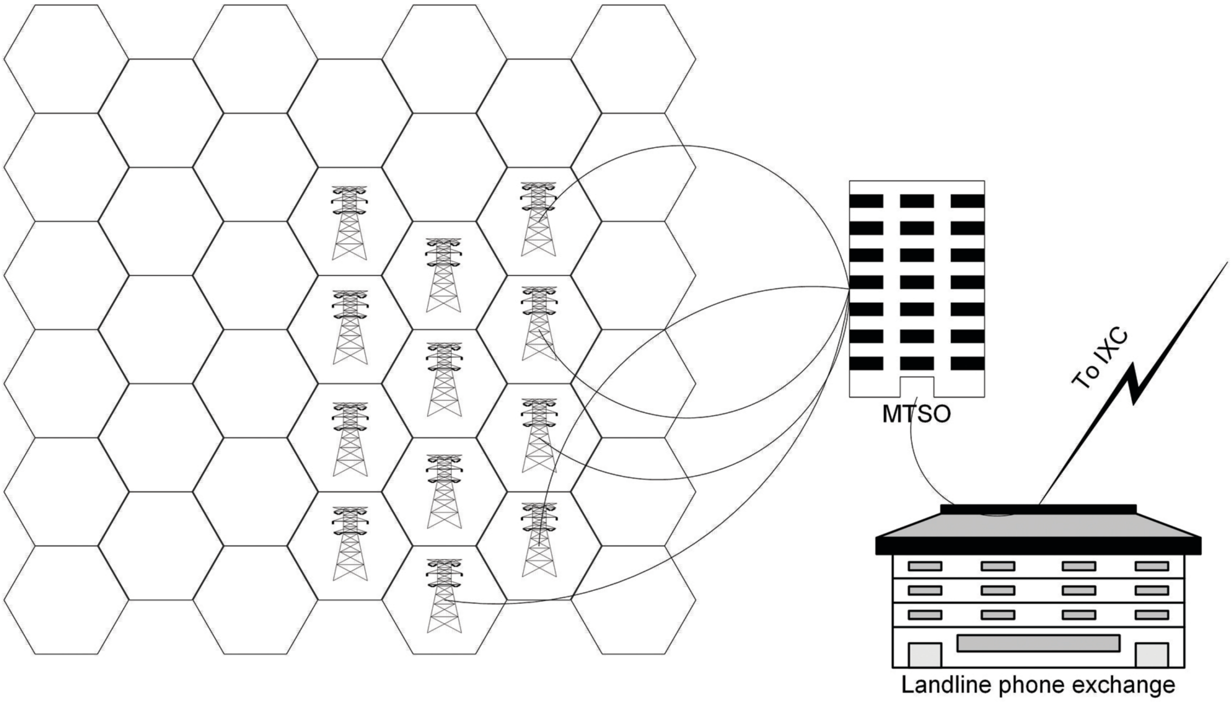


Figure 16-7. Cell-phone system architecture

Tower accidents

The buildout and maintenance of the cell phone tower network may be one of the most dangerous contemporary industries. There were 10 fatalities in 2014 and 14 in 2013. This number fell to 4 in 2015.[9,10]

There are also concerns about radiation exposure to workers who work near these towers for extended periods, as well as residents directly in front of these antennas. As of 2014, an estimated 10% of towers violated regulations regarding barricades and signage near towers warning people of possible danger from exposure.[11]

Yet another hazard in working on cell towers is that ospreys like to use these towers as nesting spots to gain a clear view of hunting spots in the distance. The nests can weigh as much as half a ton, and their presence slows down construction work. Ospreys are protected birds, and there are strict regulations on how their nests can be handled.

Frequency Reuse

Why do we divide the service area into cells? Why can't we just have one tower to serve all customers in an entire metro area? Such a system would cost far less than the current system that uses multiple towers to cover a single metro area. At first glance, it might appear that multiple cells are created to provide strong signals. We saw the case of 802.11 networks in Chapter 15, where signals from a base station only reach about 100 meters away. Is this because signals from a cell phone base station can only reach about one to two miles in each direction from the base station, for a cell diameter of two to four miles?

Actually, this is not the case. The frequencies used for cell phone communication have excellent propagation properties. Signals from the base station can reach as far as 40 miles away. Therefore, a single cell tower can indeed serve an entire metro area. In fact, base stations are designed to limit the signal range within the boundaries of the cell served by the base station. Then why do we create cells and deliberately raise the costs of the cell phone system?

The cellular design of the cell phone system is primarily motivated by the need to serve a large number of users employing the limited amount of wireless bandwidth that is allocated for cell phone service. We have seen that spectrum auctions are the mechanism by which service providers obtain bandwidth, and frequencies are extremely expensive to obtain at these auctions. To efficiently use the available bandwidth, cell phone networks reuse frequencies across cells. We know that a single frequency can serve one customer at a time in any single area covered by one cell phone tower. By reusing the same frequency in different nonadjacent towers, the same frequency can serve multiple customers within a metro region. Frequencies are not reused in adjacent cells to avoid mutual interference. Thus the network in Figure 16-7 may use a frequency reuse pattern, as shown in Figure 16-8. In this figure, the network uses four frequencies—f1, f2, f3, f4—and allocates them in cells so that no two adjacent cells use the same frequency. By creating appropriate cell patterns, these frequencies can be reused as often as necessary to cover the entire service area.

Ethan: Two antennas met on a roof, fell in love, and got married.

Grace: How was the ceremony?

Ethan: I don't know, but the reception was terrific!

Source: *Boys' Life* magazine, July 2014

[9] http://wirelessestimator.com/content/fatalities (accessed Feb. 2016).

[10] Ryan Knutson, "A New Spate of Deaths in the Wireless Industry," *Wall Street Journal*, Aug. 21, 2013.

[11] Ianthe Dugan and Ryan Knuttson, "Cellphone Boom Spurs Antenna Safety Worries," *Wall Street Journal*, Oct. 2, 2014.

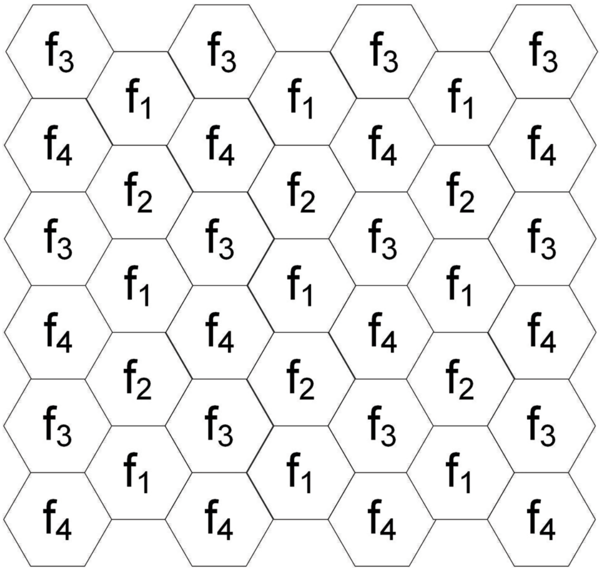


Figure 16-8. Cell-phone frequency reuse pattern example

A simple example can help us understand the benefits of frequency reuse. Say you operate a cell phone network and are allocated enough bandwidth to serve about 1,000 subscribers. If you have just one tower to serve your entire territory, only 1,000 of your subscribers may call at any given time. Suppose, instead, that you use the frequency-reuse pattern shown in Figure 16-8, and you divide your available bandwidth into four frequency sub-bands. Each sub-band would be capable of serving 250 (1,000/4) subscribers. Thus each cell would be capable of supporting 250 simultaneous conversations.

But since the network now reuses frequencies, the same sub-band can support another 250 conversations in another cell. Thus, by dividing the service area into cells as necessary, cell phone operators can support as many subscribers as they can sign up. For example, if Figure 16-8 represents the layout of cells in your network, you will have 39 cells. With each cell supporting 250 simultaneous conversations, you will be able to support 39 * 250 = 9,750 simultaneous cell phone calls. It is expensive and complicated to divide a service area into cells, but by doing so, there is no limit to the number of subscribers who may be supported using a limited amount of bandwidth. If a particular cell becomes too busy, you can simply subdivide it into more cells. Figure 16-9 shows an example.

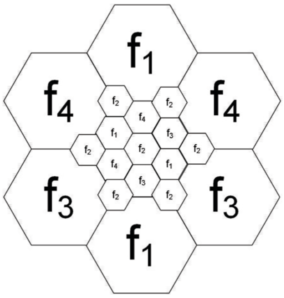


Figure 16-9. Resizing cells to accommodate subscribers

Though cells are most frequently shown as regular hexagons, in practice, cell shapes and sizes are influenced by population, terrain, buildings, hills, and so forth. Figure 16-10 shows the actual locations of cell towers in Pomona, California.[12] We see that the arrangement of cells is not very regular. Areas with a high density of users have more towers, with each tower covering a smaller cell. The figure shows that the regular hexagonal pattern commonly used to represent cell phone coverage areas of cell phone towers is only a convenient representation of the actual pattern of cells.

M&A activity driven by spectrum needs

In most industries, merger and acquisition (M&A) activities are triggered by an ambition to expand to a new market, eliminate a competitor, or acquire new customers. While all these drivers are present in the cellular industry, too, one of the primary drivers has been a need to acquire wireless spectrum from competitors. For example, AT&T offered to buy T-Mobile in 2012 for almost $40 billion in order to gain access to T-Mobile's spectrum.[13] Regulators blocked the deal, for fear of its potential to reduce competition and lead to higher consumer prices. The founder of Dish Network, Charlie Ergen, spent upward of $20 billion to acquire unused spectrum, with the hope of monetizing the asset, perhaps with 5G network deployments.[14]

Since then, some spectrum has been freed from TV broadcasting and made available to cellular providers in spectrum auctions.

[12] Retrieved from http://www.antennasearch.com.

[13] Gina Chon, Anton Troianovski, and Anupreeta Das, "AT&T Hunts Spectrum," *Wall Street Journal*, Feb. 16, 2012.

[14] https://www.cnbc.com/2018/07/27/dish-chairman-charlie-ergen-wireless-spectrum-bet.html (accessed Feb. 2020).

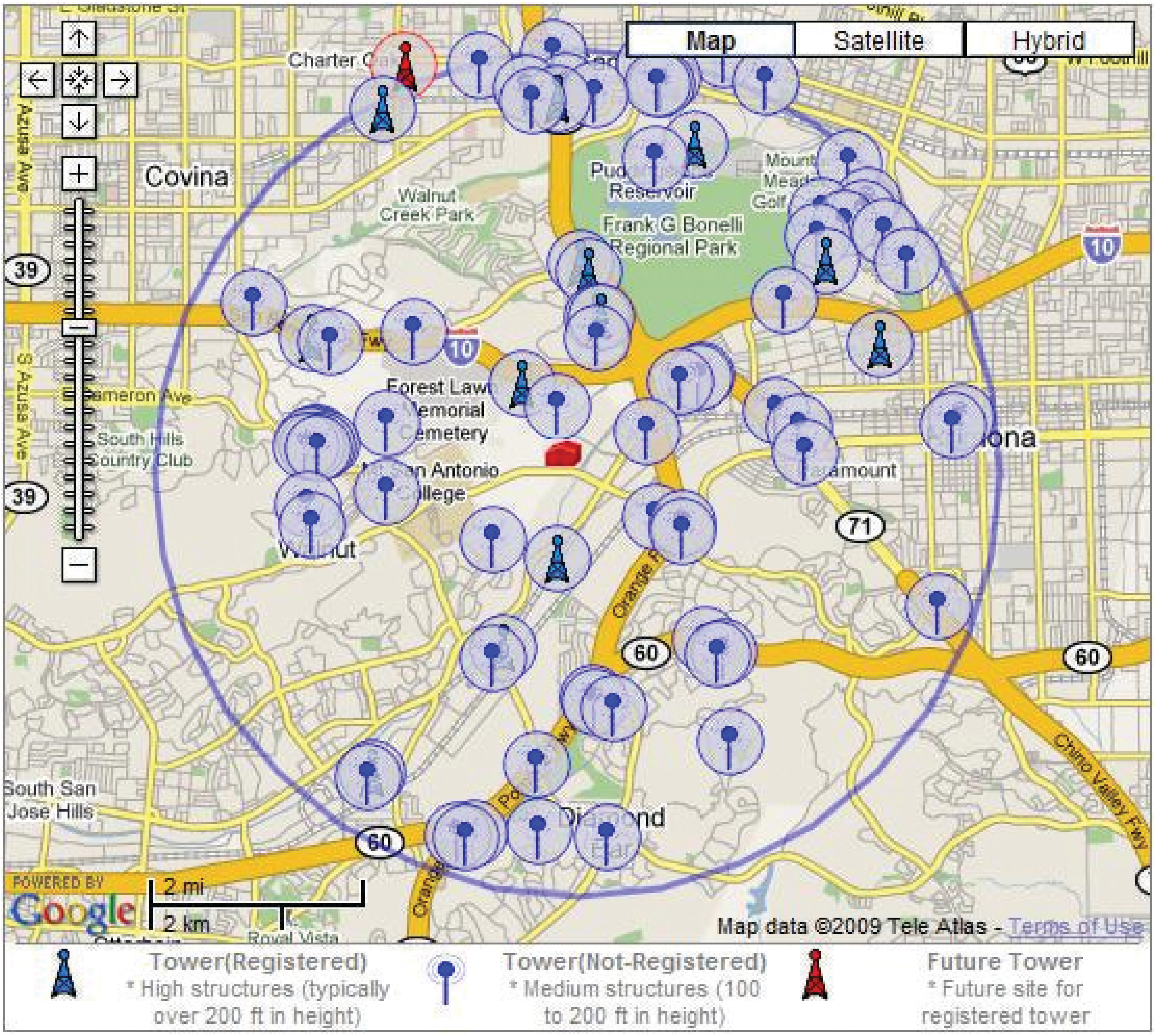


Figure 16-10. Cell phone towers in Pomona, California (Source: www.antennasearch.com)

Roaming and the Role of the MTSO

The mobile telephone switching office (MTSO) is the nerve center of cell phone service, as shown in Figure 16-7. *The MTSO is the switching office that connects all of the individual cell towers to the Central Office (CO). The MTSO is responsible for monitoring the relative signal strength of cell phones as reported by each of the cell towers, and switching conversations to the cell towers with the best possible reception.*

Every cell phone has a home MTSO. When a cell phone is switched on, it periodically broadcasts its

Network investment expenses

Building out the voice and data network is incredibly expensive and one of the largest capital-intensive projects in the country. In 2019, AT&T invested almost $20 billion in its network.[15] Verizon invested approx. $18 billion on its network during 2019.[16]

[15] https://about.att.com/story/2020/2019_earnings.html (accessed Feb. 2020).

[16] “2019 Investor Meeting Report,” Verizon, Feb. 21, 2019 (accessed Feb. 2020).

presence. These broadcasts are received by all the towers in its coverage area, and all the towers send the signal strength and other information about the phone to the MTSO. The MTSO identifies the tower receiving the best signal from the phone and instructs the tower to handle calls from the phone. The selected MTSO also informs the home MTSO of the mobile phone user that the phone is located within its service area. For example, the home MTSO for a cell phone number with area code 407 will be in Orlando, Florida. If the cell phone is currently located in Kingston, Rhode Island, the local MTSO in Kingston, Rhode Island, will inform the home MTSO of the cell phone in Orlando that the cell phone is located in Kingston. When someone dials the 407 number, the call is first connected to the user's home MTSO in Orlando. The home MTSO will then direct the call to the MTSO in Kingston, Rhode Island, where the user is located. This MTSO will forward the call to the tower that is responsible for the phone.[17]

If the user moves away from a cell, its signals to its current cell tower weaken. Simultaneously, the signals get stronger at a tower in a neighboring cell. The MTSO uses these differences in signal strengths to transfer the responsibility of handling the call to the appropriate neighboring tower. This process is called *handoff*. *Handoff is the process of transferring a phone call in progress from one cell transmitter and frequency pair to another cell transmitter and receiver, using a different frequency pair without interruption of the call.*

Code Division Multiple Access (CDMA)

Third-generation cell phones use a kind of multiplexing method that we have not seen before. It is called *CDMA*, which stands for *code division multiple access*. *CDMA is a coding scheme, used as a modulation technique, in which multiple channels are independently coded for transmission over a single wideband channel. Several transmissions can occur simultaneously within the same bandwidth, with the mutual interference reduced by the use of unique codes in each transmission.*

We saw frequency-division multiplexing (FDM), where different signals are sent at different frequencies. By tuning into one frequency, the receiver can obtain the signal at the frequency, eliminating the signals being transmitted at all other frequencies. Another common multiplexing technique is time-division multiplexing (TDM). In this scheme, each station is allowed to transmit in an allocated time slot. The receiver only listens to transmissions at the specified time slots and ignores transmissions made at other times. SONET is an example of TDM.

Whereas multiplexing schemes such as FDM and TDM are useful, they have a major limitation. Each channel is only allowed to use a fraction of the transmitting capacity of the medium. Thus, if an FDM scheme has 10 frequency slots, it can serve at most 10 users. The eleventh user will have to wait until one of the earlier users hangs up.

CDMA eliminates this limitation. CDMA allows an almost unlimited number of users to transmit signals at any time using the entire bandwidth of the medium. Each communication is allocated a unique chipping code. Before transmission, signals are processed using the assigned chipping code. Analogous to the FDM example, the receiver processes the incoming signal using the same chipping code used by the sender. This extracts the communication of interest and eliminates most of the information in all other signals. Any number of chipping codes may be generated, and therefore any number of users may be added to a cell. The only limitation is that, as used in the cell phone system, CDMA does not eliminate all information from other signals. As a result, the background noise level in CDMA increases when the number of users increases. Eventually this can make the sound quality in the cell unacceptable. When this happens, the operator divides a large cell into smaller cells and adds new towers to serve the new cells. This is why, as shown in Figure 16-9, densely populated areas have many towers in close proximity. The hands-on exercise uses a spreadsheet example to demonstrate how four users can use CDMA to send and receive 5 bits each.

[17] For a good early read on optimizations being attempted in the industry, see Scott Woolley, "The $10 Phone Bill," *Forbes*, Nov. 2009.

Summary

This chapter provided a high-level overview of landline and cell-phone networks. Though many functions of the phone system are moving over to computerized communication technologies, such as email, IM, and VoIP, the phone continues to be an important medium for business and personal communication. The phone system transmits signals in the frequency range of 300–3,400 Hz. Each end user is connected to the nearest end office of the phone company using a dedicated pair of copper wires called the *local loop*. The phone system is used in many parts of the world to offer high-speed Internet service using a technology called the digital subscriber line (DSL).

The phone system in the United States was initially operated by one company, AT&T. In 1984, the company agreed in a court settlement to focus exclusively on providing long-distance phone service. AT&T's local phone service networks were divested as seven local phone companies. Later, in 1996, Congress passed the Telecommunications Act, which opened up all sectors of telephony to competition. The Telecommunications Act also required incumbent local phone companies with established phone networks to allow competitors to use the incumbent phone company's networks at reasonable prices to compete with the incumbent.

Of late, cellular telephony is becoming increasingly popular, even in developing countries. In many cases, developing countries actually find it cheaper to set up a cell phone network than to set up a landline phone network. Cell phone networks divide the coverage area into small cells. Users in each cell are served by a cell phone tower located in the cell. As users move from cell to cell, their calls are handed off to the most appropriate cell phone tower in the area. The division of the coverage area into cells permits frequency reuse, which allows a small set of frequencies to be used to serve as many subscribers as necessary.

In which of the following contexts were the words, "Watson, come in here please," used in homage to Alexander Graham Bell?[18]

(a) March 21, 2006, Biz Stone's response, on seeing the first tweet by Jack Dorsey
(b) April 8, 2008, Noah Glass to Evan Williams, on hearing about the Twitter IPO
(c) November 6, 2008, Evan Williams to Noah Glass, on being offered a position at Twitter
(d) January 23, 2009, Noah Glass to Biz Stone, on being fired from Twitter

Cell phone technologies have evolved in three distinct phases. Each phase is called a *generation*. We are currently using the third generation of cell phone technologies, creatively called 3G. The third generation uses a multiplexing technology called CDMA, which is very efficient in using bandwidth. CDMA enables cell phone networks to offer high-data-rate network connections to a large number of end users using the very limited bandwidth available for 3G networks.

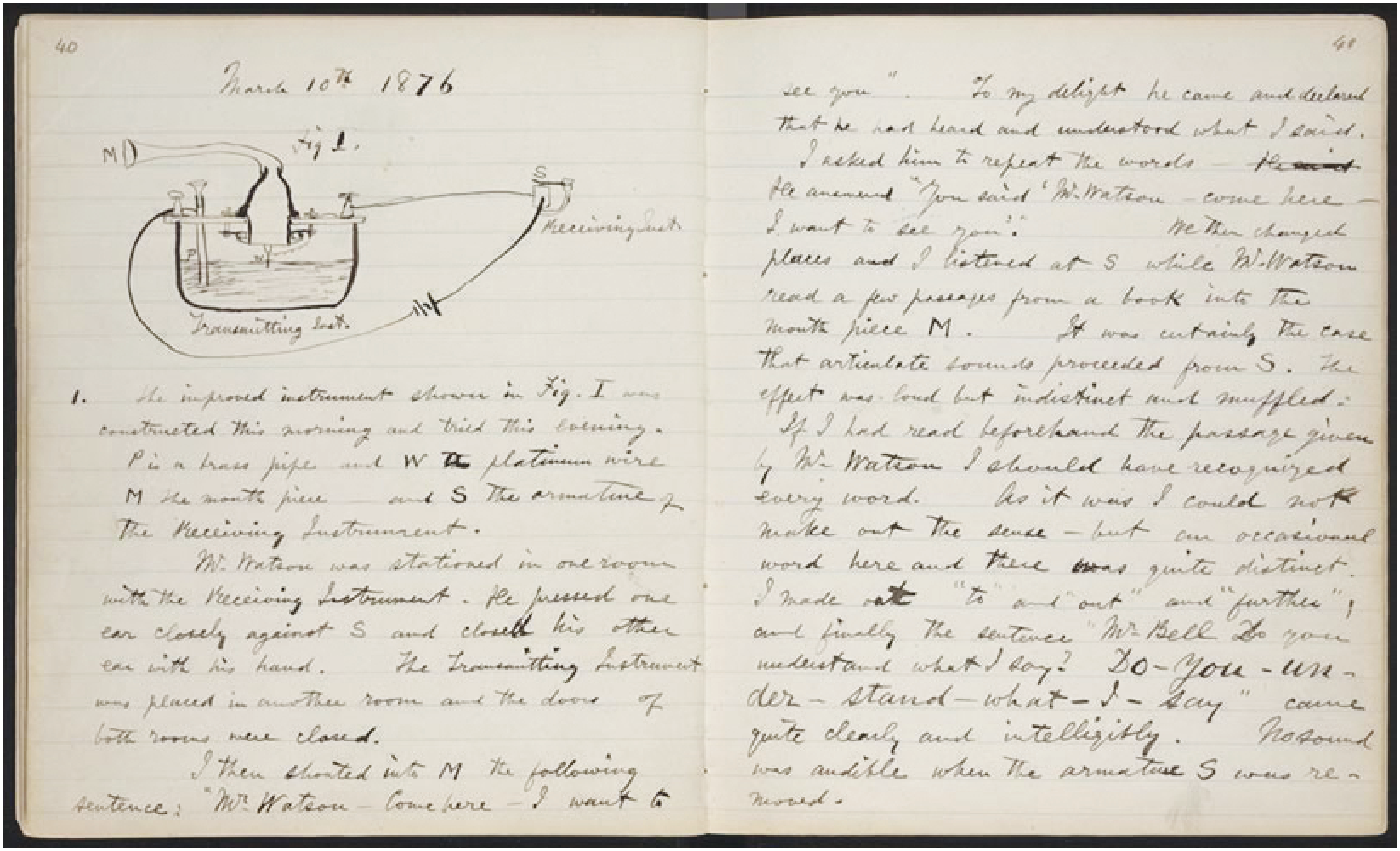
About the Colophon

First impressions are lasting impressions. It is therefore not surprising that the first phrase spoken on the telephone by the inventor of the technology is also one of the most memorable phrases ever communicated using the technology.

Alexander Graham Bell maintained meticulous diaries of his experiments to create the telephone. In his diary entry of March 10, 1876 (Figure 16-11), he described his first successful communication using the telephone.[19] When he spoke through the instrument, his assistant, Thomas A. Watson, was in the next room. In the diary, Graham Bell wrote, "I then shouted into M the following sentence: 'Mr. Watson—come here—I want to see you.' To my delight, he came and declared that he had heard and understood what I said."

[18] Answer: (a). Source: Nick Bilton, *Hatching Twitter* (Penguin, 2013). All the individuals mentioned here are the co-founders of Twitter.

[19] https://www.loc.gov/resource/magbell.25300201/?sp=22 (accessed Feb. 2020).



40

March 10th 1876

Fig 1.
M
S
P
W
Receiving Inst.
Transmitting Inst.

1. The improved instrument shown in Fig. I was constructed this morning and tried this evening.

P is a brass pipe and W ~~the~~ platinum wire M the mouth piece and S the armature of The Receiving Instrument.

Mr. Watson was stationed in one room with the Receiving Instrument. He pressed one ear closely against S and closed his other ear with his hand. The Transmitting Instrument was placed in another room and the doors of both rooms were closed.

I then shouted into M the following sentence: "Mr. Watson – Come here – I want to

41

see you". To my delight he came and declared that he had heard and understood what I said.

I asked him to repeat the words – ~~He said~~ He answered "You said 'Mr. Watson – come here – I want to see you'". We then changed places and I listened at S while Mr. Watson read a few passages from a book into the mouth piece M. It was certainly the case that articulate sounds proceeded from S. The effect was loud but indistinct and muffled.

If I had read beforehand the passage given by Mr. Watson I should have recognized every word. As it was I could not make out the sense – but an occasional word here and there ~~was~~ quite distinct. I made ~~out~~ "to" and "out" and "further"; and finally the sentence "Mr. Bell Do you understand what I say? Do – you – un-der – stand – what – I – say" came quite clearly and intelligibly. No sound was audible when the armature S was removed.

Figure 16-11. Alexander Graham Bell's diary, March 10, 1876 (Source: www.antennasearch.com)

Review Questions

1. What are some of the ways in which phone networks have been important for data communications over the years?
2. What are some of the important landmarks in the development of phone service?
3. What is *circuit switching*?
4. How has the adoption of landline telephony evolved in the last decade?
5. What is the *local loop* in the context of phone service?
6. What is the *central office* or the *end office*?
7. What are *interexchange carriers* (IXC)? What are some similarities and differences between the local loop and IXC links?
8. What is *hertz*?
9. Why is the phone system designed to carry signals in the frequency range of 33–3,400 Hz?
10. What is a *digital subscriber line* (DSL)?
11. What was the motivation for the development of DSL technology?
12. What are the three kinds of signals on a cable providing DSL and phone service? What are the frequency ranges used by the three signals?
13. Why do most ISPs provide much higher downstream data rates than upstream data rates?
14. Why is the modified final judgment important to the development of phone service in the United States?
15. What was the outcome of the modified final judgment?
16. What were the circumstances that led to the Telecommunications Act of 1996?
17. What were some of the implications of the Telecommunications Act of 1996?
18. What is *cellular telephony*?
19. What are the three generations of cellular telephony service? Describe the important features of each generation of cell phone service.
20. What is *frequency reuse* in the context of cellular telephony? Why is frequency reuse necessary for cellular telephony?
21. Why are service areas divided into small cells for cellular telephony?
22. What is the *MTSO* in cell-phone service? What are the important roles of the MTSO?
23. What is a *handoff*? Why is a handoff important? How does it work?
24. What is *CDMA*?
25. How is CDMA better suited than TDM or FDM for cellular telephony?

EXAMPLE CASE: *Cellphones and Global Development*

Mo Ibrahim founded MSI Cellular Investments (later Celtel) in Africa in 1998, when the continent with a population of 950 million people had about 2 million phones. Recognizing the opportunity for a business that connected people, Mo went about building his network and company, investing more than $750 million in seven years. In 2005, when the company was bought by MTC Kuwait for $3.4 billion, it had 24 million mobile subscribers in 15 different African countries. This made it one of Africa's most successful businesses at the time. Today, Africa has more than 350 unique mobile subscribers. Mo has now invested in a global satellite provider, O3B, which aims to bring cell phone service to remote locations in developing countries. Mo believes Africa can offer well-run businesses growth rates of more than 30% each year, a rate unattainable anywhere else in the world.

In spite of this growth, data usage continues to be expensive and unaffordable for most users in Africa. Cell phones in Africa are primarily prepaid, and prepaid data costs about $11/GB. Where daily incomes for large sections of the population are below $1.25/day, $11/GB is a very heavy price to pay for Internet access.

This has created other business opportunities and the mobility-driven eco-system continues to evolve. Entrepreneurs are outfitting mini-buses with free Wi-Fi access to the Internet. These mini-buses are a popular mode of transportation in Africa, and buses that offer this service have an advantage compared to their competitors on the same routes—gathering more fare-paying passengers who use the time to browse the web and update their social media status. Each bus pays about $25/month for data, but at every stop, passengers choose Wi-Fi outfitted buses over the competition.

In another well-known underdeveloped part of the world, Afghanistan, there were only 10,000 fixed-line phones in 2001. Services such as television were nonexistent at the time. Today, about 20 million of the 30 million people in the country use mobile phones—about the same number of people who regularly watch television. Similarly, the road network has grown from 32 paved miles to more than 7,500 paved miles. This connectivity will hopefully bring political accountability and improve lives.

In China, where development is more rapid and advanced, mobile telephones are bringing about other changes. In 2005, most consumer transactions in the country were conducted by cash. In 2015, on the other hand, more than 350 million people used their cell phones for payment—a growth of more than 60% compared to 2014. The volume of payments made through mobile phones was estimated at $2.5 trillion in 2015. Alibaba and WeChat are the leading players in the industry in China. Alibaba charges a maximum service fee of 0.6% for each transaction. By comparison, credit card swipe fees in the United States can exceed 1.5% and are the second or third highest cost for retailers, after wages and health care benefits. Alibaba and WeChat have established protections similar to those offered by US credit card companies to establish reputations for trust among both buyers and sellers. Alibaba even offers investment instruments, such as money market funds that offer higher interest rates than state-owned banks, so customers also use these applications for their wealth-management solutions.

Example Case Questions

1. What are some of the most promising investment opportunities in Africa and the developing world today?
2. If you were appointed as the regional manager for a cell phone company in Africa, with a responsibility to increase sales, what are some specific steps you would take? Briefly justify your recommendations.
3. Have you used mobile payments? Briefly describe your experience with mobile payments.

References

1. Mohseni, Saad. "The Untold Story of Afghan Progress." *Wall Street Journal*, March 17, 2013.
2. Salvaterra, Neanda. "Mobile Pioneer Sees Rich Promise in Africa." *Wall Street Journal*, April 18, 2011.
3. Vogt, Heidi. "No Wi-Fi at Home? Then Take a Bus." *Wall Street Journal*, April 15, 2014.
4. Yuan, Li. "How Mobile Payments Reshape Lifestyles." *Wall Street Journal*, February 24, 2016.

HANDS-ON EXERCISE: *CDMA*

Creating a hands-on exercise for telephony is not simple because end users have no access to phone carrier networks. Therefore, instead of trying to poke into phone company networks, the hands-on exercise for this chapter will give you the opportunity to learn about one of the most important recent developments in telephony—CDMA. You will create CDMA codes and use these CDMA codes to multiplex data transmission. You will also decode the data for reception at the receiver.[20]

In the following discussion, we find it convenient to use –1 to represent binary 0. This makes it easier to show the computations involved.

In CDMA, transmitters use mutually orthogonal codes, called *chipping codes*, to process data. The *orthogonality* of chipping codes refers to the fact that the dot product of any two chipping codes is 0. The dot product, or inner product, of two codes is calculated by multiplying the respective elements of the two codes and taking the sum of the products. Table 16-1 shows

Table 16-1. Dot product of two codes

Element	Code 1	Code 2	Product of elements
1	1	1	1
2	1	−1	−1
Dot product = Sum of the product of elements			0

[20] There is some level of math involved in this exercise, but every attempt has been made to keep it simple.

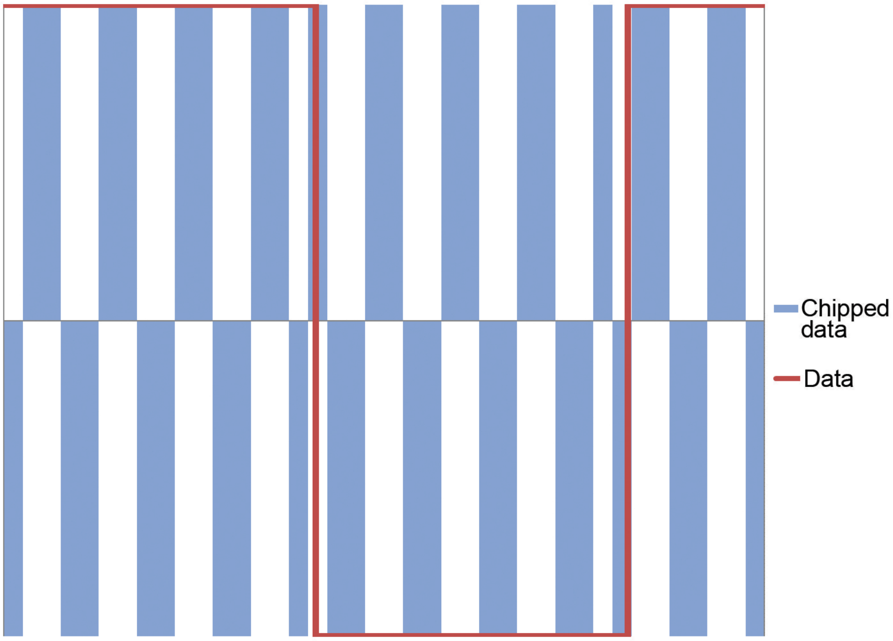


Figure 16-12. Data and chipped signal example

an example. The two chipping codes in the example are [1 1] and [1 –1]. Writing the codes in columns, as in Table 16-1, we can evaluate the dot product of the two codes and confirm that it is zero.

Techniques exist to create chipping codes of any length.[21] The chip rate is higher than the bit rate. Once a chipping code is selected, processing data for transmission involves calculating the exclusive OR of each bit of the data with every element of the chipping code. The result is transmitted into the medium.

The receivers multiply the resulting signal in the medium with the transmitter's chipping code and add the result for each bit period. This recovers the transmitted data. This is a basic CDMA operation.

The spreadsheet CDMA.xlsx in the readings for this chapter on the student resources website has an example of four pairs of users transmitting 5 bits each. Figure 16-12 shows the data and chipped signal for user D in the spreadsheet example.

The README worksheet walks you through the spreadsheet to see the data, chipping codes, sender processing, and receiver processing. You will find it convenient to also download the CDMA.xlw workspace and open the workspace. This will show you both the README and data worksheets simultaneously.

You may find it useful to use Excel's Trace precedents feature to visualize how the values are computed.[22]

Hands-on Exercise Questions

Use the CDMA.xlsx worksheet to answer the following:

1. Use Wikipedia or another resource to write a brief summary of CDMA.
2. Use Wikipedia or another information resource to write a Walsh matrix of size 4 * 4.
3. Pick any two different codes in the 8 * 8 Walsh matrix used in the data.xlsx worksheet (cells A[28]–H[36]). Using a procedure similar to Table 16-1, confirm that the dot product of the two codes is zero.
4. Repeat Question 3, but use the chipping code for B for both codes. Confirm that the dot product of a code with itself is not zero.
5. The example in the spreadsheet uses the chipping codes from the first four columns in the Walsh codes of size 8. Replace the chipping code for D with one of the unused chipping

[21] For one such technique, please search Wikipedia or another resource for *Walsh codes*.

[22] William Rogers, a student in the University of South Florida's fall 2013 class, suggested this.

codes (any code in columns E[29–36] … H[29–36]). Confirm that the data is recovered correctly with the new code (contents in cells A22–D26 should not change).

6. You saw in Question 4 above that the dot product of a code with itself is not zero. To see the impact of this, repeat Question 5 above, but this time, reuse the chipping code for user B for user D. Confirm that the data is not recovered correctly.

CRITICAL THINKING EXERCISE: *Other 3 Billionaires*

The company we saw in the example case, O3B, selected its name to represent the "other 3 billion" people on the planet who did not have Internet access. However, given the range of its customers, the company is sometimes mocked as representing the "other 3 billionaires," for bringing high-speed network connectivity to billionaires' yachts, or the "other 3 battle groups," for bringing high-speed network connectivity to military ships.

1. Should O3B protect the purity of its business by discontinuing service to profitable but wealthy customers? Justify your response.

Reference

1. http://spacenews.com/2014-top-fixed-satellite-service-operators-once-mocked-03b-investment-now-force-multiplier-for-ses/ (accessed Feb. 2020).

IT INFRASTRUCTURE DESIGN EXERCISE: *Switch to VoIP*

While a Mumbai location uses VoIP for all its voice traffic, a Singapore location is trying to decide whether to switch its users from traditional phones to VoIP phones. Answer the following questions:

1. Using the Internet or another information resource, compare the advantages and disadvantages of VoIP compared to traditional phone service (PSTN) along dimensions such as cost, performance, and reliability.
2. Based on the above, would you recommend that TrendyWidgets switch from PSTN to VoIP?

Supplementary Chapter 17

Services Delivery

He's making a list
And checking it twice.
—Haven Gillespie, "Santa Claus Is Coming to Town," 1934

Introduction

The computing infrastructure introduced in the previous chapter, together with the communications infrastructure discussed in prior chapters, is used to deliver technology services to end users. While the previous chapters have focused on the technology design choices to maximize efficiency in such service delivery, managerial design choices are equally necessary to maximize efficiency. In recent years, organizations have increasingly focused on these types of choices. This chapter introduces the key managerial design choices used in industry for the efficient delivery of technology services to end users. At the end of this chapter you should know:

- the key managerial design choices used in delivering technology services to end users;
- the popular design choices for maintaining high availability of services; and
- the essential practices used to maintain service delivery during extreme events.

IT Services Management

IT services management (ITSM) is the set of activities performed by an organization to deliver IT services to end users. ITSM aims to focus the IT organization on using technology to meet the organization's business needs. ITSM is a relatively new domain of research and practice and reflects managerial design choices in services delivery.

Careful consideration and adaptation of ITSM recommendations can be very useful. Most students reading this text will have no control over their technology design choices. For example, students of this text are very unlikely to develop a new network layer protocol. However, most students will have substantial control over their managerial design choices, on budgeting, training, and capacity planning, for example. Poor choices can cause significant delivery inefficiencies, in spite of deploying the best technology infrastructure. Therefore, while knowledge of the underlying technology design principles is important, knowledge of ITSM recommendations can help students actually influence the design of technology solutions.

ITSM Frameworks

A number of frameworks have been developed over the years for technology services delivery, each reflecting the perspectives of the leading sponsors of the framework. These frameworks include COBIT, ISO 20000, and ITIL. COBIT (Control Objectives for IT) is sponsored by ISACA (Information Systems Audit and Control Association) and, reflecting the perspective of auditors, has a strong focus on compliance and governance of the IT function. ISO 20000 has been developed by the International Standards Organization (the same organization that developed the OSI model) and has a strong focus on defining the requirements that an IT service provider should meet. ITIL (IT infrastructure library) was initially developed by the UK government as best practice recommendations so that the various departments of the government could manage their IT functions consistently and efficiently. ITIL thus has a focus on providing detailed guidance on implementing processes for services delivery. ITIL has become increasingly popular as a framework for discussions between business and IT, and the recommendations introduced in this chapter are inspired by ITIL.

IT Services Delivery Background

Historically, business users, solution developers, and the IT operations team worked independently. Business users demanded applications, solution developers developed these applications to the best of their ability, and the operations team did what it could to keep the applications running without interruption over the life of the application. Application developers were often unaware of the capabilities and limitations of the IT infrastructure maintained by the operations team until the application was ready to be deployed. This was acceptable until around 2000, when computer applications were largely internal and most applications had very few users.

However, beginning in around 2000, applications went online, and popular applications began to attract hundreds of thousands of users each day. Most application developers, who generally tested their applications on their local machines, had no idea of concepts such as threading, memory management, and security that were needed for their applications to serve this many users. The old model became indefensible in this new internet-driven environment. It became necessary for the systems development function to be aware of the operational prerequisites and to embed these prerequisites in every new systems build.

Bill Gates's trustworthy computing memo and Microsoft's code moratorium

On January 18, 2002, Microsoft's then chairman, Bill Gates, wrote a famous memo titled "Trustworthy Computing."[1] In the memo, he indicated the need for Microsoft's products to focus on security to the same extent that they historically focused on adding new features. Later, in February 2002, Microsoft put a moratorium on new code development in Windows in order to train its developers on security.

The overall cost of this effort exceeded $100 million. It is considered an important landmark in the evolution of software development to integrate security details at the time of technology development.

The ITSM recommendations aim to make the IT organization strictly focus on client needs, using well-designed IT processes. As a result, the IT organization concentrates on the services required by the customer, rather than focusing on the technology fashions of the day. In the next section, we look at some key managerial design recommendations for efficient service delivery. These have been drawn from ITILv2 and focus on the steady-state operation of services, which (in the opinion of the authors) is conceptually simpler for fresh graduates to understand. ITILv3,

[1] https://www.windowscentral.com/bill-gates-memo-industry-practice-story-security-development-cycle (accessed Feb. 2020).

introduced in 2007, replaced this model with a life cycle model, focused on ongoing improvement of services. If your work involves services delivery, you will learn about these issues in great detail at work. This chapter is intended to introduce you to services delivery and its associated concerns.

Service Delivery Disciplines

The five areas of managerial discipline identified by ITILv2 are listed as follows and briefly described further in this section. The general idea is that an entity at the service provider (called the *service desk*) enforces these disciplines:

- Service-level management
- Capacity management
- Contingency planning
- Availability management
- IT financial management

Service-Level Management

Service-level management refers to the maintenance of a catalog of all services offered, together with binding agreements with both the provider and the customer for performance. Effective service-level management reflects technological feasibility and client budgets and can minimize conflicts. The agreements with providers and customers are called *service level agreements* (*SLAs*).

SLAs are output-based contracts between customers and providers that define the service the customer can expect to receive in return for the fees paid. A typical SLA includes definitions for the service under consideration, the reliability and other performance metrics the user can expect, the procedures for reporting problems, the responsiveness a client can expect when reporting a problem, the consequences for not meeting service levels, and any escape clauses that may invalidate the SLA. Typical escape clauses include extreme events such as floods, earthquakes, terrorist events, and so forth.[2]

Developing SLAs can lead to other benefits beyond reduced conflicts between clients and vendors. During the course of developing the SLA, both client and vendor can better understand each other's priorities—especially any priorities that are unique to the other party. Also, clients often better understand the importance of quantitatively measuring all aspects of their services in order to be able to report problems.

Capacity Management

Capacity management is the managerial discipline for maintaining IT infrastructure at the right size to meet current and anticipated business needs in a cost-effective manner. Capacity management requires rigorous performance measurements of current services to detect bottlenecks and unused resources. These measurements can help reduce the capacity-planning uncertainties related to anticipated growth.

[2] Sample SLAs can be found on the Internet. An example is https://www.bmc.com/blogs/sla-template-examples/ (accessed Sept. 2020).

Global Crossing

One of the best-known cases of capacity management gone bad in any industry relates to the telecom firm Global Crossing. Between 1997 and 2002, the firm incurred more than $12 billion in debt to build out a global submarine optical fiber network in anticipation of growth in data and voice traffic resulting from the growth of the Internet. When the demand did not materialize, the firm had to declare bankruptcy.[3] Thousands of technology executives were affected.

There was likely an element of greed in the rushed buildout, but unwise emotions often accompany poor capacity planning.

Continuity Management

Continuity management is the discipline of planning to ensure that IT services can recover and continue, should a serious unexpected incident occur. Continuity management includes proactive measures to reduce the likelihood of unexpected incidents, in addition to reactive measures to be taken when the incidents occur. As unexpected incidents such as terrorist attacks and financial setbacks become more frequent, most large organizations require continuity management from IT providers before doing business with them.[4]

An entire section of this chapter is devoted to key elements of continuity management—business continuity and disaster recovery. At a high level, however, continuity management involved prioritizing the business processes to be recovered if an incident occurs, minimizing the risks facing high priority services, and developing options for recovery under different circumstances. Once developed, these plans are tested and revised periodically.

Availability Management

Availability management is the discipline of ensuring that resources such as IT infrastructure and personnel are appropriate for meeting the service-level agreements in place. On a day-to-day basis, end users are generally most concerned about the availability of IT resources; therefore weaknesses in availability management can be extremely stressful to both the client and the service provider. This puts a premium on developing maintainable software and components so problems can be remedied quickly and designing resilience into the technology so that common usage scenarios do not cause the system to fail.

IT Financial Management

IT financial management is the discipline of accurate accounting of IT services and using this information to deliver IT services in the most cost-effective manner possible. IT financial management improves the viability of the infrastructure supporting service delivery by establishing sound client prices, and it also helps in identifying areas where productive investments are possible. Financial management is usually not very interesting to IT staff, and so it is, unfortunately, neglected, which often results in technology underinvestment.

The financial management exercise can help organizations identify costs that may have escaped notice, such as building costs, depreciation costs, external service-provider costs, and software-licensing costs.

[3] For a story on the heady start of the company, please see http://www.forbes.com/forbes/1999/0419/6308242a.html (accessed Sept. 2020).

[4] FEMA has a website with templates for many disciplines in services delivery: https://www.fema.gov/emergency-managers/national-preparedness/continuity/toolkit (accessed Sept. 2020).

These disciplines are evolving with technology and end-user expectations, as well as with the industry's experiences with deploying these disciplines. It is anticipated that with the right managerial discipline, the IT organization can efficiently deliver services to agreed standards. The standards themselves would be developed in dialog with customers. Changes in end-user needs would be met with minimal disruptions to ongoing services and would leverage existing infrastructures. Suppliers would be kept in the loop as changes are planned so that suppliers could build and deliver the necessary components on time, with minimal disruptions to their own production schedules.

The next two sections in this chapter take a closer look at two components of service delivery in more detail—high availability and business continuity. These components have been chosen for their salience to clients on a day-to-day basis.

High Availability Concepts

Today's businesses demand 24/7 availability of information in order to serve customers, make smart decisions, push innovation, take advantage of opportunities, and stay one step ahead of the competition. *High availability is the ability of a system to remain operational for a duration that is significantly higher than normal*. High availability is an important goal of service delivery, and in this section, we look at some of the important ways in which high availability is achieved.

Characteristics of High Availability

IT systems that support business must be online and accessible through both planned and unplanned events. However, organizations need to achieve high availability within budgetary constraints, carefully determining the right level of availability for each part of the IT infrastructure, and striking a balance between the costs and risks associated with the nonavailability of information. Ongoing availability management aims to optimize this balance.

High availability (HA) is assessed from the perspective of an application's end user. End users are disappointed when needed data is unavailable, and they can easily switch their loyalty to a competitor's brand when system availability becomes unacceptable, for whatever reason. Availability failures due to higher-than-expected usage create the same adverse outcomes as failures due to outages in critical components in the solution.

High availability system designs typically share four characteristics: reliability, recoverability, error-detection, and continuous operations. Reliable hardware and software components facilitate the design of HA solutions. Software components include the database, web servers, and applications. Recoverability involves the design choices that facilitate recovering from a failure if one occurs. This involves anticipating important failures and planning to recover from those failures in the time that meets business requirements. For example, if a critical table is accidentally deleted from the database, how would you recover it? Error detection refers to quick awareness of a problem so that recovery procedures can be adopted. Continuous operations refer to keeping all maintenance operations invisible to the end user.

High Availability Requirements

Since IT systems run throughout the year, even small failures of availability add to downtime significantly, as shown in Table 17-1.

Table 17-1. Downtime as a function of availability

Availability percentage	Approximate downtime per year
95%	18 days
99%	4 days
99.9%	9 hours
99.99%	1 hour
99.999%	5 minutes

While 99% performance would be adequate in most contexts, it leads to four days of downtime in a year—an unacceptable level of downtime for any business. However, implementing high-availability solutions is expensive. High-availability implementations involve more fault-tolerant (i.e., expensive) and redundant systems for business technology components and also require greater ongoing expenses in IT staff, processes, and services to reduce downtime. The transition to high availability can involve tasks such as

- retiring legacy systems,
- investing in more sophisticated and robust systems and facilities,
- redesigning the overall IT architecture to adapt to a high-availability model,
- redesigning business processes to support high availability, and
- hiring and training personnel.

Therefore, businesses start by identifying technology components that require high availability. An analysis of the business requirements for high availability and an understanding of the accompanying costs help businesses achieve availability within budgetary constraints.

Framework for Determining High-Availability Requirements

Two considerations are commonly used to identify the business and technology components that most benefit from high availability: business impact analysis and cost of downtime.

Business Impact Analysis

A rigorous business impact analysis identifies the critical business processes within an organization, calculates the quantifiable loss risk for unplanned and planned IT outages affecting each of these business processes, and outlines the less tangible impacts of these outages. It takes into consideration essential business functions, people and system resources, government regulations, and internal and external business dependencies. This analysis is done using objective and subjective data gathered from interviews with knowledgeable and experienced personnel, reviewing business practice history, financial reports, IT systems logs, and so on.

The business impact analysis categorizes business processes based on the severity of the impact of IT-related outages. For example, at a semiconductor manufacturer, with chip design centers located worldwide, an internal corporate system providing access to human resources, business expenses, and internal procurement is not likely to be considered as mission-critical as the customer-facing website. Any downtime of the customer-facing website is likely to severely affect the ability of the organization to obtain customer orders, which in turn can have a material financial impact on the company. At a consulting organization, on the other hand, the internal HR system is mission critical, since a problem with the HR system at the organization will prevent the organization from assigning and tracking its consultants working on different projects. The customer-facing website may be less critical in the short run. This leads us to the next element in the high-availability requirements framework—cost of downtime.

Cost of Downtime

A well-implemented business impact analysis provides insights into the costs that result from unplanned and planned downtimes of the IT systems supporting the various business processes. Understanding this cost is essential because this has a direct influence on the high-availability technology chosen to minimize the downtime risk.

Various reports have been published documenting the costs of downtime across industry verticals. These costs range from millions of dollars per hour for brokerage operations and credit card sales, to tens of thousands of dollars per hour for package shipping services.

While these numbers are staggering, the reasons are quite obvious. The Internet has brought millions of customers directly to the businesses' electronic storefronts. Critical and interdependent business issues such as customer relationships, competitive advantages, legal obligations, industry reputation, and shareholder confidence are even more critical now because of their increased vulnerability to business and technology disruptions. The higher the downtime costs arising from a business process, the greater the justification for implementing high-availability technology solutions to support the business process.

Upon completion of the business impact analysis, businesses can define service-level agreements (SLAs) in terms of high availability for critical aspects of its business. Commonly, business processes are categorized into several HA tiers:

- Tier 1 business processes have maximum business impact. They have the most stringent HA requirements, and the systems supporting them need to be available on a continuous basis. For a business with a high-volume e-commerce presence, this may be the web-based customer interaction system.
- Tier 2 business processes can have slightly relaxed HA requirements. The second tier of an e-commerce business may be their supply chain/merchandising systems. For example, these systems do not need to maintain 99.999% availability. Thus the HA systems and technologies chosen to support tier 1 and tier 2 are likely to be different.
- Tier 3 business processes may be related to internal development and quality assurance processes. Systems supporting these processes need not have the rigorous HA requirements of the other tiers.

High-Availability Architectures

Contemporary high-availability architectures can be categorized into local high-availability solutions and disaster-recovery solutions. Local high-availability solutions provide high availability at a single location—typically a single data center. Local high-availability solutions can protect against threats such as process, node, and media failures, as well as human errors. Disaster recovery solutions are usually geographically distributed and provide high-availability during disasters such as floods, hurricanes, or regional network outages. Disaster recovery solutions can protect against local disasters that affect an entire data center.

A number of technologies and best practices are used to achieve high availability. The most common mechanism, though, is redundancy, whereby redundant systems and components are used to process user requests. *All necessary resources organized in the manner most suitable to deliver user services is called a system instance*. High-availability solutions are built by organizing system instances appropriately. Active system instances are instances involved with handling user requests. Passive system instances are fully configured instances that are not currently handling user requests. Local high-availability solutions are categorized into active-active solutions and active-passive solutions by their level of redundancy (Figure 17-1):

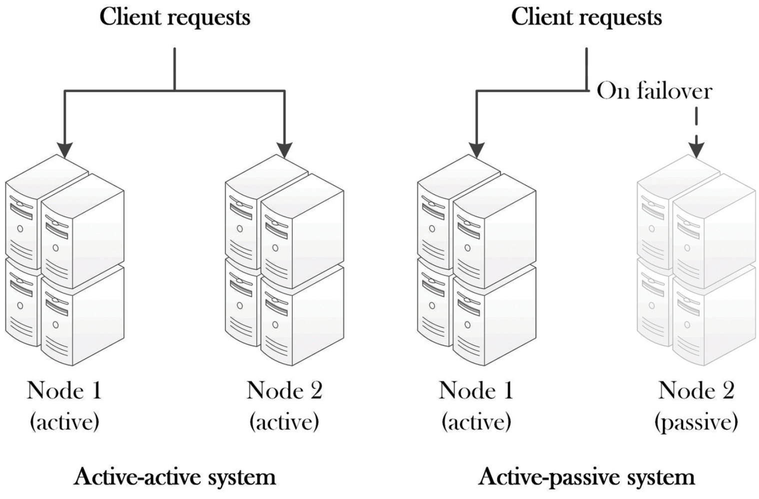


Figure 17-1. Comparing active-active and active-passive high-availability solutions

- Active-active solutions deploy two or more active system instances and can be used to improve scalability as well as provide high availability. In active-active deployments, all instances handle requests concurrently.

- Active-passive solutions deploy an active instance that handles requests and a passive instance that is on standby. In addition, a heartbeat mechanism is set up between these two instances. The heartbeat mechanisms are vendor-specific operating system technologies to automatically monitor and failover between cluster nodes, so that when the active instance fails, an agent shuts down the active instance completely, brings up the passive instance, and application services can successfully resume processing. As a result, the active-passive roles are now switched. The same procedure can be done manually for planned or unplanned downtime. Active-passive solutions are also called cold failover clusters.

Achieving High Availability

High availability is commonly achieved from the hardware. Additionally, operating systems and applications can help with achieving high availability.

The most basic high-availability strategy is to ensure that the hardware is as robust as possible, minimizing failures in the first place. Application servers also usually have many built-in high-availability features. For example, servers can provide the means to replicate data between multiple instances of applications to maintain service and data availability. Enterprise server operating systems have features to provide failover clustering, which keeps both applications and operating systems highly available.

Achieving High Availability at the Hardware Level

High availability generally begins with the hardware. An effective hardware strategy can significantly improve system availability. Hardware strategies can range from simply adopting commonsense practices with inexpensive hardware to using expensive fault-tolerant equipment.

High Availability at the Hardware Level using Robust Hardware

Unplanned downtime can be reduced significantly with high-quality, reliable components that are less likely to fail in the first place. Redundant components can then be added to take over in case of a hardware failure. An effective hardware strategy involves standardized components, accessible spares, and careful maintenance of the environment.

Fault-Tolerant Servers

Fault-tolerant servers have high or complete redundancy across all hardware components. This includes power supplies, fans, hard disks, memory, and CPUs. When a component such as a power supply fails, secondary components continue to seamlessly serve user workloads. As such, fault-tolerant systems "operate through" a component failure without loss of data or application state.

Server equipment varies in its level of fault-tolerance. Most high-end servers employ at least some redundant components, especially to eliminate common points of failure, but they will still fail when a component that is not redundant, such as a microprocessor or memory controller, fails. True fault-tolerant servers use complete redundancy across all system components, ensuring that no single point of failure can compromise system availability. Some fault-tolerant server designs extend this level of redundancy across data-center boundaries by letting the server's redundant subsystems be installed in separate, yet connected, locations. Support for fault tolerance in enterprise servers is handled completely at the operating system kernel and hardware abstraction layer—a method that makes it transparent to applications.

Dynamic Hardware Partitioning

We have seen earlier (in the discussion of virtual machines) that contemporary server hardware and software can generally be configured into one or more isolated hardware partitions where each hardware partition runs its own instance of the operating system and associated applications. Hardware resources on any partition are isolated from the other partitions on the same server. These partitions can even be changed dynamically. *Dynamic hardware partitioning is the capability of servers that allows system administrators to change the allocation of hardware partition units to each partition while the servers are still running*.

Dynamic hardware partitioning further enhances server availability and fault tolerance. On servers that can be dynamically partitioned, administrators can hot replace or hot add additional processors and memory to partitions without restarting the operating system or applications running on the hardware partition as needed.[5] This significantly increases the reliability, availability, and serviceability of servers. For example, memory chips that show signs of failing can be replaced, or spare processors can be added to partitions as demand increases.

High Availability at the Hardware Level Using Unreliable Commodity PC Architecture

High reliability also can be achieved by creating a reliable computing infrastructure from clusters of unreliable commodity PCs. This architecture was originally promulgated by popular technology leaders such as Google by replicating services across many different low-cost PC machines (instead of high-cost fault-tolerant servers) and automatically detecting and handling failures.

In this architecture, the cost advantages of using inexpensive, Intel PC-based clusters over high-end multiprocessor servers can be quite substantial, at least for highly parallelizable applications. A common feature of these applications is that they use unstructured data and primarily perform read operations on the database. These architectures leverage commodity PCs following a few key design principles, including the following.[6]

Software Reliability

Instead of fault-tolerant hardware features such as redundant power supplies and high-quality components, the software used in this architecture detects failures and directs requests appropriately. The architecture leverages the fact that most commodity PC hardware can now be assumed to be quite reliable. You may observe parallels between this assumption and the assumption made when using cyclic redundancy check (CRC)—that most modern networks can be assumed to be quite reliable so it is acceptable for the underlying network to discard defective packets.

Replication

When commodity PCs are used to serve large workloads, a large number of machines are inherently necessary to process the user requests. Therefore, replication is already built into the system design from the ground up. When the software can detect and respond to failures, high availability is achieved at almost no additional cost.

Price-Focused Design Is Preferable to Performance-Focused Design

Hardware that provides the best computing performance per unit price is preferable to hardware that provides the best absolute performance. Generally, this allows more computational resources to be directed at user tasks for a given budget.

[5] In electronics, the term *hot* refers to systems that are currently powered or active. By contrast, *cold* systems are not currently powered or active.

[6] L. A. Barroso, J. Dean, and U. Holzle, "Web Search for a Planet: The Google Cluster Architecture," *IEEE Micro*, 23(2) March–April 2003: 22–28.

Achieving High Availability at the Application/Middleware Level

While the high-availability hardware designs prevent catastrophes that can take longer times (days or weeks) to recover from, high-availability designs at the application and database level prevent data loss in running applications. High-availability design at the application and middleware level includes paying attention to network disconnections between the application server and end users, application server failure, and database server failure.

Application Server High-Availability Architecture

Redundancy continues to be the basic premise of high-availability application server architectures. Redundancy is commonly achieved by using clusters consisting of redundant application server nodes with failover policies like "1 of N," whereby the application server runs on 1 out of N available server machines; or "static," whereby the application server runs on a dedicated server machine. *A cluster is a collection of servers (called nodes), any of which can run the workloads common to the cluster*.

Within the cluster, application servers provide continuous replication, which provides data availability for the system. This is done by automatically backing up transaction log files to one or more copies located on a second set of disks, preferably on another server that is preferably in another location. The copies of the database transaction log files are used to replay the completed transactions on a copy of the database, thereby keeping the databases consistent without having to be synchronous with each other.

When a server in the cluster experiences a hardware or software failure, the management software in the cluster detects it and starts this service on another node in the cluster. Applications running on clusters need to be cluster-aware and capable of responding to these signals from the management software. They need to be capable of taking over immediately, without any need for server restarts and replaying transactions, other than those that were "in-flight" at the point in time the primary server failed.

A representative architecture for achieving high-application server availability is illustrated in Figure 17-2.[7] Five nodes are organized in two clusters and one server running both the cluster manager and logs. The nodes on the left are the primary application servers, while the nodes on the right serve as backups. The shared server is mounted to all nodes so that the transaction log files are accessible from all nodes, with appropriate permissions. Provided that both nodes of a cluster do not fail simultaneously, and that each node is independently capable of serving user needs for the applications served by the cluster, the architecture will improve availability significantly.

Database Server High-Availability Features

A typical high-availability configuration for a web application is shown in Figure 17-3. We see that a failure in the database can independently hurt availability even if the application servers are available. Therefore, enterprise database servers have a comprehensive set of HA capabilities. High-availability features in the database typically include data protection, disaster recovery, real-time data integration and replication, and support for multiple hardware and operating-system platforms.

Three mechanisms commonly support high-availability database operations: database mirroring, failover clustering, and replication.

[7] Udo Pletat, "High Availability in a J2EE Enterprise Application Environment," http://ceur-ws.org/Vol-141/paper13.pdf (accessed Feb. 2020).

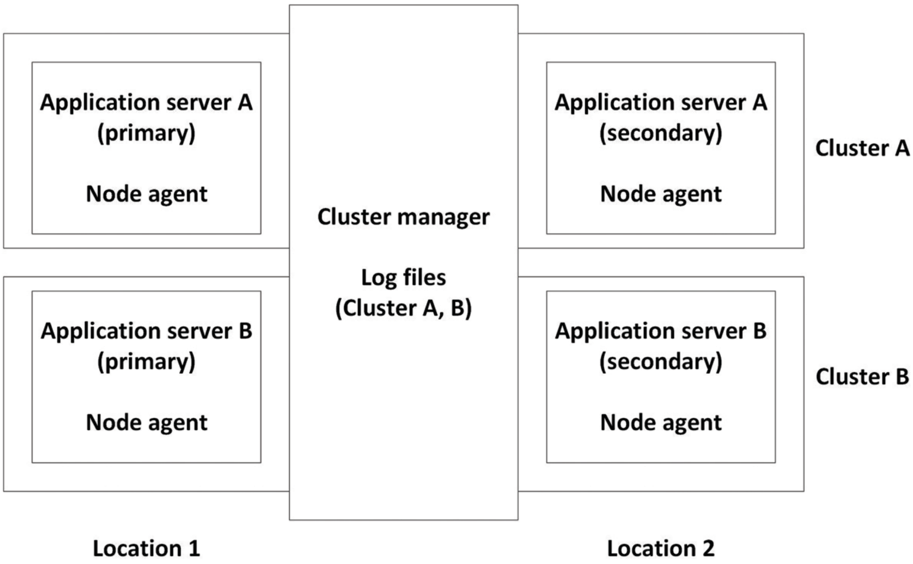


Figure 17-2. High-availability application server configuration example

Database Mirroring

Database mirroring is a software solution for providing almost instantaneous failover with no loss of committed data. It is now a standard feature of most enterprise database server software to provide high availability. Database mirroring can be used to maintain a single standby database (mirror database) for a corresponding production database (principal database). The mirror database is often also used for various reporting needs without disturbing the customer-facing principal database. Mirror databases are also for snapshots, to obtain read-only access to data as it existed at the time when the snapshot was created.

In industrial use, database mirroring is run in either synchronous operation in high-safety mode or asynchronous operation in high-performance mode. In high-performance mode, the transactions commit without waiting for the mirror server to write the log to disk, which maximizes performance. In high-safety mode, a transaction is completed only when it is committed on both partners, but this usually reduces the transaction speeds of the database.

Failover Clustering

Failover clustering is an architecture that provides redundancy through a configuration in which other servers essentially act as clones of the main production system. A failover cluster comprises one or more servers (nodes) with a set of shared cluster disks. The shared disk array is configured to allow all nodes access to the disk resources, but with only one node actively processing data. When a server node fails, the failover cluster automatically moves control of the shared resources to a working node. This configuration allows seamless failover capabilities in the event of a CPU, memory, or other hardware failure that does not affect storage.

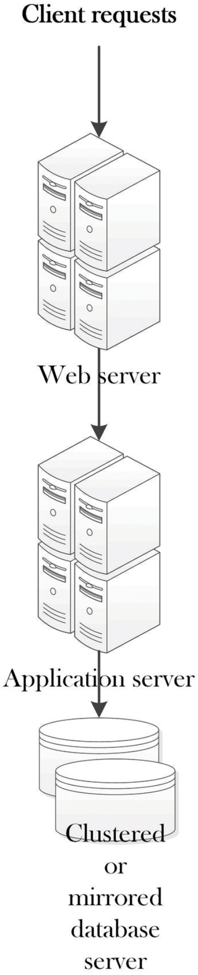


Figure 17-3. Typical web application architecture for high availability

Replication

Database replication allows two or more database servers to stay "in sync" so that the secondary servers can answer queries and potentially actually change data. If data on the secondary servers is changed, it is merged during synchronization.

Replication can be used to allow "slices" of a database to be replicated between several sites. The "slice" can be a set of database objects (i.e., tables) or even parts of a table, such as only certain specific rows (horizontal slicing) or only certain columns. While replication is primarily a technology for making data available off-site and to consolidate data to central sites, it also can be used for high availability or for disaster recoverability.

High Availability at the Operating System Level

Failover clusters, a technology we saw in the context of databases, also can be provided by the operating system for high availability by reducing single points of failure. Failover clusters are often used for key databases, file sharing on a network, business applications, and customer services such as e-commerce websites. There are two basic types of clustering technologies at the operating system (OS) level: network load balancing clusters and failover clusters. The choice of which technology to use generally depends upon the service being delivered by the cluster. For the purpose of availability, services may be characterized as stateless or stateful. *Stateless transactions are self-contained transactions, requiring no awareness by the server of the prior history of transactions. Stateful workloads require awareness of the history of the transaction to complete successfully*. Since stateful services require an awareness of the history of a transaction, they are more challenging for availability.

Network Load Balancing for Stateless Workloads

Network load balancing is an effective, scalable means for achieving high availability for stateless server workloads. Client requests handled before a given client request have no impact on that current transaction. The standard example of a stateless transaction is a web request. The http protocol has been designed such that each request for a web page is self-contained, and when responding to the request, the web server gathers all necessary information to present the page to the client. Upon delivering the response, the server discards all these resources collected to respond to the request. When the user clicks on a link on the page to continue the transaction, the click contains all necessary information for the web server to respond, and it is treated as a new stand-alone request from the client. In this situation, since each request supplies all the information needed by the server to fulfill the request, any given request can be processed by any instance of a server with access to the same underlying data.

Network load-balanced clusters leverage this flexibility to automatically distribute incoming requests among available computers. If a server in a NLB cluster fails unexpectedly, only the active connections to the failed server are lost. However, more commonly, NLB greatly simplifies bringing hosts down intentionally for planned maintenance. Upon completion of maintenance, servers can be added back to the cluster to resume sharing workloads.

Failover Clusters for Stateful Workloads

The prototypical example of a stateful workload is a database transaction. When responding to a database request, the server could potentially alter the data. However, since reading and writing information from and to storage is slow, a request involving a series of transactions is first completed in memory and the final results are recorded back in storage. If there are problems during the sequence, all transactions can be reverted. Thus previous client requests in a sequence can influence subsequent transactions if all transactions in the sequence are to be consistent with each other.

Operating systems can facilitate failover clusters. If the primary node in a clustered application fails or that node is taken offline for maintenance, the cluster manager will start the clustered application on a backup cluster node. The operating system can now immediately redirect requests for resources to the backup cluster node to minimize the impact of the failure.

Multisite Clusters

While failover clusters improve availability, servers in close proximity to each other are vulnerable to natural disasters, power failures, and wide-area network (WAN) outages (which can themselves be caused by construction accidents). In these situations, all cluster nodes can be disabled at once. Multisite clusters can mitigate this risk. Clusters nodes can be connected through LAN or WAN links, providing high availability and disaster recovery. If any given node in a multisite cluster fails, subsequent requests are directed to a node at one of the available sites.

Business Continuity and Disaster Recovery

"Snow Blizzard Shutting Down NYC," "Earthquake in California Damaging Buildings," "Super Storm Sandy Wiping Out the New Jersey Boardwalk": These headlines are all too common these days, and the media narrative makes it appear that the storms are getting larger, more frequent, and more destructive. How does this affect you as an IT professional?

As an IT professional, one of your primary responsibilities is IT service delivery, as we saw earlier in this chapter. IT is in every corner of just about every organization today. In small businesses, IT may be as simple as a router provided by the ISP, some servers for data storage, and a handful of desktops or laptops and printers. In larger organizations such as your university, IT can include hundreds of applications running on hundreds of servers across multiple load-balanced locations.

Regardless of how simple or complex your IT environment is, you need to plan for business disruptions. Without such a plan, events such as a local power outage or a tornado, hurricane, or earthquake can keep you disconnected from the market for extended periods, while your clients shift to your better-prepared competitors, hurting your business. Unfortunately, the data suggests that companies may be unprepared for disasters. For example, almost 75% of respondents in a recent survey had no backup plan in case their phone lines went down.[8] There are many reasons for this, including a lack of awareness, time, resources, or sense of urgency. Hopefully, by the end of this chapter, you will understand the importance of being better prepared.

Business continuity planning (BCP) and disaster recovery are processes used by IT organizations to prepare for emergencies. *Business continuity planning is the methodology used to create and validate a plan for maintaining continuous business operations before, during, and after disruptive events*. The subset of BCP that is used to restore the affected services is called disaster recovery. *Disaster recovery is the set of procedures used to restore technology services that were disrupted during an extreme event*.

BCP is an important component in evaluating competing technology choices in many high-value industries. For example, large financial institutions, utility companies, health care organizations, credit card processing companies, mainstream media companies, and high-volume online retailers may decide that they cannot tolerate even a few minutes of downtime under any circumstances. Such downtime could cause large losses for businesses or could put lives at stake at the hospital. These operational requirements will therefore justify the costs of fully redundant systems as part of BCP.

Disaster recovery usually involves several discreet steps in the planning stages, though those steps blur quickly during implementation because the situation during a crisis is almost never

[8] https://www.information-age.com/two-thirds-companies-dont-have-telecoms-back-plan-123458335/ (accessed Feb. 2020).

exactly as planned. Disaster recovery involves stopping the effects of the disaster as quickly as possible and addressing the immediate aftermath. This might include shutting down systems that have been breached, evaluating which systems are impacted by a flood or earthquake, and determining the best way to proceed.

Figure 17-4 shows the timeline for business continuity and disaster recovery activities around a disaster event. Business continuity activities are ongoing throughout the event. During active disaster recovery, business continuity activities include determining where to set up temporary systems, how to procure replacement systems or parts, and how to set up security in a new location. Finally, at the end of the event, once normal operations are resumed, the lessons learned are used to revise the continuity and recovery plans for more efficient response during the next event.

Planning for Business Continuity and Disaster Recovery

The role of IT professionals is unique in BC/DR. On the one hand, they are not responsible for the company's comprehensive BC/DR planning, but on the other hand, technology is so integral to most corporate operations that continued IT operations is one of the core concerns of BC/DR. As a result, a holistic view of the organization allows IT to determine the most appropriate role of the IT group within the organization. Many elements of a BC/DR plan extend beyond the walls of the IT department. For example, corporate communications will keep stakeholders informed, the power company can try to restore power to critical buildings, and so on. The BC/DR project team thus requires expertise in several areas. Some typical areas are shown in Table 17-2.

The BC/DR plans are usually saved as documents and include the anticipated scenarios, contact information for the key personnel involved, detailed recovery procedures for the identified scenarios, and target recovery times.

Understanding BC/DR Requirements

When you are ready to develop a BC/DR plan, you can invoke ideas from your systems analysis and design class. An effective plan starts with understanding the client's requirements. The BC/DR requirements rely strongly on the organization's specific location(s), industry, and operations-specific details to shortlist the types of disasters and events most relevant to the plan. As the plan is being developed, information that is likely to be helpful includes the following:

Table 17-2. Subject matter expertise needed for BC/DR planning

IT	Facilities	Specialty
Data centers	Office spaces	Off-site data storage
IT infrastructure	Production facilities	Critical data/records
End-user IT (laptops)	Manufacturing facilities	Critical equipment
Voice and data communications	Inventory storage areas	Critical equipment

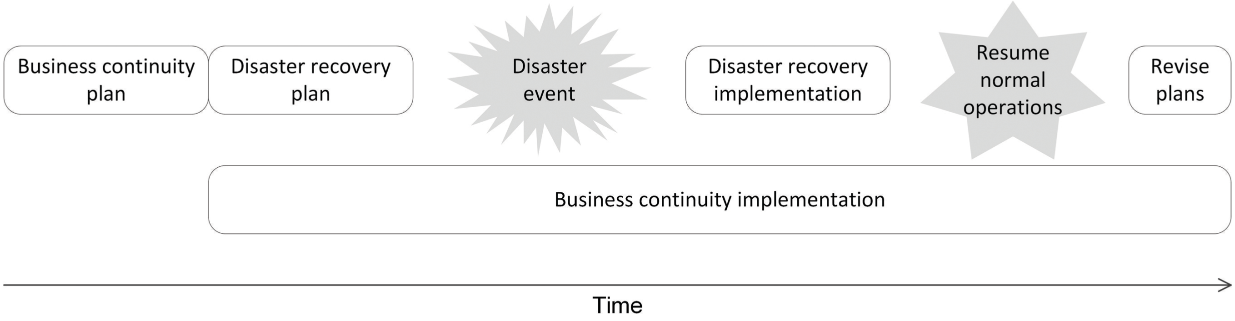


Figure 17-4. Business continuity and disaster-recovery cycle

- Which system functions are critical?
- How much downtime is tolerable?
- How much data loss is tolerable in the event of a server or facility failure (if logs that would enable point-in-time recovery are lost)?
- Is the client willing to pay for an off-site recovery site that's always available?
- Could the client accept only critical functions being available in the case of a serious disaster? Which ones and for how long?

You may already know many of these answers from organizational documentation, but often, the BC/DR planning exercise is the first time the organization develops clarity on these concerns. An open dialog with the client can be helpful in developing this information.

BC/DR Plan Maintenance

Once the high-level BC/DR plan is developed, the disaster recovery components can be developed, which include detailed information that will likely be needed during a disaster (e.g., backup and recovery scripts, parameter files, etc.). The full package (backup/recovery plan plus recovery components) should be jointly reviewed by business stakeholders and IT to ensure that the BC/DR package is consistent and coherent. The package can then be used when required for a real emergency.

The BC/DR package should be tested at initial implementation time and regularly thereafter for as many scenarios as possible. The tests often require cooperation from other teams, so they are complex exercises that cannot be done frivolously.

Summary

In this chapter, we looked at the core concepts underlying industry-scale delivery of IT services. Some popular frameworks were introduced and the underlying ideas behind a popular service management framework, ITIL, were discussed. We identified high availability as a key concern for businesses and looked at how high availability is commonly achieved at the hardware, middleware, and operating system levels.

Finally, we introduced the idea of planning for disasters during peacetime. Business continuity planning is an ongoing activity that aims to help the organization continue to serve customers when unanticipated problems happen.

About the Colophon

When it comes to meeting high availability expectations, even our mythological characters take no chances. IT professionals can learn a lesson or two from our favorite saint.

Review Questions

1. What is *IT services management*? What is its objective?
2. What are some common frameworks that can be used for ITSM? What are their principal advantages and disadvantages?
3. What are the five service delivery disciplines as defined by ITILv2? In your opinion, which of these disciplines is most important for efficient service delivery? Why?
4. What is *service-level management*? Why is it important?
5. What is a *service level agreement*? Why is it useful?
6. What are some of the important components of a SLA? In your opinion, which of these

components is most important for effective relations between provider and client? Why?

7. What is *capacity management*? Why is it important?
8. What is *continuity management*? Why is it important?
9. What is *availability management*? Why is it important?
10. What is IT financial management? Why is it important?
11. What is *high availability*? Why is it important?
12. What are the four characteristics of high-availability system designs?
13. What are some of the important costs associated with transitioning to high-availability systems?
14. What is *business impact analysis*? How does it affect high availability?
15. What is the cost of downtime? How does it affect high availability?
16. How are local high-availability solutions different from disaster-recovery solutions? How are they similar?
17. What are active-active systems? When would you use an active-active high-availability system over an active-passive high-availability system?
18. What are *fault-tolerant servers*?
19. What is an *application server cluster*? How does it improve availability?
20. What are the common ways of achieving high-availability database operations?
21. What is *business continuity planning*?
22. What is *disaster recovery*? How is it related to business continuity?
23. What is the distinction between stateful and stateless workloads? How do they affect high-availability design?
24. What are *multisite clusters*? How do they improve availability?
25. What is the natural disaster most likely to occur in your area? If you were preparing a business continuity plan for your organization, what are some of the most useful items of information you can find?

EXAMPLE CASE: *Chaos Monkey at Netflix*

It is now common knowledge that many of the most interesting consumer services that have emerged in recent years have been developed by technology companies that have brought together some of the smartest minds on the planet to build products and services used by tens of millions of users around the world. Google, Facebook, and Amazon are examples of these organizations.

Less well known is the fact that many of these companies also maintain interesting blogs that document the engineering that goes behind making these services highly available. These blogs are very detailed, and the students reading this book are bound to find them very interesting. Since they are generally so well-written, we will, for the most part, simply refer students to these blogs and limit this discussion to providing context where appropriate.

Chaos Monkey sounds like an odd name for a technology designed to improve availability. But this is the name given by Netflix to technology it has developed to deliberately fail components in live customer-facing systems and observe the behavior of their infrastructure in response. Chaos Monkey helps Netflix discover problems in its infrastructure and anticipate availability problems before they occur. The team summarizes the technology with the line, “Do you think your applications can handle a troop of mischievous monkeys loose in your infrastructure? Now you can find out.” You can read more about Chaos Monkey on the Netflix tech blog.[9] You can also download Chaos Monkey from GitHub.[10]

Example Case Exercise Questions

1. What is *Chaos Monkey*? What are some of its features?
2. How can technology such as Chaos Monkey help improve service delivery?
3. Look at the tech blog of one of the leading Internet technology companies that maintain such a blog. Briefly summarize the most recent article on the blog.

[9] https://netflixtechblog.com/netflix-chaos-monkey-upgraded-1d679429be5d (accessed Feb. 2020).

[10] https://netflix.github.io/chaosmonkey/ (accessed Feb. 2020).

HANDS-ON EXERCISE: *Device Uptime*

In this simple exercise, we will use command-line utilities to obtain the uptime for common operating systems. In Windows, use the command `net stats srv`, and on the Mac, use the command `uptime`.

Hands-on Exercise Question

1. What is the uptime on your machine?

CRITICAL THINKING EXERCISE: *Personal High Availability*

Are you personally prepared for high availability? Let us complete this exercise with the goal of helping you achieve your most important current priority. For most readers of this chapter, this priority is likely to be to find a job that can lead to a productive career. Use the following questions as a framework to minimize hurdles in achieving this goal:

1. What is your most important priority right now?
2. What are the top three hurdles you can foresee in achieving this priority?
3. What would be the most effective and cost-effective ways by which you can minimize these hurdles?

IT INFRASTRUCTURE DESIGN EXERCISE: *Including Active Replication*[11]

Initially, TrendyWidget assumed that AWS guaranteed high availability. But after its CIO read about AWS's NoSQL outage,[12] it decided to maintain a two-region active-active replication of its data.

Answer the following questions and update your infrastructure diagram as suggested:

1. What are the locations of some AWS regions?
2. Say TrendyWidgets decides to locate its servers in the US East and US West locations. Update your infrastructure diagram to indicate these locations. For simplicity, show a direct connection from TrendyWidget to the US East location and use any informal representation (say dotted lines) to indicate the internal AWS network.

[11] A popular resource for information on data centers, an essential part of modern IT infrastructures, is the data center knowledge blog: http://www.datacenterknowledge.com (accessed Apr. 2016).

[12] http://www.techrepublic.com/article/aws-outage-how-netflix-weathered-the-storm-by-preparing-for-the-worst/ (accessed Feb. 2016).