2

THE PHYSICAL LAYER

In this chapter, we look at the lowest layer in our reference model, the physical layer. It defines the electrical, timing, and other interfaces by which bits are sent as signals over channels. The physical layer is the foundation on which the network is built. The properties of different kinds of physical channels determine the performance (e.g., throughput, latency, and error rate) so it is a good place to start our journey into network-land.

We will begin by introducing three kinds of transmission media: guided or wired (e.g., copper, coaxial cable, fiber optics), wireless (terrestrial radio), and satellite. Each of these technologies has different properties that affect the design and performance of the networks that use them. This material provides background information on the key transmission technologies used in modern networks.

We then cover a theoretical analysis of data transmission, only to discover that Mother (Parent?) Nature puts some limits on what can be sent over a communications channel (i.e., a physical transmission medium used to send bits). Next comes digital modulation, which is all about how analog signals are converted into digital bits and back. After that we will look at multiplexing schemes, exploring how multiple conversations can be put on the same transmission medium at the same time without interfering with one another.

Finally, we will look at three examples of communication systems used in practice for wide area computer networks: the (fixed) telephone system, the mobile phone system, and the cable television system. Each of these is important in practice, so we will devote a fair amount of space to each one.

2.1 GUIDED TRANSMISSION MEDIA

The purpose of the physical layer is to transport bits from one machine to another. Various physical media can be used for the actual transmission. Transmission media that rely on a physical cable or wire are often called **guided transmission media** because the signal transmissions are guided along a path with a physical cable or wire. The most common guided transmission media are copper cable (in the form of coaxial cable or twisted pair) and fiber optics. Each type of guided transmission media has its own set of trade-offs in terms of frequency, bandwidth, delay, cost, and ease of installation and maintenance. Bandwidth is a measure of the carrying capacity of a medium. It is measured in **Hz** (or MHz or GHz). It is named in honor of the German physicist Heinrich Hertz. We will discuss this in detail later in this chapter.

2.1.1 Persistent Storage

One of the most common ways to transport data from one device to another is to write them onto persistent storage, such as magnetic or solid-state storage (e.g., recordable DVDs), physically transport the tape or disks to the destination machine, and read them back in again. Although this method is not as sophisticated as using a geosynchronous communication satellite, it is often more cost effective, especially for applications where a high data rate or cost per bit transported is the key factor.

A simple calculation will make this point clear. An industry-standard Ultrium tape can hold 30 terabytes. A box 60 × 60 × 60 cm can hold about 1000 of these tapes, for a total capacity of 800 terabytes, or 6400 terabits (6.4 petabits). A box of tapes can be delivered anywhere in the United States in 24 hours by Federal Express and other companies. The effective bandwidth of this transmission is 6400 terabits/86,400 sec, or a bit over 70 Gbps. If the destination is only an hour away by road, the bandwidth is increased to over 1700 Gbps. No computer network can even approach this. Of course, networks are getting faster, but tape densities are increasing, too.

If we now look at cost, we get a similar picture. The cost of an Ultrium tape is around $40 when bought in bulk. A tape can be reused at least 10 times, so the tape cost is maybe $4000 per box per usage. Add to this another $1000 for shipping (probably much less), and we have a cost of roughly $5000 to ship 800 TB. This amounts to shipping a gigabyte for a little over half a cent. No network can beat that. The moral of the story is:

> *Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway.*

For moving *very* large amounts of data, this is often the best solution. Amazon has what it calls the ‘‘Snowmobile,’’ which is a large truck filled with thousands of

hard disks, all connected to a high-speed network inside the truck. The total capacity of the truck is 100 PB (100,000 TB or 100 million GB). When a company has a huge amount of data to move, it can have the truck come to its premises and plug into the company's fiber-optic network, then suck out all the data into the truck. Once that it is done, the truck drives to another location and disgorges all the data. For example, a company wishing to replace its own massive datacenter with the Amazon cloud might be interested in this service. For very large volumes of data, no other method of data transport can even approach this.

2.1.2 Twisted Pairs

Although the bandwidth characteristics of persistent storage are excellent, the delay characteristics are poor: Transmission time is measured in hours or days, not milliseconds. Many applications, including the Web, video conferencing, and online gaming, rely on transmitting data with low delay. One of the oldest and still most common transmission media is **twisted pair**. A twisted pair consists of two insulated copper wires, typically about 1 mm thick. The wires are twisted together in a helical form, similar to a DNA molecule. Two parallel wires constitute a fine antenna; when the wires are twisted, the waves from different twists cancel out, so the wire radiates less effectively. A signal is usually carried as the difference in voltage between the two wires in the pair. Transmitting the signal as the difference between the two voltage levels, as opposed to an absolute voltage, provides better immunity to external noise because the noise tends to affect the voltage traveling through both wires in the same way, leaving the differential relatively unchanged.

The most common application of the twisted pair is the telephone system. Nearly all telephones are connected to the telephone company (telco) office by a twisted pair. Both telephone calls and ADSL Internet access run over these lines. Twisted pairs can run several kilometers without amplification, but for longer distances the signal becomes too attenuated and repeaters are needed. When many twisted pairs run in parallel for a substantial distance, such as all the wires coming from an apartment building to the telephone company office, they are bundled together and encased in a protective sheath. The pairs in these bundles would interfere with one another if it were not for the twisting. In parts of the world where telephone lines run on poles above ground, it is common to see bundles several centimeters in diameter.

Twisted pairs can be used for transmitting either analog or digital information. The bandwidth depends on the thickness of the wire and the distance traveled, but hundreds of megabits/sec can be achieved for a few kilometers, in many cases, and more when various tricks are used. Due to their adequate performance, widespread availability, and low cost, twisted pairs are widely used and are likely to remain so for years to come.

Twisted-pair cabling comes in several varieties. One common variety of twisted-pair cables now deployed in many buildings is called **Category 5e** cabling, or

''Cat 5e.'' A Category 5e twisted pair consists of two insulated wires gently twisted together. Four such pairs are typically grouped in a plastic sheath to protect the wires and keep them together. This arrangement is shown in Fig. 2-1.

Figure 2-1. Category 5e UTP cable with four twisted pairs. These cables can be used for local area networks.

Different LAN standards may use the twisted pairs differently. For example, 100-Mbps Ethernet uses two (out of the four) pairs, one pair for each direction. To reach higher speeds, 1-Gbps Ethernet uses all four pairs in both directions simultaneously, which requires the receiver to factor out the signal that is transmitted.

Some general terminology is now in order. Links that can be used in both directions at the same time, like a two-lane road, are called **full-duplex** links. In contrast, links that can be used in either direction, but only one way at a time, like a single-track railroad line, are called **half-duplex** links. A third category consists of links that allow traffic in only one direction, like a one-way street. They are called **simplex** links.

Returning to twisted pair, Cat 5 replaced earlier **Category 3** cables with a similar cable that uses the same connector, but has more twists per meter. More twists result in less crosstalk and a better-quality signal over longer distances, making the cables more suitable for high-speed computer communication, especially 100-Mbps and 1-Gbps Ethernet LANs.

New wiring is more likely to be **Category 6** or even **Category 7**. These categories have more stringent specifications to handle signals with greater bandwidths. Some cables in Category 6 and above can support the 10-Gbps links that are now commonly deployed in many networks, such as in new office buildings. **Category 8** wiring runs at higher speeds than the lower categories, but operates only at short distances of around 30 meters and is thus only suitable in data centers. The Category 8 standard has two options: Class I, which is compatible with Category 6A; and Class II, which is compatible with Category 7A.

Through Category 6, these wiring types are referred to as **UTP** (**Unshielded Twisted Pair**) as they consist simply of wires and insulators. In contrast to these, Category 7 cables have shielding on the individual twisted pairs, as well as around the entire cable (but inside the plastic protective sheath). Shielding reduces the susceptibility to external interference and crosstalk with other nearby cables to meet demanding performance specifications. The cables are reminiscent of the

high-quality, but bulky and expensive shielded twisted pair cables that IBM introduced in the early 1980s. However, these did not prove popular outside of IBM installations. Evidently, it is time to try again.

2.1.3 Coaxial Cable

Another common transmission medium is the **coaxial cable** (known to its many friends as just ''coax'' and pronounced ''co-ax''). It has better shielding and greater bandwidth than unshielded twisted pairs, so it can span longer distances at higher speeds. Two kinds of coaxial cable are widely used. One kind, 50-ohm cable, is commonly used when it is intended for digital transmission from the start. The other kind, 75-ohm cable, is commonly used for analog transmission and cable television. This distinction is based on historical, rather than technical, factors (e.g., early dipole antennas had an impedance of 300 ohms, and it was easy to use existing 4:1 impedance-matching transformers). Starting in the mid-1990s, cable TV operators began to provide Internet access over cable, which has made 75-ohm cable more important for data communication.

A coaxial cable consists of a stiff copper wire as the core, surrounded by an insulating material. The insulator is encased by a cylindrical conductor, often as a closely woven braided mesh. The outer conductor is covered in a protective plastic sheath. A cutaway view of a coaxial cable is shown in Fig. 2-2.

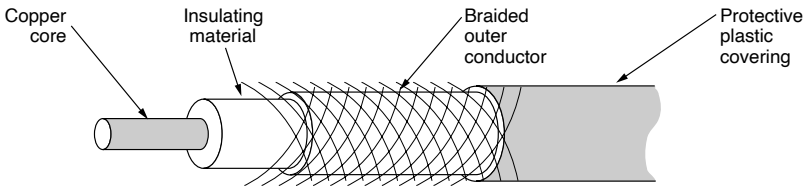


Figure 2-2. A coaxial cable.

The construction and shielding of the coaxial cable give it a good combination of high bandwidth and excellent noise immunity (e.g., from garage door openers, microwave ovens, and more). The bandwidth possible depends on the cable quality and length. Coaxial cable has extremely wide bandwidth; modern cables have a bandwidth of up to 6 GHz, thus allowing many conversations to be simultaneously transmitted over a single coaxial cable (a single television program might occupy approximately 3.5 MHz). Coaxial cables were once widely used within the telephone system for long-distance lines but have now largely been replaced by fiber optics on long-haul routes. Coax is still widely used for cable television and metropolitan area networks and is also used for delivering high-speed Internet connectivity to homes in many parts of the world.

2.1.4 Power Lines

The telephone and cable television networks are not the only sources of wiring that can be reused for data communication. There is a yet more common kind of wiring: electrical power lines. Power lines deliver electrical power to houses, and electrical wiring within houses distributes the power to electrical outlets.

The use of power lines for data communication is an old idea. Power lines have been used by electricity companies for low-rate communication such as remote metering for many years, as well in the home to control devices (e.g., the X10 standard). In recent years there has been renewed interest in high-rate communication over these lines, both inside the home as a LAN and outside the home for broadband Internet access. We will concentrate on the most common scenario: using electrical wires inside the home.

The convenience of using power lines for networking should be clear. Simply plug a TV and a receiver into the wall, which you must do anyway because they need power, and they can send and receive movies over the electrical wiring. This configuration is shown in Fig. 2-3. There is no other plug or radio. The data signal is superimposed on the low-frequency power signal (on the active or ''hot'' wire) as both signals use the wiring at the same time.

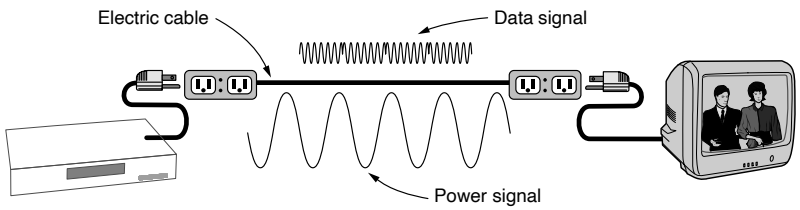


Figure 2-3. A network that uses household electrical wiring.

The difficulty with using household electrical wiring for a network is that it was designed to distribute power signals. This task is quite distinct from distributing data signals, at which household wiring does a horrible job. Electrical signals are sent at 50–60 Hz and the wiring attenuates the much higher frequency (MHz) signals needed for high-rate data communication. The electrical properties of the wiring vary from one house to the next and change as appliances are turned on and off, which causes data signals to bounce around the wiring. Transient currents when appliances switch on and off create electrical noise over a wide range of frequencies. And without the careful twisting of twisted pairs, electrical wiring acts as a fine antenna, picking up external signals and radiating signals of its own. This behavior means that to meet regulatory requirements, the data signal must avoid licensed frequencies such as the amateur radio bands.

Despite these difficulties, it is practical to send at least 500 Mbps short distances over typical household electrical wiring by using communication schemes that resist impaired frequencies and bursts of errors. Many products use proprietary standards for power-line networking, but standards are being developed.

2.1.5 Fiber Optics

More than a few people in the computer industry take enormous pride in how fast computer technology is improving as it follows Moore's law, which predicts a doubling of the number of transistors per chip roughly every 2 years (Kuszyk and Hammoudeh, 2018). The original (1981) IBM PC ran at a clock speed of 4.77 MHz. Forty years later, PCs could run a four-core CPU at 3 GHz. This increase is of a factor of around 2500. Impressive.

In the same period, wide area communication links went from 45 Mbps (a T3 line in the telephone system) to 100 Gbps (a modern long-distance line). This gain is similarly impressive, more than a factor of 2000, while at the same time the error rate went from 10^{-5} per bit to almost zero. In the past decade, single CPUs have approached physical limits, which is why the number of CPU cores per chip is being increased. In contrast, the achievable bandwidth with fiber technology is in excess of 50,000 Gbps (50 Tbps) and we are nowhere near reaching these limits. The current practical limit of around 100 Gbps is simply due to our inability to convert between electrical and optical signals any faster. To build higher-capacity links, many channels are simply carried in parallel over a single fiber.

In this section, we will study fiber optics to learn how that transmission technology works. In the ongoing race between computing and communication, communication may yet win because of fiber-optic networks. The implication of this would be essentially infinite bandwidth and a new conventional wisdom that computers are hopelessly slow so that networks should try to avoid computation at all costs, no matter how much bandwidth that wastes. This change will take a while to sink in to a generation of computer scientists and engineers taught to think in terms of the low transmission limits imposed by copper wires.

Of course, this scenario does not tell the whole story because it does not include cost. The cost to install fiber over the last mile to reach consumers and bypass the low bandwidth of wires and limited availability of spectrum is tremendous. It also costs more energy to move bits than to compute. We may always have islands of inequities where either computation or communication is essentially free. For example, at the edge of the Internet we apply computation and storage to the problem of compressing and caching content, all to make better use of Internet access links. Within the Internet, we may do the reverse, with companies such as Google moving huge amounts of data across the network to where it is cheaper to perform storage or computation.

Fiber optics are used for long-haul transmission in network backbones, high-speed LANs (although so far, copper has often managed to catch up eventually),

and high-speed Internet access such as fiber to the home. An optical transmission system has three key components: the light source, the transmission medium, and the detector. Conventionally, a pulse of light indicates a 1 bit and the absence of light indicates a 0 bit. The transmission medium is an ultra-thin fiber of glass. The detector generates an electrical pulse when light falls on it. By attaching a light source to one end of an optical fiber and a detector to the other, we have a unidirectional (i.e., simplex) data transmission system that accepts an electrical signal, converts and transmits it by light pulses, and then reconverts the output to an electrical signal at the receiving end.

This transmission system would leak light and be useless in practice were it not for an interesting principle of physics. When a light ray passes from one medium to another—for example, from fused silica (glass) to air—the ray is refracted (bent) at the silica/air boundary, as shown in Fig. 2-4(a). Here we see a light ray incident on the boundary at an angle α_1 emerging at an angle β_1. The amount of refraction depends on the properties of the two media (in particular, their indices of refraction). For angles of incidence above a certain critical value, the light is refracted back into the silica; none of it escapes into the air. Thus, a light ray incident at or above the critical angle is trapped inside the fiber, as shown in Fig. 2-4(b), and can propagate for many kilometers with virtually no loss.

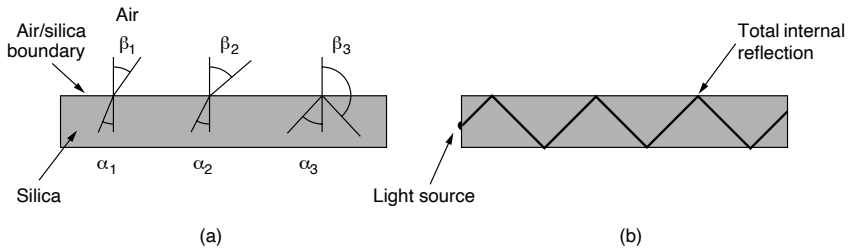


Figure 2-4. (a) Three examples of a light ray from inside a silica fiber impinging on the air/silica boundary at different angles. (b) Light trapped by total internal reflection.

The sketch of Fig. 2-4(b) shows only one trapped ray, but since any light ray incident on the boundary above the critical angle will be reflected internally, many different rays will be bouncing around at different angles. Each ray is said to have a different mode, so a fiber having this property is called a **multimode fiber**. If the fiber's diameter is reduced to a few wavelengths of light (less than 10 microns, as opposed to more than 50 microns for multimode fiber), the fiber acts like a waveguide and the light can propagate only in a straight line, without bouncing, yielding a **single-mode fiber**. Single-mode fibers are more expensive but are widely used for longer distances; they can transmit signals approximately 50 times

farther than multimode fibers. Currently available single-mode fibers can transmit data at 100 Gbps for 100 km without amplification. Even higher data rates have been achieved in the laboratory for shorter distances. The choice between single-mode or multimode fiber depends on the application. Multimode fiber can be used for transmissions of up to about 15 km and can allow the use of relatively less expensive fiber-optic equipment. On the other hand, the bandwidth of multimode fiber becomes more limited as distance increases.

Transmission of Light Through Fiber

Optical fibers are made of glass, which, in turn, is made from sand, an inexpensive raw material available in unlimited amounts. Glassmaking was known to the ancient Egyptians, but their glass had to be no more than 1 mm thick or the light could not shine through. Glass transparent enough to be useful for windows was developed during the Renaissance. The glass used for modern optical fibers is so transparent that if the oceans were full of it instead of water, the seabed would be as visible from the surface as the ground is from an airplane on a clear day.

The attenuation of light through glass depends on the wavelength of the light (as well as on some of the physical properties of the glass). It is defined as the ratio of input to output signal power. For the kind of glass used in fibers, the attenuation is shown in Fig. 2-5 in units of decibels (dB) per linear kilometer of fiber. As an example, a factor of two loss of signal power corresponds to an attenuation of $10 \log_{10} 2 = 3$ dB. We will discuss decibels shortly. In brief, it is a logarithmic way to measure power ratios, with 3 dB meaning a factor of two power ratio. The figure shows the near-infrared part of the spectrum, which is what is used in practice. Visible light has slightly shorter wavelengths, from about 0.4 to 0.7 microns. (1 micron is 10^{-6} meters.) The true metric purist would refer to these wavelengths as 400 nm to 700 nm, but we will stick with traditional usage.

Three wavelength bands are most commonly used at present for optical communication. They are centered at 0.85, 1.30, and 1.55 microns, respectively. All three bands are 25,000 to 30,000 GHz wide. The 0.85-micron band was used first. It has higher attenuation and so is used for shorter distances, but at that wavelength the lasers and electronics could be made from the same material (gallium arsenide). The last two bands have good attenuation properties (less than 5% loss per kilometer). The 1.55-micron band is now widely used with erbium-doped amplifiers that work directly in the optical domain.

Light pulses sent down a fiber spread out in length as they propagate. This spreading is called **chromatic dispersion**. The amount of it is wavelength dependent. One way to keep these spread-out pulses from overlapping is to increase the distance between them, but this can be done only by reducing the signaling rate. Fortunately, it has been discovered that making the pulses in a special shape related to the reciprocal of the hyperbolic cosine causes nearly all the dispersion effects to cancel out, so it is now possible to send pulses for thousands of kilometers without

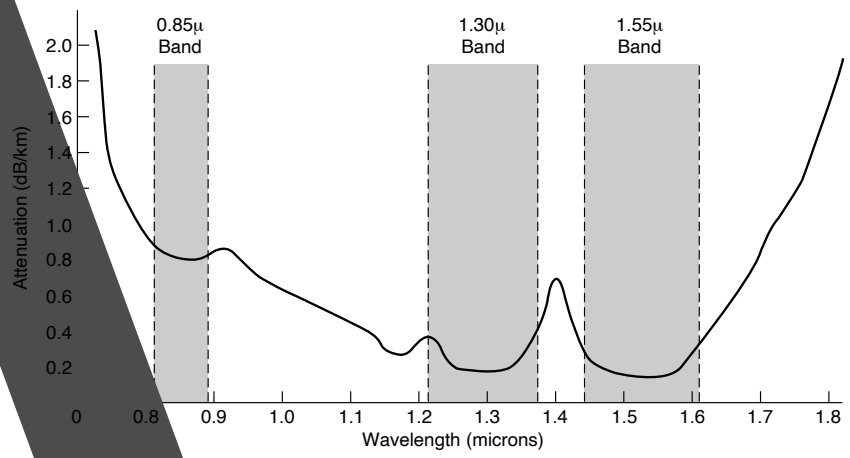


Figure 2-5. Attenuation of light through fiber in the infrared region.

appreciable shape distortion. These pulses are called **solitons**. They are starting to be widely used in practice.

Fiber Cables

Fiber-optic cables are similar to coax, except without the braid. Figure 2-6(a) shows a single fiber viewed from the side. At the center is the glass core through which the light propagates. In multimode fibers, the core is typically around 50 microns in diameter, about the thickness of a human hair. In single-mode fibers, the core is 8 to 10 microns.

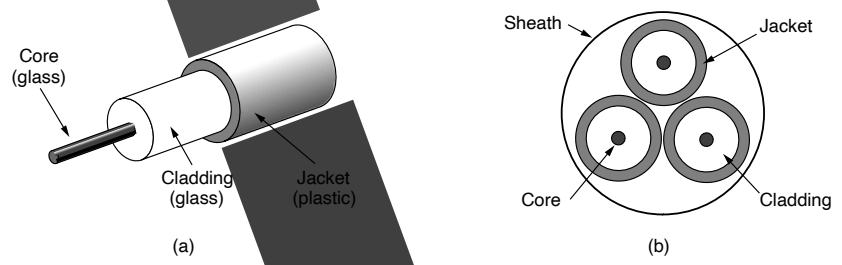


Figure 2-6. (a) Side view of a single fiber. (b) End view of a sheath with three fibers.

The core is surrounded by a glass cladding with a lower index of refraction than the core, to keep all the light in the core. Next comes a thin plastic jacket to

protect the cladding. Fibers are typically grouped in bundles, protected by an outer sheath. Figure 2-6(b) shows a sheath with three fibers.

Terrestrial fiber sheaths are normally laid in the ground within a meter of the surface, where they are occasionally subject to attacks by backhoes or gophers. Near the shore, transoceanic fiber sheaths are buried in trenches by a kind of sea-plow. In deep water, they just lie on the bottom, where they can be snagged by fishing trawlers or attacked by a giant squid.

Fibers can be connected in three different ways. First, they can terminate in connectors and be plugged into fiber sockets. Connectors lose about 10 to 20% of the light, but they make it easy to reconfigure systems. Second, they can be spliced mechanically. Mechanical splices just lay the two carefully cut ends next to each other in a special sleeve and clamp them in place. Alignment can be improved by passing light through the junction and then making small adjustments to maximize the signal. Mechanical splices take trained personnel about 5 minutes and result in a 10% light loss. Third, two pieces of fiber can be fused (melted) to form a solid connection. A fusion splice is almost as good as a single drawn fiber, but even here, a small amount of attenuation occurs. For all three kinds of splices, reflections can occur at the point of the splice and the reflected energy can interfere with the signal.

Two kinds of light sources are typically used to do the signaling: LEDs (Light Emitting Diodes) and semiconductor lasers. They have different properties, as shown in Fig. 2-7. They can be tuned in wavelength by inserting Fabry-Perot or Mach-Zehnder interferometers between the source and the fiber. Fabry-Perot interferometers are simple resonant cavities consisting of two parallel mirrors. The light is incident perpendicular to the mirrors. The length of the cavity selects out those wavelengths that fit inside an integral number of times. Mach-Zehnder interferometers separate the light into two beams. The two beams travel slightly different distances. They are recombined at the end and are in phase for only certain wavelengths.

Item	LED	Semiconductor laser
Data rate	Low	High
Fiber type	Multi-mode	Multi-mode or single-mode
Distance	Short	Long
Lifetime	Long life	Short life
Temperature sensitivity	Minor	Substantial
Cost	Low cost	Expensive

Figure 2-7. A comparison of semiconductor diodes and LEDs as light sources.

The receiving end of an optical fiber consists of a photodiode, which gives off an electrical pulse when struck by light. The response time of photodiodes, which convert the signal from the optical to the electrical domain, limits data rates to

about 100 Gbps. Thermal noise is also an issue, so a pulse of light must carry enough energy to be detected. By making the pulses powerful enough, the error rate can be made arbitrarily small.

Comparison of Fiber Optics and Copper Wire

It is instructive to compare fiber to copper. Fiber has many advantages. To start with, it can handle much higher bandwidths than copper. This alone would require its use in high-end networks. Due to the low attenuation, repeaters are needed only about every 50 km on long lines, versus about every 5 km for copper, resulting in a big cost saving. Fiber also has the advantage of not being affected by power surges, electromagnetic interference, or power failures. Nor is it affected by corrosive chemicals in the air, important for harsh factory environments.

Oddly enough, telephone companies like fiber for a completely different reason: it is thin and lightweight. Many existing cable ducts are completely full, so there is no room to add new capacity. Removing all the copper and replacing it with fiber empties the ducts, and the copper has excellent resale value to copper refiners who regard it as very high-grade ore. Also, fiber is much lighter than copper. One thousand twisted pairs 1 km long weigh 8000 kg. Two fibers have more capacity and weigh only 100 kg, which reduces the need for expensive mechanical support systems that must be maintained. For new routes, fiber wins hands down due to its much lower installation cost. Finally, fibers do not leak light and are difficult to tap. These properties give fiber good security against wiretappers.

On the downside, fiber is a less familiar technology requiring skills not all engineers have, and fibers can be damaged easily by being bent too much. Since optical transmission is inherently unidirectional, two-way communication requires either two fibers or two frequency bands on one fiber. Finally, fiber interfaces cost more than electrical interfaces. Nevertheless, the future of all fixed data communication over more than short distances is clearly with fiber. For a discussion of many aspects of fiber optics and their networks, see Pearson (2015).

2.2 WIRELESS TRANSMISSION

Many people now have wireless connectivity to many devices, from laptops and smartphones, to smart watches and smart refrigerators. All of these devices rely on wireless communication to transmit information to other devices and endpoints on the network.

In the following sections, we will look at wireless communication in general, which has many other important applications besides providing connectivity to users who want to surf the Web from the beach. Wireless has advantages for even fixed devices in some circumstances. For example, if running a fiber to a building is difficult due to the terrain (mountains, jungles, swamps, etc.), wireless may be

more appropriate. It is noteworthy that modern wireless digital communication began as a research project of Prof. Norman Abramson of the University of Hawaii in the 1970s where the Pacific Ocean separated the users from their computer center, and the telephone system was inadequate. We will discuss this system, ALOHA, in Chap. 4.

2.2.1 The Electromagnetic Spectrum

When electrons move, they create electromagnetic waves that can propagate through space (even in a vacuum). These waves were predicted by the British physicist James Clerk Maxwell in 1865 and first observed by the German physicist Heinrich Hertz in 1887. The number of oscillations per second of a wave is called its **frequency**, f, and is measured in Hz. The distance between two consecutive maxima (or minima) is called the **wavelength**, which is universally designated by the Greek letter λ (lambda).

When an antenna of the appropriate size is attached to an electrical circuit, the electromagnetic waves can be broadcast efficiently and received by a receiver some distance away. All wireless communication is based on this principle.

In a vacuum, all electromagnetic waves travel at the same speed, no matter what their frequency. This speed, usually called the **speed of light**, c, is approximately 3×10^8 m/sec, or about 1 foot (30 cm) per nanosecond. (A case could be made for redefining the foot as the distance light travels in a vacuum in 1 nsec rather than basing it on the shoe size of some long-dead king.) In copper or fiber, the speed slows to about 2/3 of this value and becomes slightly frequency dependent. The speed of light is the universe's ultimate speed limit. No object or signal can ever move faster than it.

The fundamental relation between f, λ, and c (in a vacuum) is

$$\lambda f = c \tag{2-1}$$

Since c is a constant, if we know f, we can find λ, and vice versa. As a rule of thumb, when λ is in meters and f is in MHz, $\lambda f \approx 300$. For example, 100-MHz waves are about 3 meters long, 1000-MHz waves are 0.3 meters long, and 0.1-meter waves have a frequency of 3000 MHz.

The electromagnetic spectrum is shown in Fig. 2-8. The radio, microwave, infrared, and visible light portions of the spectrum can all be used for transmitting information by modulating the amplitude, frequency, or phase of the waves. Ultraviolet light, X-rays, and gamma rays would be even better, due to their higher frequencies, but they are hard to produce and modulate, do not propagate well through buildings, and are dangerous to living things.

The bands listed at the bottom of Fig. 2-8 are the official ITU (International Telecommunication Union) names and are based on the wavelengths, so the LF band goes from 1 km to 10 km (approximately 30 kHz to 300 kHz). The terms LF,

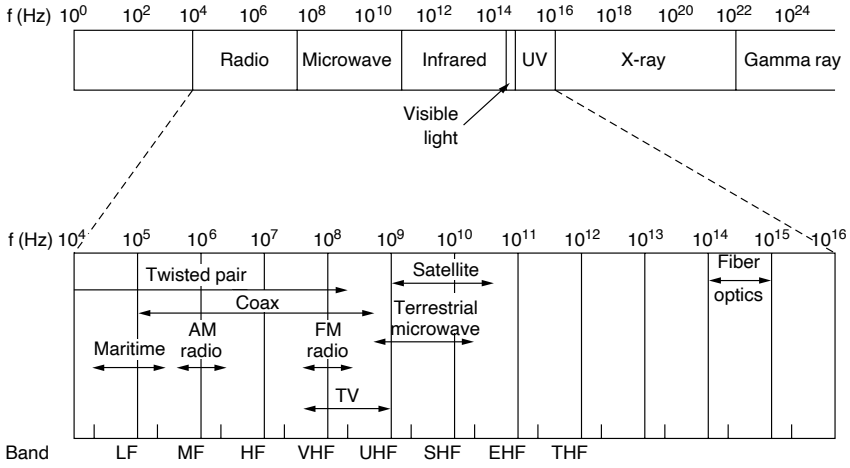


Figure 2-8. The electromagnetic spectrum and its uses for communication.

MF, and HF refer to Low, Medium, and High Frequency, respectively. Clearly, when the names were assigned nobody expected to go above 10 MHz, so the higher bands were later named the Very, Ultra, Super, Extremely, and Tremendously High Frequency bands. Beyond that, there are no names, but Incredibly, Astonishingly, and Prodigiously High Frequency (IHF, AHF, and PHF) would sound nice. Above 10^{12} Hz, we get into the infrared, where the comparison is typically to light, not radio.

The theoretical basis for communication, which we will discuss later in this chapter, tells us the amount of information that a signal such as an electromagnetic wave can carry depends on the received power and is proportional to its bandwidth. From Fig. 2-8, it should now be obvious why networking people like fiber optics so much. Many GHz of bandwidth are available to tap for data transmission in the microwave band, and even more bandwidth is available in fiber because it is further to the right in our logarithmic scale. As an example, consider the 1.30-micron band of Fig. 2-5, which has a width of 0.17 microns. If we use Eq. (2-1) to find the start and end frequencies from the start and end wavelengths, we find the frequency range to be about 30,000 GHz. With a reasonable signal-to-noise ratio of 10 dB, this is 300 Tbps.

Most transmissions use a relatively narrow frequency band, in other words, $\Delta f/f \ll 1$). They concentrate their signal power in this narrow band to use the spectrum efficiently and obtain reasonable data rates by transmitting with enough power. The rest of this section describes three different types of transmission that make use of wider frequency bands.

2.2.2 Frequency Hopping Spread Spectrum

In **frequency hopping spread spectrum**, a transmitter hops from frequency to frequency hundreds of times per second. It is popular for military communication because it makes transmissions hard to detect and next to impossible to jam. It also offers good resistance to fading due to signals taking different paths from source to destination and interfering after recombining. It also offers resistance to narrowband interference because the receiver will not be stuck on an impaired frequency for long enough to shut down communication. This robustness makes it useful for crowded parts of the spectrum, such as the ISM bands we will describe shortly. This technique is used commercially, for example, in Bluetooth and older versions of 802.11.

As a curious footnote, the technique was co-invented by the Austrian-born film star Hedy Lamarr, who was famous for acting in European films in the 1930s under her birth name of Hedwig (Hedy) Kiesler. Her first husband was a wealthy armaments manufacturer who told her how easy it was to block the radio signals then used to control torpedoes. When she discovered that he was selling weapons to Hitler, she was horrified, disguised herself as a maid to escape him, and fled to Hollywood to continue her career as a movie actress. In her spare time, she invented frequency hopping to help the Allied war effort.

Her scheme used 88 frequencies, the number of keys (and frequencies) on the piano. For their invention, she and her friend, the musical composer George Antheil, received U.S. patent 2,292,387. However, they were unable to convince the U.S. Navy that their invention had any practical use and never received any royalties. Only years after the patent expired was the technique rediscovered and used in mobile electronic devices rather than for blocking signals to torpedoes during war time.

2.2.3 Direct Sequence Spread Spectrum

A second form of spread spectrum, **direct sequence spread spectrum**, uses a code sequence to spread the data signal over a wider frequency band. It is widely used commercially as a spectrally efficient way to let multiple signals share the same frequency band. These signals can be given different codes, a method called code division multiple access that we will return to later in this chapter. This method is shown in contrast with frequency hopping in Fig. 2-9. It forms the basis of 3G mobile phone networks and is also used in GPS (Global Positioning System). Even without different codes, direct sequence spread spectrum, like frequency hopping spread spectrum, can tolerate interference and fading because only a fraction of the desired signal is lost. It is used in this role in older versions of the 802.11b wireless LANs protocol. For a fascinating and detailed history of spread spectrum communication, see Walters (2013).

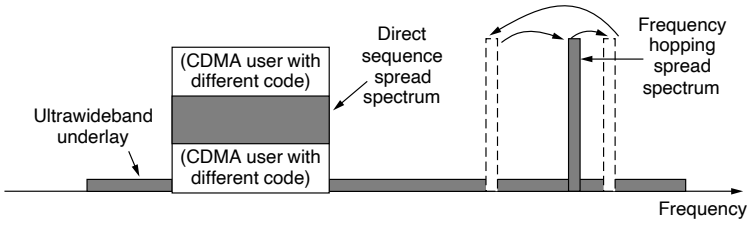


Figure 2-9. Spread spectrum and ultra-wideband (UWB) communication.

2.2.4 Ultra-Wideband Communication

UWB (**Ultra-WideBand**) communication sends a series of low-energy rapid pulses, varying their carrier frequencies to communicate information. The rapid transitions lead to a signal that is spread thinly over a very wide frequency band. UWB is defined as signals that have a bandwidth of at least 500 MHz or at least 20% of the center frequency of their frequency band. UWB is also shown in Fig. 2-9. With this much bandwidth, UWB has the potential to communicate at several hundred megabits per second. Because it is spread across a wide band of frequencies, it can tolerate a substantial amount of relatively strong interference from other narrowband signals. Just as importantly, since UWB has very little energy at any given frequency when used for short-range transmission, it does not cause harmful interference to those other narrowband radio signals. In contrast to spread spectrum transmission, UWB transmits in ways that do not interfere with the carrier signals in the same frequency band. It can also be used for imaging through solid objects (ground, walls, and bodies) or as part of precise location systems. The technology is popular for short-distance indoor applications, as well as precision radar imaging and location-tracking technologies.

2.3 USING THE SPECTRUM FOR TRANSMISSION

We will now discuss how the various parts of the electromagnetic spectrum of Fig. 2-8 are used, starting with radio. We will assume that all transmissions use a narrow frequency band unless otherwise stated.

2.3.1 Radio Transmission

Radio frequency (RF) waves are easy to generate, can travel long distances, and can penetrate buildings easily, so they are widely used for communication, both indoors and outdoors. Radio waves also are omnidirectional, meaning that

they travel in all directions from the source, so the transmitter and receiver do not have to be carefully aligned physically.

Sometimes omni-directional radio is good, but sometimes it is bad. In the 1970s, General Motors decided to equip all its new Cadillacs with computer-controlled anti-lock brakes. When the driver stepped on the brake pedal, the computer pulsed the brakes on and off instead of locking them on hard. One fine day an Ohio Highway Patrolman began using his new mobile radio to call headquarters, and suddenly the Cadillac next to him began behaving like a bucking bronco. When the officer pulled the car over, the driver claimed that he had done nothing and that the car had gone crazy.

Eventually, a pattern began to emerge: Cadillacs would sometimes go berserk, but only on major highways in Ohio and then only when the Highway Patrol was there watching. For a long, long time General Motors could not understand why Cadillacs worked fine in all the other states and also on minor roads in Ohio. Only after much searching did they discover that the Cadillac's wiring made a fine antenna for the frequency used by the Ohio Highway Patrol's new radio system.

The properties of radio waves are frequency dependent. At low frequencies, radio waves pass through obstacles well, but the power falls off sharply with distance from the source—at least as fast as $1/r^2$ in air—as the signal energy is spread more thinly over a larger surface. This attenuation is called **path loss**. At high frequencies, radio waves tend to travel in straight lines and bounce off obstacles. Path loss still reduces power, though the received signal can depend strongly on reflections as well. High-frequency radio waves are also absorbed by rain and other obstacles to a larger extent than are low-frequency ones. At all frequencies, radio waves are subject to interference from motors and other electrical equipment.

It is interesting to compare the attenuation of radio waves to that of signals in guided media. With fiber, coax, and twisted pair, the signal drops by the same fraction per unit distance, for example, 20 dB per 100 m for twisted pair. With radio, the signal drops by the same fraction as the distance doubles, for example 6 dB per doubling in free space. This behavior means that radio waves can travel long distances, and interference between users is a problem. For this reason, all governments tightly regulate the use of radio transmitters, with few notable exceptions, which are discussed later in this chapter.

In the VLF, LF, and MF bands, radio waves follow the ground, as illustrated in Fig. 2-10(a). These waves can be detected for perhaps 1000 km at the lower frequencies, less at the higher ones. AM radio broadcasting uses the MF band, which is why the ground waves from Boston AM radio stations cannot be heard easily in New York. Radio waves in these bands pass through buildings easily, which is why radios work indoors. The main problem with using these bands for data communication is their low bandwidth.

In the HF and VHF bands, the ground waves tend to be absorbed by the earth. However, the waves that reach the ionosphere, a layer of charged particles circling the earth at a height of 100 to 500 km, are refracted by it and sent back to earth, as

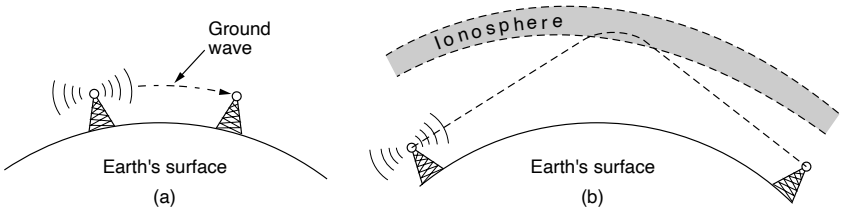


Figure 2-10. (a) In the VLF, LF, and MF bands, radio waves follow the curvature of the earth. (b) In the HF band, they bounce off the ionosphere.

shown in Fig. 2-10(b). Under certain atmospheric conditions, the signals can bounce several times. Amateur radio operators (hams) use these bands to talk long distance. The military also uses the HF and VHF bands for communication.

2.3.2 Microwave Transmission

Above 100 MHz, the waves travel in nearly straight lines and can therefore be narrowly focused. Concentrating all the energy into a small beam by means of a parabolic antenna (like the familiar satellite TV dish) gives a much higher signal-to-noise ratio, but the transmitting and receiving antennas must be accurately aligned with each other. In addition, this directionality allows multiple transmitters lined up in a row to communicate with multiple receivers in a row without interference, provided some minimum spacing rules are observed. Before fiber optics, for decades these microwaves formed the heart of the long-distance telephone transmission system. In fact, MCI, one of AT&T's first competitors after it was deregulated, built its entire system with microwave communications passing between towers tens of kilometers apart. Even the company's name reflected this (MCI stood for Microwave Communications, Inc.). MCI has since gone over to fiber and through a long series of corporate mergers and bankruptcies in the telecommunications shuffle has become part of Verizon.

Microwaves are **directional**: they travel in a straight line, so if the towers are too far apart, the earth will get in the way (think about a Seattle-to-Amsterdam link). Thus, repeaters are needed periodically. The higher the towers are, the farther apart they can be. The distance between repeaters goes up roughly with the square root of the tower height. For 100-meter towers, repeaters can be 80 km apart.

Unlike radio waves at lower frequencies, microwaves do not pass through buildings well. In addition, even though the beam may be well focused at the transmitter, there is still some divergence in space. Some waves may be refracted off low-lying atmospheric layers and may take slightly longer to arrive than the

direct waves. The delayed waves may arrive out of phase with the direct wave and thus cancel the signal. This effect is called **multipath fading** and is often a serious problem. It is weather and frequency dependent. Some operators keep 10% of their channels idle as spares to switch on when multipath fading temporarily wipes out a particular frequency band.

The demand for higher data rates is driving wireless network operators to yet higher frequencies. Bands up to 10 GHz are now in routine use, but at around 4 GHz, a new problem sets in: absorption by water. These waves are only a few centimeters long and are absorbed by rain. This effect would be fine if one were planning to build a huge outdoor microwave oven for roasting passing birds, but for communication it is a severe problem. As with multipath fading, the only solution is to shut off links that are being rained on and route around them.

In summary, microwave communication is so widely used for long-distance telephone communication, mobile phones, television distribution, and other purposes that a severe shortage of spectrum has developed. It has several key advantages over fiber. The main one is that no right of way is needed to lay down cables. By buying a small plot of ground every 50 km and putting a microwave tower on it, one can bypass the telephone system entirely. This is how MCI managed to get started as a new long-distance telephone company so quickly. (Sprint, another early competitor to the deregulated AT&T, went a completely different route: it was formed by the Southern Pacific Railroad, which already owned a large amount of right of way and just buried fiber next to the tracks.)

Microwave is also relatively inexpensive. Putting up two simple towers (which can be just big poles with four guy wires) and putting antennas on each one may be cheaper than burying 50 km of fiber through a congested urban area or up over a mountain, and it may also be cheaper than leasing the telephone company's fiber, especially if the telephone company has not yet even fully paid for the copper it ripped out when it put in the fiber.

2.3.3 Infrared Transmission

Unguided infrared waves are widely used for short-range communication. The remote controls used for televisions, Blu-ray players, and stereos all use infrared communication. They are relatively directional, cheap, and easy to build but have a major drawback: they do not pass through solid objects. (Try standing between your remote control and your television and see if it still works.) In general, as we go from long-wave radio toward visible light, the waves behave more and more like light and less and less like radio.

On the other hand, the fact that infrared waves do not pass through solid walls well is also a plus. It means that an infrared system in one room of a building will not interfere with a similar system in adjacent rooms or buildings: you cannot control your neighbor's television with your remote control. Furthermore, security of infrared systems against eavesdropping is better than that of radio systems on

account of this reason. Therefore, no government license is needed to operate an infrared system, in contrast to radio systems, which must be licensed outside the ISM bands. Infrared communication has a limited use on the desktop, for example, to connect notebook computers and printers with the **IrDA** (**Infrared Data Association**) standard, but it is not a major player in the communication game.

2.3.4 Light Transmission

Unguided optical signaling or **free-space optics** has been in use for centuries. Paul Revere used binary optical signaling from the Old North Church just prior to his famous ride. A more modern application is to connect the LANs in two buildings via lasers mounted on their rooftops. Optical signaling using lasers is inherently unidirectional, so each end needs its own laser and its own photodetector. This scheme offers very high bandwidth at very low cost and is relatively secure because it is difficult to tap a narrow laser beam. It is also relatively easy to install and, unlike microwave transmission, does not require a license from the **FCC** (**Federal Communications Commission**) in the United States and analogous government bodies in other countries.

The laser's strength, a very narrow beam, is also its weakness here. Aiming a laser beam 1 mm wide at a target the size of a pin head 500 meters away requires the marksmanship of a latter-day Annie Oakley. Usually, lenses are put into the system to defocus the beam slightly. To add to the difficulty, wind and temperature changes can distort the beam and laser beams also cannot penetrate rain or thick fog, although they normally work well on sunny days. However, many of these factors are not an issue when the use is to connect two spacecraft.

One of the authors (AST) once attended a conference at a modern hotel in Europe in the 1990s at which the conference organizers thoughtfully provided a room full of terminals to allow the attendees to read their email during boring presentations. Since the local phone company was unwilling to install a large number of telephone lines for just 3 days, the organizers put a laser on the roof and aimed it at their university's computer science building a few kilometers away. They tested it the night before the conference and it worked perfectly. At 9 A.M. the next day, which was bright and sunny, the link failed completely and stayed down all day. The pattern repeated itself the next 2 days. It was not until after the conference that the organizers discovered the problem: heat from the sun during the daytime caused convection currents to rise up from the roof of the building, as shown in Fig. 2-11. This turbulent air diverted the beam and made it dance around the detector, much like a shimmering road on a hot day. The lesson here is that to work well in difficult conditions as well as good conditions, unguided optical links need to be engineered with a sufficient margin of error.

Unguided optical communication may seem like an exotic networking technology today, but it might soon become much more prevalent. In many places, we are surrounded by cameras (that sense light) and displays (that emit light using LEDs

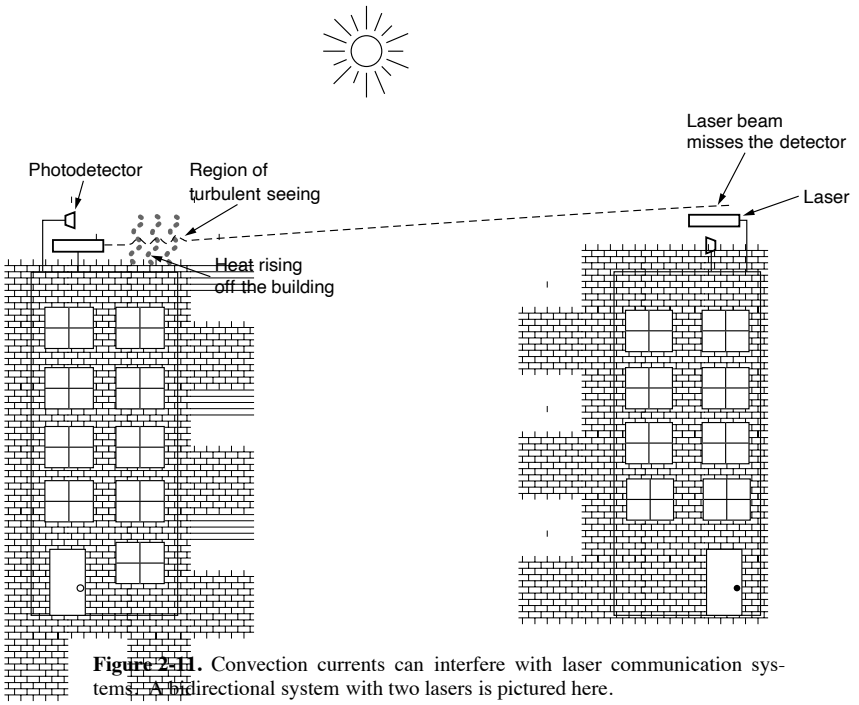


Figure 2-11. Convection currents can interfere with laser communication systems. A bidirectional system with two lasers is pictured here.

and other technology). Data communication can be layered on top of these displays by encoding information in the pattern at which LEDs turn on and off that is below the threshold of human perception. Communicating with visible light in this way is inherently safe and creates a low-speed network in the immediate vicinity of the display. This could enable all sorts of fanciful ubiquitous computing scenarios. The flashing lights on emergency vehicles might alert nearby traffic lights and vehicles to help clear a path. Informational signs might broadcast maps. Even festive lights might broadcast songs that are synchronized with their display.

2.4 FROM WAVEFORMS TO BITS

In this section, we describe how signals are transmitted over the physical media we have discussed. We begin with a discussion of the theoretical basis for data communication, and follow with a discussion of modulation (the process of converting analog waveforms to bits) and multiplexing (which allows a single physical medium to carry multiple simultaneous transmissions).

2.4.1 The Theoretical Basis for Data Communication

Information can be transmitted on wires by varying some physical property such as voltage or current. By representing the value of this voltage or current as a single-valued function of time, $f(t)$, we can model the behavior of the signal and analyze it mathematically. This analysis is the subject of the following sections.

Fourier Analysis

In the early 19th century, the French mathematician Jean-Baptiste Fourier proved that any reasonably behaved periodic function, $g(t)$ with period T, can be constructed as the sum of a (possibly infinite) number of sines and cosines:

$$g(t) = \frac{1}{2}c + \sum_{n=1}^{\infty} a_n \sin(2\pi nft) + \sum_{n=1}^{\infty} b_n \cos(2\pi nft) \qquad (2\text{-}2)$$

where $f = 1/T$ is the fundamental frequency, a_n and b_n are the sine and cosine amplitudes of the nth **harmonics** (terms), and c is a constant that determines the mean value of the function. Such a decomposition is called a **Fourier series**. From the Fourier series, the function can be reconstructed. That is, if the period, T, is known and the amplitudes are given, the original function of time can be found by performing the sums of Eq. (2-2).

A data signal that has a finite duration, which all of them do, can be handled by just imagining that it repeats the entire pattern over and over forever (i.e., the interval from T to $2T$ is the same as from 0 to T, etc.).

The a_n amplitudes can be computed for any given $g(t)$ by multiplying both sides of Eq. (2-2) by $\sin(2\pi kft)$ and then integrating from 0 to T. Since

$$\int_0^T \sin(2\pi kft)\sin(2\pi nft)\,dt = \begin{cases} 0 \text{ for } k \neq n \\ T/2 \text{ for } k = n \end{cases}$$

only one term of the summation survives: a_n. The b_n summation vanishes completely. Similarly, by multiplying Eq. (2-2) by $\cos(2\pi kft)$ and integrating between 0 and T, we can derive b_n. By just integrating both sides of the equation as it stands, we can find c. The results of performing these operations are as follows:

$$a_n = \frac{2}{T}\int_0^T g(t)\sin(2\pi nft)\,dt \qquad b_n = \frac{2}{T}\int_0^T g(t)\cos(2\pi nft)\,dt \qquad c = \frac{2}{T}\int_0^T g(t)\,dt$$

Bandwidth-Limited Signals

The relevance of all of this to data communication is that real channels affect different frequency signals differently. Let us consider a specific example: the transmission of the ASCII character ''b'' encoded in an 8-bit byte. The bit pattern

that is to be transmitted is 01100010. The left-hand part of Fig. 2-12(a) shows the voltage output by the transmitting computer. The Fourier analysis of this signal yields the coefficients:

$$a_n = \frac{1}{\pi n}\left[\cos(\pi n/4) - \cos(3\pi n/4) + \cos(6\pi n/4) - \cos(7\pi n/4)\right]$$

$$b_n = \frac{1}{\pi n}\left[\sin(3\pi n/4) - \sin(\pi n/4) + \sin(7\pi n/4) - \sin(6\pi n/4)\right]$$

$$c = 3/4.$$

The root-mean-square amplitudes, $\sqrt{a_n^2 + b_n^2}$, for the first few terms are shown on the right-hand side of Fig. 2-12(a). These values are of interest because their squares are proportional to the energy transmitted at the corresponding frequency.

No transmission facility can transmit signals without losing some power in the process. If all the Fourier components were equally diminished, the resulting signal would be reduced in amplitude but not distorted [i.e., it would have the same nice squared-off shape as Fig. 2-12(a)]. Unfortunately, all transmission facilities diminish different Fourier components by different amounts, thus introducing distortion. Usually, for a wire, the amplitudes are transmitted mostly undiminished from 0 up to some frequency f_c (measured in Hz) with all frequencies above this cutoff frequency attenuated. The width of the frequency range transmitted without being strongly attenuated is called the **bandwidth**. In practice, the cutoff is not really sharp, so often the quoted bandwidth is from 0 to the frequency at which the received power has fallen by half.

The bandwidth is a physical property of the transmission medium that depends on, for example, the construction, thickness, length, and material of a wire or fiber. Filters are often used to further limit the bandwidth of a signal. 802.11 wireless channels generally use roughly 20 MHz, for example, so 802.11 radios filter the signal bandwidth to this size (although in some cases an 80-MHz band is used).

As another example, traditional (analog) television channels occupy 6 MHz each, on a wire or over the air. This filtering lets more signals share a given region of spectrum, which improves the overall efficiency of the system. It means that the frequency range for some signals will not start at zero, but at some higher number. However, this does not matter. The bandwidth is still the width of the band of frequencies that are passed, and the information that can be carried depends only on this width and not on the starting and ending frequencies. Signals that run from 0 up to a maximum frequency are called **baseband** signals. Signals that are shifted to occupy a higher range of frequencies, as is the case for all wireless transmissions, are called **passband** signals.

Now let us consider how the signal of Fig. 2-12(a) would look if the bandwidth were so low that only the lowest frequencies were transmitted [i.e., if the function were being approximated by the first few terms of Eq. (2-2)]. Figure 2-12(b) shows the signal that results from a channel that allows only the first harmonic (the

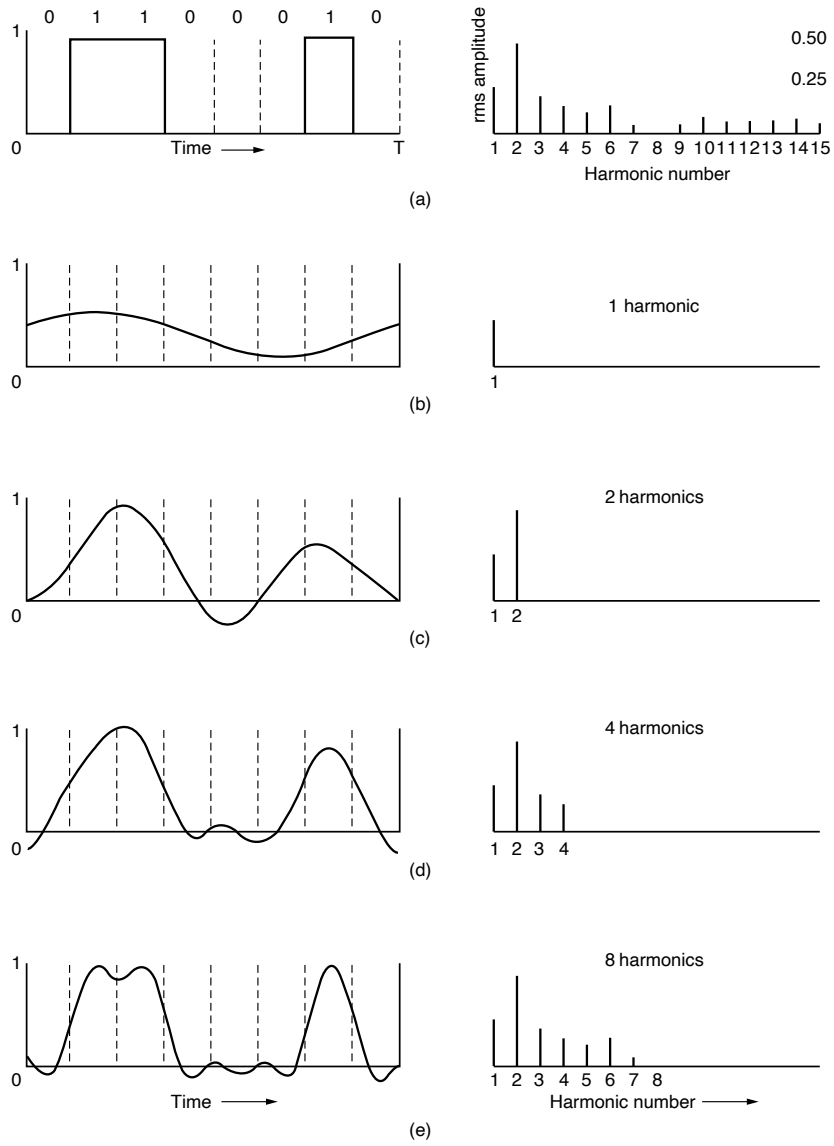


Figure 2-12. (a) A binary signal and its root-mean-square Fourier amplitudes. (b)–(e) Successive approximations to the original signal.

fundamental, *f*) to pass through. Similarly, Fig. 2-12(c)–(e) show the spectra and reconstructed functions for higher-bandwidth channels. For digital transmission, the goal is to receive a signal with just enough fidelity to reconstruct the sequence of bits that was sent. We can already do this easily in Fig. 2-12(e), so it is wasteful to use more harmonics to receive a more accurate replica.

Given a bit rate of b bits/sec, the time required to send the 8 bits in our example 1 bit at a time is $8/b$ sec, so the frequency of the first harmonic of this signal is $b/8$ Hz. An ordinary telephone line, often called a **voice-grade line**, has an artificially introduced cutoff frequency just above 3000 Hz. The presence of this restriction means that the number of the highest harmonic passed through is roughly $3000/(b/8)$, or $24,000/b$ (the cutoff is not sharp).

For some data rates, the numbers work out as shown in Fig. 2-13. From these numbers, it is clear that trying to send at 9600 bps over a voice-grade telephone line will transform Fig. 2-12(a) into something looking like Fig. 2-12(c), making accurate reception of the original binary bit stream tricky. It should be obvious that at data rates much higher than 38.4 kbps, there is no hope at all for *binary* signals, even if the transmission facility is completely noiseless. In other words, limiting the bandwidth limits the data rate, even for perfect channels. However, coding schemes that make use of several voltage levels do exist and can achieve higher data rates. We will discuss these later in this chapter.

Bps	T (msec)	First harmonic (Hz)	# Harmonics sent
300	26.67	37.5	80
600	13.33	75	40
1200	6.67	150	20
2400	3.33	300	10
4800	1.67	600	5
9600	0.83	1200	2
19200	0.42	2400	1
38400	0.21	4800	0

Figure 2-13. Relation between data rate and harmonics for our very simple example.

There is much confusion about bandwidth because it means different things to electrical engineers and to computer scientists. To electrical engineers, (analog) bandwidth is (as we have described above) a quantity measured in Hz. To computer scientists, (digital) bandwidth is the maximum data rate of a channel, a quantity measured in bits/sec. That data rate is the end result of using the analog bandwidth of a physical channel for digital transmission, and the two are related, as we discuss next. In this book, it will be clear from the context whether we mean analog bandwidth (Hz) or digital bandwidth (bits/sec).

2.4.2 The Maximum Data Rate of a Channel

As early as 1924, an AT&T engineer, Harry Nyquist, realized that even a perfect channel has a finite transmission capacity. He derived an equation expressing the maximum data rate for a finite-bandwidth noiseless channel. In 1948, Claude Shannon carried Nyquist's work further and extended it to the case of a channel subject to random (i.e., thermodynamic) noise (Shannon, 1948). This paper is the most important paper in all of information theory. We will just briefly summarize their now classical results here.

Nyquist proved that if an arbitrary signal has been run through a low-pass filter of bandwidth B, the filtered signal can be completely reconstructed by making only $2B$ (exact) samples per second. Sampling the line faster than $2B$ times per second is pointless because the higher-frequency components that such sampling could recover have already been filtered out. If the signal consists of V discrete levels, Nyquist's theorem states:

$$\text{Maximum data rate} = 2B \log_2 V \text{ bits/sec} \tag{2-3}$$

For example, a noiseless 3-kHz channel cannot transmit binary (i.e., two-level) signals at a rate exceeding 6000 bps.

So far we have considered only noiseless channels. If random noise is present, the situation deteriorates rapidly. And there is always random (thermal) noise present due to the motion of the molecules in the system. The amount of thermal noise present is measured by the ratio of the signal power to the noise power, called the **SNR** (**Signal-to-Noise Ratio**). If we denote the signal power by S and the noise power by N, the signal-to-noise ratio is S/N. Usually, the ratio is expressed on a log scale as the quantity $10 \log_{10} S/N$ because it can vary over a tremendous range. The units of this log scale are called **decibels** (**dB**), with ''deci'' meaning 10 and ''bel'' chosen to honor Alexander Graham Bell, who first patented the telephone. An S/N ratio of 10 is 10 dB, a ratio of 100 is 20 dB, a ratio of 1000 is 30 dB, and so on. The manufacturers of stereo amplifiers often characterize the bandwidth (frequency range) over which their products are linear by giving the 3-dB frequency on each end. These are the points at which the amplification factor has been approximately halved (because $10 \log_{10} 0.5 \approx -3$).

Shannon's major result is that the maximum data rate or **capacity** of a noisy channel whose bandwidth is B Hz and whose signal-to-noise ratio is S/N, is given by:

$$\text{Maximum data rate} = B \log_2 (1 + S/N) \textit{bits/sec} \tag{2-4}$$

This equation tells us the best capacities that real channels can have. For example, ADSL (Asymmetric Digital Subscriber Line), which provides Internet access over normal telephone lines, uses a bandwidth of around 1 MHz. The SNR depends strongly on the distance of the home from the telephone exchange, and an SNR of around 40 dB for short lines of 1 to 2 km is very good. With these characteristics,

the channel can never transmit much more than 13 Mbps, no matter how many or how few signal levels are used and no matter how often or how infrequently samples are taken. The original ADSL was specified up to 12 Mbps, though users sometimes saw lower rates. This data rate was actually very good for its time, with over 60 years of communications techniques having greatly reduced the gap between the Shannon capacity and the capacity of real systems.

Shannon's result was derived from information-theory arguments and applies to any channel subject to thermal noise. Counterexamples should be treated in the same category as perpetual motion machines. For ADSL to exceed 12 Mbps, it must either improve the SNR (for example by inserting digital repeaters in the lines closer to the customers) or use more bandwidth, as is done with the evolution to ASDL2+.

2.4.3 Digital Modulation

Now that we have studied the properties of wired and wireless channels, we turn our attention to the problem of sending digital information. Wires and wireless channels carry analog signals such as continuously varying voltage, light intensity, or sound intensity. To send digital information, we must devise analog signals to represent bits. The process of converting between bits and signals that represent them is called **digital modulation**.

We will start with schemes that directly convert bits into a signal. These schemes result in **baseband transmission**, in which the signal occupies frequencies from zero up to a maximum that depends on the signaling rate. It is common for wires. Then we will consider schemes that regulate the amplitude, phase, or frequency of a carrier signal to convey bits. These schemes result in **passband transmission**, in which the signal occupies a band of frequencies around the frequency of the carrier signal. It is common for wireless and optical channels for which the signals must reside in a given frequency band.

Channels are often shared by multiple signals. After all, it is much more convenient to use a single wire to carry several signals than to install a wire for every signal. This kind of sharing is called **multiplexing**. It can be accomplished in several different ways. We will present methods for time, frequency, and code division multiplexing.

The modulation and multiplexing techniques we describe in this section are all widely used for wires, fiber, terrestrial wireless, and satellite channels.

Baseband Transmission

The most straightforward form of digital modulation is to use a positive voltage to represent a 1 bit and a negative voltage to represent a 0 bit, as can be seen in

Fig. 2-14(a). For an optical fiber, the presence of light might represent a 1 and the absence of light might represent a 0. This scheme is called **NRZ** (**Non-Return-to-Zero**). The odd name is for historical reasons, and simply means that the signal follows the data. An example is shown in Fig. 2-14(b).

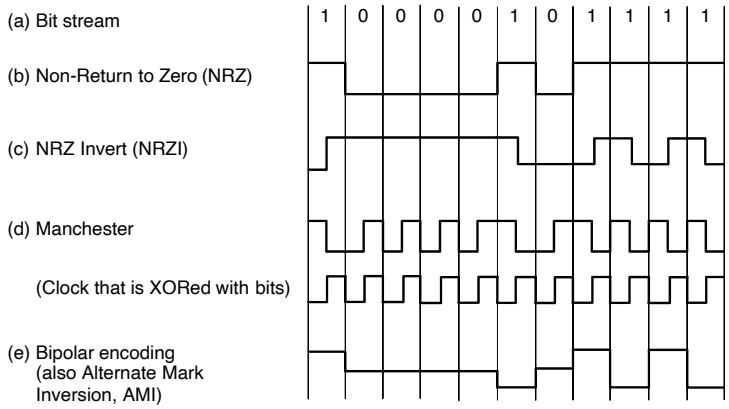


Figure 2-14. Line codes: (a) Bits, (b) NRZ, (c) NRZI, (d) Manchester, (e) Bipolar or AMI.

Once sent, the NRZ signal propagates down the wire. At the other end, the receiver converts it into bits by sampling the signal at regular intervals of time. This signal will not look exactly like the signal that was sent. It will be attenuated and distorted by the channel and noise at the receiver. To decode the bits, the receiver maps the signal samples to the closest symbols. For NRZ, a positive voltage will be taken to indicate that a 1 was sent and a negative voltage will be taken to indicate that a 0 was sent.

NRZ is a good starting point for our studies because it is simple, but it is seldom used by itself in practice. More complex schemes can convert bits to signals that better meet engineering considerations. These schemes are called **line codes**. Below, we describe line codes that help with bandwidth efficiency, clock recovery, and DC balance.

Bandwidth Efficiency

With NRZ, the signal may cycle between the positive and negative levels up to every 2 bits (in the case of alternating 1s and 0s). This means that we need a bandwidth of at least $B/2$ Hz when the bit rate is B bits/sec. This relation comes from the Nyquist rate [Eq. (2-3)]. It is a fundamental limit, so we cannot run NRZ faster without using additional bandwidth. Bandwidth is often a limited resource, even

for wired channels. Higher-frequency signals are increasingly attenuated, making them less useful, and higher-frequency signals also require faster electronics.

One strategy for using limited bandwidth more efficiently is to use more than two signaling levels. By using four voltages, for instance, we can send 2 bits at once as a single **symbol**. This design will work as long as the signal at the receiver is sufficiently strong to distinguish the four levels. The rate at which the signal changes is then half the bit rate, so the needed bandwidth has been reduced.

We call the rate at which the signal changes the **symbol rate** to distinguish it from the **bit rate**. The bit rate is the symbol rate multiplied by the number of bits per symbol. An older name for the symbol rate, particularly in the context of devices called telephone modems that convey digital data over telephone lines, is the **baud rate**. In the literature, the terms ''bit rate'' and ''baud rate'' are often used incorrectly.

Note that the number of signal levels does not need to be a power of two. Often it is not, with some of the levels used for protecting against errors and simplifying the design of the receiver.

Clock Recovery

For all schemes that encode bits into symbols, the receiver must know when one symbol ends and the next symbol begins to correctly decode the bits. With NRZ, in which the symbols are simply voltage levels, a long run of 0s or 1s leaves the signal unchanged. After a while, it is hard to tell the bits apart, as 15 zeros look much like 16 zeros unless you have a very accurate clock.

Accurate clocks would help with this problem, but they are an expensive solution for commodity equipment. Remember, we are timing bits on links that run at many megabits/sec, so the clock would have to drift less than a fraction of a microsecond over the longest permitted run. This might be reasonable for slow links or short messages, but it is not a general solution.

One strategy is to send a separate clock signal to the receiver. Another clock line is no big deal for computer buses or short cables in which there are many lines in parallel, but it is wasteful for most network links since if we had another line to send a signal we could use it to send data. A clever trick here is to mix the clock signal with the data signal by XORing them together so that no extra line is needed. The results are shown in Fig. 2-14(d). The clock makes a clock transition in every bit time, so it runs at twice the bit rate. When it is XORed with the 0 level, it makes a low-to-high transition that is simply the clock. This transition is a logical 0. When it is XORed with the 1 level it is inverted and makes a high-to-low transition. This transition is a logical 1. This scheme is called **Manchester encoding** and was used for classic Ethernet.

The downside of Manchester encoding is that it requires twice as much bandwidth as NRZ due to the clock, and we have learned that bandwidth often matters. A different strategy is based on the idea that we should code the data to ensure that

there are enough transitions in the signal. Consider that NRZ will have clock recovery problems only for long runs of 0s and 1s. If there are frequent transitions, it will be easy for the receiver to stay synchronized with the incoming stream of symbols.

As a step in the right direction, we can simplify the situation by coding a 1 as a transition and a 0 as no transition, or vice versa. This coding is called **NRZI** (**Non-Return-to-Zero Inverted**), a twist on NRZ. An example is shown in Fig. 2-14(c). The popular **USB** (**Universal Serial Bus**) standard for connecting computer peripherals uses NRZI. With it, long runs of 1s do not cause a problem.

Of course, long runs of 0s still cause a problem that we must fix. If we were the telephone company, we might simply require that the sender not transmit too many 0s. Older digital telephone lines in the United States, called T1 lines (discussed later) did, in fact, require that no more than 15 consecutive 0s be sent for them to work correctly. To really fix the problem, we can break up runs of 0s by mapping small groups of bits to be transmitted so that groups with successive 0s are mapped to slightly longer patterns that do not have too many consecutive 0s.

A well-known code to do this is called **4B/5B**. Every 4 bits is mapped into a 5-bit pattern with a fixed translation table. The five bit patterns are chosen so that there will never be a run of more than three consecutive 0s. The mapping is shown in Fig. 2-15. This scheme adds 25% overhead, which is better than the 100% overhead of Manchester encoding. Since there are 16 input combinations and 32 output combinations, some of the output combinations are not used. Putting aside the combinations with too many successive 0s, there are still some codes left. As a bonus, we can use these nondata codes to represent physical layer control signals. For example, in some uses, ''11111'' represents an idle line and ''11000'' represents the start of a frame.

Data (4B)	Codeword (5B)	Data (4B)	Codeword (5B)
0000	11110	1000	10010
0001	01001	1001	10011
0010	10100	1010	10110
0011	10101	1011	10111
0100	01010	1100	11010
0101	01011	1101	11011
0110	01110	1110	11100
0111	01111	1111	11101

Figure 2-15. 4B/5B mapping.

An alternative approach is to make the data look random, known as scrambling. In this case, it is very likely that there will be frequent transitions. A **scrambler** works by XORing the data with a pseudorandom sequence before it is transmitted. This kind of mixing will make the data themselves as random as the

pseudorandom sequence (assuming it is independent of the pseudorandom sequence). The receiver then XORs the incoming bits with the same pseudorandom sequence to recover the real data. For this to be practical, the pseudorandom sequence must be easy to create. It is commonly given as the seed to a simple random number generator.

Scrambling is attractive because it adds no bandwidth or time overhead. In fact, it often helps to condition the signal so that it does not have its energy in dominant frequency components (caused by repetitive data patterns) that might radiate electromagnetic interference. Scrambling helps because random signals tend to be ''white,'' or have energy spread across the frequency components.

However, scrambling does not guarantee that there will be no long runs. It is possible to get unlucky occasionally. If the data are the same as the pseudorandom sequence, they will XOR to all 0s. This outcome does not generally occur with a long pseudorandom sequence that is difficult to predict. However, with a short or predictable sequence, it might be possible for malicious users to send bit patterns that cause long runs of 0s after scrambling and cause links to fail. Early versions of the standards for sending IP packets over SONET links in the telephone system had this defect (Malis and Simpson, 1999). It was possible for users to send certain ''killer packets'' that were guaranteed to cause problems.

Balanced Signals

Signals that have as much positive voltage as negative voltage even over short periods of time are called **balanced signals**. They average to zero, which means that they have no DC electrical component. The lack of a DC component is an advantage because some channels, such as coaxial cable or lines with transformers, strongly attenuate a DC component due to their physical properties. Also, one method of connecting the receiver to the channel called **capacitive coupling** passes only the AC portion of a signal. In either case, if we send a signal whose average is not zero, we waste energy as the DC component will be filtered out.

Balancing helps to provide transitions for clock recovery since there is a mix of positive and negative voltages. It also provides a simple way to calibrate receivers because the average of the signal can be measured and used as a decision threshold to decode symbols. With unbalanced signals, the average may drift away from the true decision level due to a density of 1s, for example, which would cause more symbols to be decoded with errors.

A straightforward way to construct a balanced code is to use two voltage levels to represent a logical 1 and a logical zero. For example, +1 V for a 1 bit and −1 V for a 0 bit. To send a 1, the transmitter alternates between the +1 V and −1 V levels so that they always average out. This scheme is called **bipolar encoding**. In telephone networks, it is called **AMI** (**Alternate Mark Inversion**), building on old terminology in which a 1 is called a ''mark'' and a 0 is called a ''space.'' An example is given in Fig. 2-14(e).

Bipolar encoding adds a voltage level to achieve balance. Alternatively, we can use a mapping like 4B/5B to achieve balance (as well as transitions for clock recovery). An example of this kind of balanced code is the **8B/10B** line code. It maps 8 bits of input to 10 bits of output, so it is 80% efficient, just like the 4B/5B line code. The 8 bits are split into a group of 5 bits, which is mapped to 6 bits, and a group of 3 bits, which is mapped to 4 bits. The 6-bit and 4-bit symbols are then concatenated. In each group, some input patterns can be mapped to balanced output patterns that have the same number of 0s and 1s. For example, ''001'' is mapped to ''1001,'' which is balanced. But there are not enough combinations for all output patterns to be balanced. For these cases, each input pattern is mapped to two output patterns. One will have an extra 1 and the alternate will have an extra 0. For example, ''000'' is mapped to both ''1011'' and its complement ''0100.'' As input bits are mapped to output bits, the encoder remembers the **disparity** from the previous symbol. The disparity is the total number of 0s or 1s by which the signal is out of balance. The encoder then selects either an output pattern or its alternate to reduce the disparity. With 8B/10B, the disparity will be at most 2 bits. Thus, the signal will never be far from balanced. There will also never be more than five consecutive 1s or 0s, to help with clock recovery.

Passband Transmission

Communication over baseband frequencies is most appropriate for wired transmissions, such as twisted pair, coax, or fiber. In other circumstances, particularly those involving wireless networks and radio transmissions, we need to use a range of frequencies that does not start at zero to send information across a channel. Specifically, for wireless channels, it is not practical to send very low frequency signals because the size of the antenna needs to be a fraction of the signal wavelength, which becomes large at high transmission frequencies. In any case, regulatory constraints and the need to avoid interference usually dictate the choice of frequencies. Even for wires, placing a signal in a given frequency band is useful to let different kinds of signals coexist on the channel. This kind of transmission is called passband transmission because an arbitrary band of frequencies is used to pass the signal.

Fortunately, our fundamental results from earlier in the chapter are all in terms of bandwidth, or the *width* of the frequency band. The absolute frequency values do not matter for capacity. This means that we can take a **baseband** signal that occupies 0 to B Hz and shift it up to occupy a passband of S to $S + B$ Hz without changing the amount of information that it can carry, even though the signal will look different. To process a signal at the receiver, we can shift it back down to baseband, where it is more convenient to detect symbols.

Digital modulation is accomplished with passband transmission by modulating a carrier signal that sits in the passband. We can modulate the amplitude, frequency, or phase of the carrier signal. Each of these methods has a corresponding name.

In **ASK** (**Amplitude Shift Keying**), two different amplitudes are used to represent 0 and 1. An example with a nonzero and a zero level is shown in Fig. 2-16(b). More than two levels can be used to encode multiple bits per symbol.

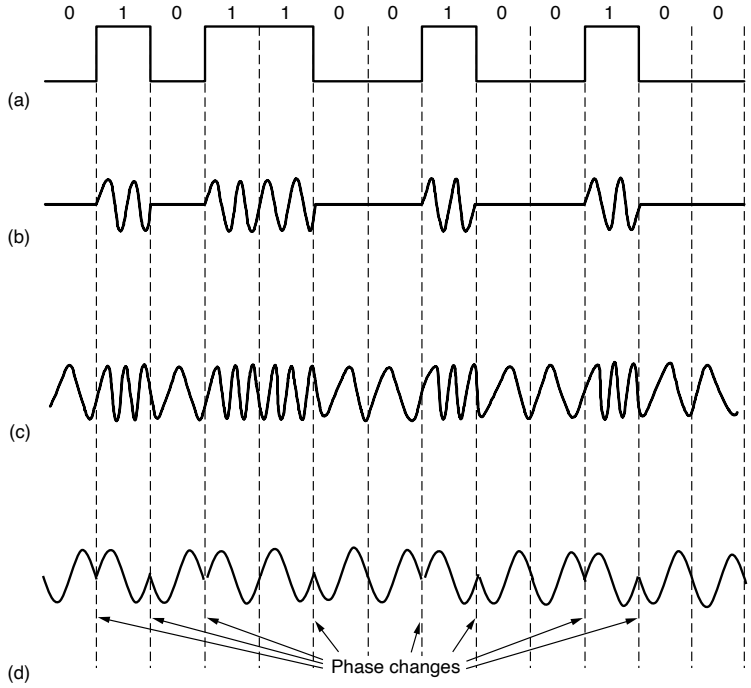


Figure 2-16. (a) A binary signal. (b) Amplitude shift keying. (c) Frequency shift keying. (d) Phase shift keying.

Similarly, with **FSK** (**Frequency Shift Keying**), two or more different tones are used. The example in Fig. 2-16(c) uses just two frequencies. In the simplest form of **PSK** (**Phase Shift Keying**), the carrier wave is systematically shifted 0 or 180 degrees at each symbol period. Because there are two phases, it is called **BPSK** (**Binary Phase Shift Keying**). ''Binary'' here refers to the two symbols, not that the symbols represent 2 bits. An example is shown in Fig. 2-16(d). A better scheme that uses the channel bandwidth more efficiently is to use four shifts, e.g., 45, 135, 225, or 315 degrees, to transmit 2 bits of information per symbol. This version is called **QPSK** (**Quadrature Phase Shift Keying**).

We can combine these schemes and use more levels to transmit more bits per symbol. Only one of frequency and phase can be modulated at a time because they

are related, with frequency being the rate of change of phase over time. Usually, amplitude and phase are modulated in combination. Three examples are shown in Fig. 2-17. In each example, the points give the legal amplitude and phase combinations of each symbol. In Fig. 2-17(a), we see equidistant dots at 45, 135, 225, and 315 degrees. The phase of a dot is indicated by the angle a line from it to the origin makes with the positive *x*-axis. The amplitude of a dot is the distance from the origin. This figure is a graphical representation of QPSK.

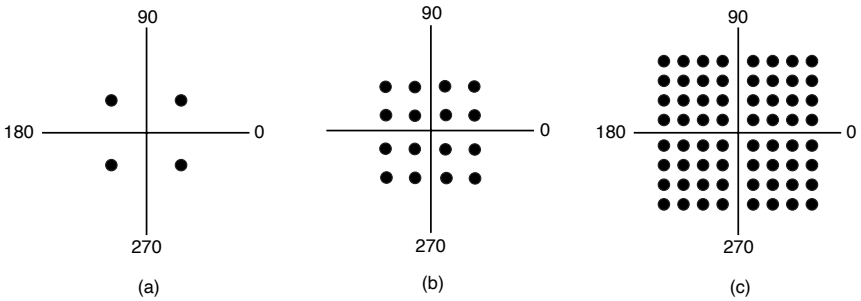
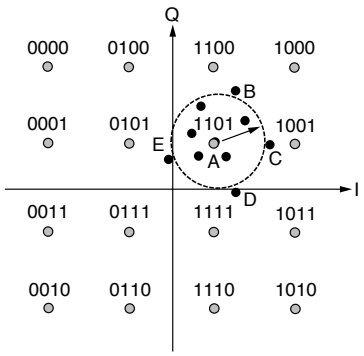


Figure 2-17. (a) QPSK. (b) QAM-16. (c) QAM-64.

This kind of diagram is called a **constellation diagram**. In Fig. 2-17(b) we see a modulation scheme with a denser constellation. Sixteen combinations of amplitudes and phase are used here, so the modulation scheme can be used to transmit 4 bits per symbol. It is called **QAM-16**, where QAM stands for **Quadrature Amplitude Modulation**. Figure 2-17(c) is a still denser modulation scheme with 64 different combinations, so 6 bits can be transmitted per symbol. It is called **QAM-64**. Even higher-order QAMs are used too. As you might suspect from these constellations, it is easier to build electronics to produce symbols as a combination of values on each axis than as a combination of amplitude and phase values. That is why the patterns look like squares rather than concentric circles.

The constellations we have seen so far do not show how bits are assigned to symbols. When making the assignment, an important consideration is that a small burst of noise at the receiver not lead to many bit errors. This might happen if we assigned consecutive bit values to adjacent symbols. With QAM-16, for example, if one symbol stood for 0111 and the neighboring symbol stood for 1000, if the receiver mistakenly picks the adjacent symbol, it will cause all of the bits to be wrong. A better solution is to map bits to symbols so that adjacent symbols differ in only 1 bit position. This mapping is called a **Gray code**. Figure 2-18 shows a QAM-16 constellation that has been Gray coded. Now if the receiver decodes the symbol in error, it will make only a single bit error in the expected case that the decoded symbol is close to the transmitted symbol.



When 1101 is sent:

Point	Decodes as	Bit errors
A	1101	0
B	1100	1
C	1001	1
D	1111	1
E	0101	1

Figure 2-18. Gray-coded QAM-16.

2.4.4 Multiplexing

The modulation schemes we have seen let us send one signal to convey bits along a wired or wireless link, but they only describe how to transmit one bitstream at a time. In practice, economies of scale play an important role in how we use networks: It costs essentially the same amount of money to install and maintain a high-bandwidth transmission line as a low-bandwidth line between two different offices (i.e., the costs come from having to dig the trench and not from what kind of cable or fiber goes into it). Consequently, multiplexing schemes have been developed to share lines among many signals. The three main ways to multiplex a single physical line are time, frequency, and code; there is also a technique called wavelength division multiplexing, which is essentially an optical form of frequency division multiplexing. We discuss each of these techniques below.

Frequency Division Multiplexing

FDM (**Frequency Division Multiplexing**) takes advantage of passband transmission to share a channel. It divides the spectrum into frequency bands, with each user having exclusive possession of some band in which to send a signal. AM radio broadcasting illustrates FDM. The allocated spectrum is about 1 MHz, roughly 500 to 1500 kHz. Different frequencies are allocated to different logical channels (stations), each operating in a portion of the spectrum, with the interchannel separation great enough to prevent interference.

For a more detailed example, in Fig. 2-19 we see three voice-grade telephone channels multiplexed using FDM. Filters limit the usable bandwidth to roughly 3100 Hz per voice-grade channel. When many channels are multiplexed together, 4000 Hz is allocated per channel. The excess bandwidth is called a **guard band**.

It keeps the channels well separated. First, the voice channels are raised in frequency, each by a different amount. Then they can be combined because no two channels now occupy the same portion of the spectrum. Notice that even though there are gaps between the channels thanks to the guard bands, there is some overlap between adjacent channels. The overlap is there because real filters do not have ideal sharp edges. This means that a strong spike at the edge of one channel will be felt in the adjacent one as nonthermal noise.

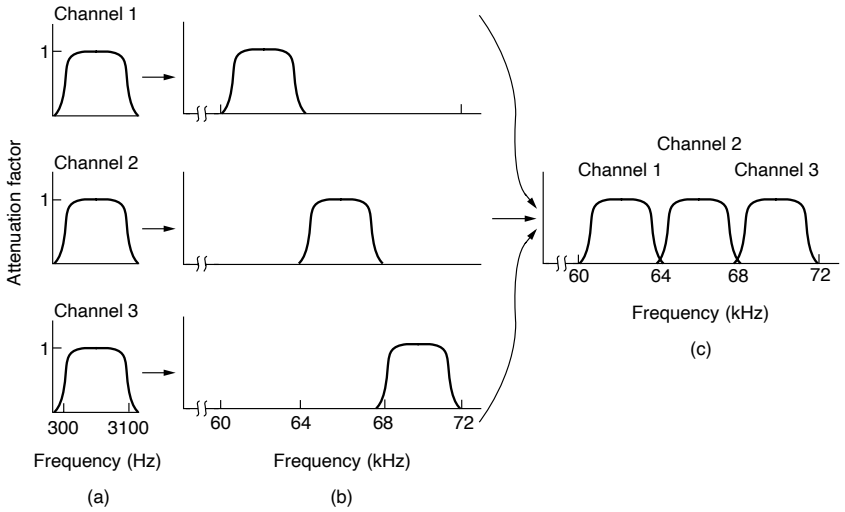


Figure 2-19. Frequency division multiplexing. (a) The original bandwidths. (b) The bandwidths raised in frequency. (c) The multiplexed channel.

This scheme has been used to multiplex calls in the telephone system for many years, but multiplexing in time is now preferred instead. However, FDM continues to be used in telephone networks, as well as cellular, terrestrial wireless, and satellite networks at a higher level of granularity.

When sending digital data, it is possible to divide the spectrum efficiently without using guard bands. In **OFDM** (**Orthogonal Frequency Division Multiplexing**), the channel bandwidth is divided into many subcarriers that independently send data (e.g., with QAM). The subcarriers are packed tightly together in the frequency domain. Thus, signals from each subcarrier extend into adjacent ones. However, as seen in Fig. 2-20, the frequency response of each subcarrier is designed so that it is zero at the center of the adjacent subcarriers. The subcarriers can therefore be sampled at their center frequencies without interference from their neighbors. To make this work, a **guard time** is needed to repeat a portion of the symbol signals in time so that they have the desired frequency response. However, this overhead is much less than is needed for many guard bands.

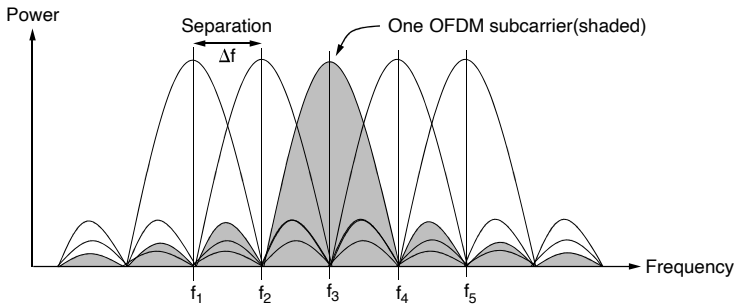


Figure 2-20. Orthogonal frequency division multiplexing (OFDM).

OFDM has been around for a long time, but it only began to be adopted in the early 2000s, following the realization that it is possible to implement OFDM efficiently in terms of a Fourier transform of digital data over all subcarriers (instead of separately modulating each subcarrier). OFDM is used in 802.11, cable networks, power-line networking, and fourth-generation (4G) cellular systems. Most often, one high-rate stream of digital information is split into a number of low-rate streams that are transmitted on the subcarriers in parallel. This division is valuable because degradations of the channel are easier to cope with at the subcarrier level; some subcarriers may be very degraded and excluded in favor of subcarriers that are received well.

Time Division Multiplexing

An alternative to FDM is **TDM** (**Time Division Multiplexing**). Here, the users take turns (in a round-robin fashion), each one periodically getting the entire bandwidth for a certain time interval. An example of three streams being multiplexed with TDM is shown in Fig. 2-21. Bits from each input stream are taken in a fixed **time slot** and output to the aggregate stream. This stream runs at the sum rate of the individual streams. For this to work, the streams must be synchronized in time. Small intervals of guard time (analogous to a frequency guard band) may be added to accommodate small timing variations.

TDM is used widely as key technique in the telephone and cellular networks. To avoid one point of confusion, let us be clear that it is quite different from the alternative **STDM** (**Statistical Time Division Multiplexing**). The prefix ''statistical'' is added to indicate that the individual streams contribute to the multiplexed stream *not* on a fixed schedule, but according to the statistics of their demand. STDM is fundamentally like packet switching under another name.

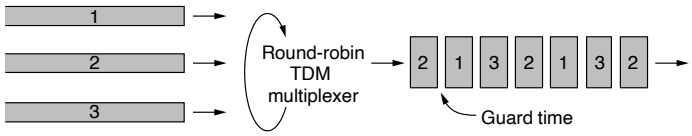


Figure 2-21. Time Division Multiplexing (TDM).

Code Division Multiplexing

There is a third kind of multiplexing that works in a completely different way than FDM and TDM. **CDM** (**Code Division Multiplexing**) is a form of **spread spectrum** communication in which a narrowband signal is spread out over a wider frequency band. This can make it more tolerant of interference, as well as allowing multiple signals from different users to share the same frequency band. Because code division multiplexing is mostly used for the latter purpose it is commonly called **CDMA** (**Code Division Multiple Access**).

CDMA allows each station to transmit over the entire frequency spectrum all the time. Multiple simultaneous transmissions are separated using coding theory. Before getting into the algorithm, let us consider an analogy: an airport lounge with many pairs of people conversing. TDM is comparable to pairs of people in the room taking turns speaking. FDM is comparable to the pairs of people speaking at different pitches, some high-pitched and some low-pitched such that each pair can hold its own conversation at the same time as but independently of the others. CDMA is somewhat comparable to each pair of people talking at once, but in a different language. The French-speaking couple just hones in on the French, rejecting everything that is not French as noise. Thus, the key to CDMA is to be able to extract the desired signal while rejecting everything else as random noise. A somewhat simplified description of CDMA follows.

In CDMA, each bit time is subdivided into m short intervals called **chips**, which are multiplied against the original data sequence (the chips are a bit sequence, but are called chips so that the are not confused with the bits of the actual message). Typically, there are 64 or 128 chips per bit, but in the example given here we will use 8 chips/bit for simplicity. Each station is assigned a unique m-bit code called a **chip sequence**. For pedagogical purposes, it is convenient to write these codes as sequences of −1 and +1. We will show chip sequences in parentheses.

To transmit a 1 bit, a station sends its chip sequence. To transmit a 0 bit, it sends the negation of its chip sequence. No other patterns are permitted. Thus, for $m = 8$, if station A is assigned the chip sequence (−1 − 1 − 1 + 1 + 1 − 1 + 1 + 1), it can send a 1 bit by transmitting the chip sequence and a 0 by transmitting its complement: (+1 + 1 + 1 − 1 − 1 + 1 − 1 − 1). It is really voltage levels that are sent, but it is sufficient for us to think in terms of the sequences.

Increasing the amount of information to be sent from b bits/sec to mb chips/sec for each station means that the bandwidth needed for CDMA is greater by a factor of m than the bandwidth needed for a station not using CDMA (assuming no changes in the modulation or encoding techniques). If we have a 1-MHz band available for 100 stations, with FDM each one would have 10 kHz and could send at 10 kbps (assuming 1 bit per Hz). With CDMA, each station uses the full 1 MHz, so the chip rate is 100 chips per bit to spread the station's bit rate of 10 kbps across the channel.

In Fig. 2-22(a) and (b), we show the chip sequences assigned to four example stations and the signals that they represent. Each station has its own unique chip sequence. Let us use the symbol **S** to indicate the m-chip vector for station S, and $\bar{\mathbf{S}}$ for its negation. All chip sequences are pairwise **orthogonal**, by which we mean that the normalized inner product of any two distinct chip sequences, **S** and **T** (written as **S•T**), is 0. It is known how to generate such orthogonal chip sequences using a method known as **Walsh codes**. In mathematical terms, orthogonality of the chip sequences can be expressed as follows:

$$\mathbf{S}\bullet\mathbf{T} \equiv \frac{1}{m}\sum_{i=1}^{m} S_i T_i = 0 \tag{2-5}$$

In plain English, as many pairs are the same as are different. This orthogonality property will prove crucial later. Note that if **S•T** = 0, then **S•**$\bar{\mathbf{T}}$ is also 0. The normalized inner product of any chip sequence with itself is 1:

$$\mathbf{S}\bullet\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m} S_i S_i = \frac{1}{m}\sum_{i=1}^{m} S_i^2 = \frac{1}{m}\sum_{i=1}^{m}(\pm 1)^2 = 1$$

0.20v This follows because each of the m terms in the inner product is 1, so the sum is m. Also, note that **S•**$\bar{\mathbf{S}}$ = – 1.

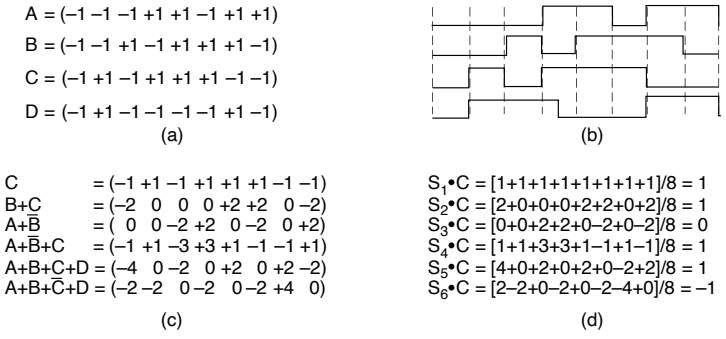


Figure 2-22. (a) Chip sequences for four stations. (b) Signals the sequences represent (c) Six examples of transmissions. (d) Recovery of station C's signal.

During each bit time, a station can transmit a 1 (by sending its chip sequence), it can transmit a 0 (by sending the negative of its chip sequence), or it can be silent and transmit nothing. We assume for now that all stations are synchronized in time, so all chip sequences begin at the same instant. When two or more stations transmit simultaneously, their bipolar sequences add linearly. For example, if in one chip period three stations output +1 and one station outputs −1, +2 will be received. One can think of this as signals that add as voltages superimposed on the channel: three stations output +1 V and one station outputs −1 V, so that 2 V is received. For instance, in Fig. 2-22(c) we see six examples of one or more stations transmitting 1 bit at the same time. In the first example, *C* transmits a 1 bit, so we just get *C*'s chip sequence. In the second example, both *B* and *C* transmit 1 bits, so we get the sum of their bipolar chip sequences, namely:

$$(-1\ -1\ +1\ -1\ +1\ +1\ +1\ -1) + (-1\ +1\ -1\ +1\ +1\ +1\ -1\ -1) = (-2\ \ 0\ \ 0\ \ 0 + 2 + 2\ \ 0 - 2)$$

To recover the bit stream of an individual station, the receiver must know that station's chip sequence in advance. It does the recovery by computing the normalized inner product of the received chip sequence and the chip sequence of the station whose bit stream it is trying to recover. If the received chip sequence is **S** and the receiver is trying to listen to a station whose chip sequence is **C**, it just computes the normalized inner product, **S•C**.

To see why this works, just imagine that two stations, *A* and *C*, both transmit a 1 bit at the same time that *B* transmits a 0 bit, as in the third example. The receiver sees the sum, $\mathbf{S} = \mathbf{A} + \bar{\mathbf{B}} + \mathbf{C}$, and computes

$$\mathbf{S}\bullet\mathbf{C} = (\mathbf{A} + \bar{\mathbf{B}} + \mathbf{C})\bullet\mathbf{C} = \mathbf{A}\bullet\mathbf{C} + \bar{\mathbf{B}}\bullet\mathbf{C} + \mathbf{C}\bullet\mathbf{C} = 0 + 0 + 1 = 1$$

The first two terms vanish because all pairs of chip sequences have been carefully chosen to be orthogonal, as shown in Eq. (2-5). Now it should be clear why this property must be imposed on the chip sequences.

To make the decoding process more concrete, we show six examples in Fig. 2-22(d). Suppose that the receiver is interested in extracting the bit sent by station *C* from each of the six signals S_1 through S_6. It calculates the bit by summing the pairwise products of the received **S** and the **C** vector of Fig. 2-22(a) and then taking 1/8 of the result (since $m = 8$ here). The examples include cases where *C* is silent, sends a 1 bit, and sends a 0 bit, individually and in combination with other transmissions. As shown, the correct bit is decoded each time. It is just like speaking French.

In principle, given enough computing capacity, the receiver can listen to all the senders at once by running the decoding algorithm for each of them in parallel. In real life, suffice it to say that this is easier said than done, and it is useful to know which senders might be transmitting.

In the ideal, noiseless CDMA system we have studied here, the number of stations that send concurrently can be made arbitrarily large by using longer chip sequences. For 2^n stations, Walsh codes can provide 2^n orthogonal chip sequences

of length 2^n. However, one significant limitation is that we have assumed that all the chips are synchronized in time at the receiver. This synchronization is not even approximately true in some applications, such as cellular networks (in which CDMA has been widely deployed starting in the 1990s). It leads to different designs.

As well as cellular networks, CDMA is used by satellites and cable networks. We have glossed over many complicating factors in this brief introduction. Engineers who want to gain a deep understanding of CDMA should read Viterbi (1995) and Harte et al. (2012). These references require quite a bit of background in communication engineering, however.

Wavelength Division Multiplexing

WDM (**Wavelength Division Multiplexing**) is a form of frequency division multiplexing that multiplexes multiple signals onto an optical fiber using different wavelengths of light. In Fig. 2-23, four fibers come together at an optical combiner, each with its energy present at a different wavelength. The four beams are combined onto a single shared fiber for transmission to a distant destination. At the far end, the beam is split up over as many fibers as there were on the input side. Each output fiber contains a short, specially constructed core that filters out all but one wavelength. The resulting signals can be routed to their destination or recombined in different ways for additional multiplexed transport.

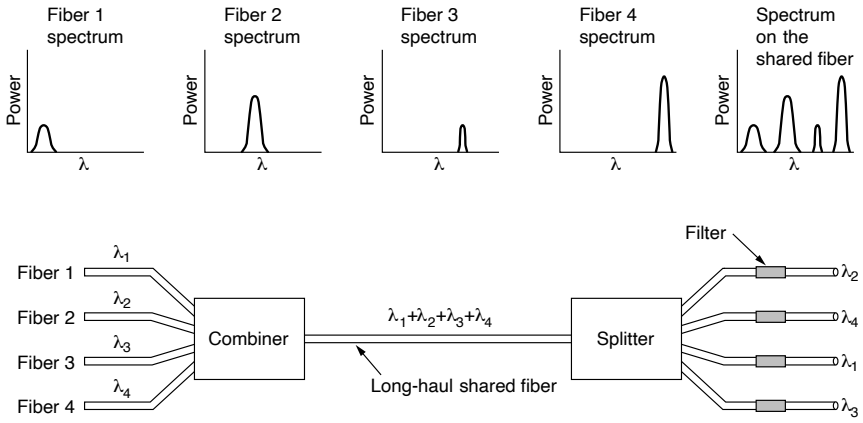


Figure 2-23. Wavelength division multiplexing.

There is really nothing new here. This way of operating is just frequency division multiplexing at very high frequencies, with the term WDM referring to the

description of fiber optic channels by their wavelength or ''color'' rather than frequency. As long as each channel has its own dedicated frequency (that is, its own wavelength) range and all the ranges are disjoint, they can be multiplexed together on the long-haul fiber. The only difference with electrical FDM is that an optical system using a diffraction grating is completely passive and thus highly reliable.

The reason WDM is popular is that the energy on a single channel is typically only a few gigahertz wide because that is the current limit of how fast we can convert between electrical and optical signals. By running many channels in parallel on different wavelengths, the aggregate bandwidth is increased linearly with the number of channels. Since the bandwidth of a single fiber band is ca. 25,000 GHz (see Fig. 2-5), there is theoretically room for 2500 10-Gbps channels even at 1 bit/Hz (and higher rates are also possible).

WDM technology has been progressing at a rate that puts computer technology to shame. WDM was invented around 1990. The first commercially available systems had eight channels of 2.5 Gbps per channel; by 1998, systems with 40 channels of 2.5 Gbps were on the market and rapidly being adopted; by 2006, there were products with 192 channels of 10 Gbps and 64 channels of 40 Gbps, capable of moving up to 2.56 Tbps; by 2019, there were systems that can handle up to 160 channels, supporting more than 16 Tbps over a single fiber pair. That is 800 times more capacity than the 1990 systems. The channels are also packed tightly on the fiber, with 200, 100, or as little as 50 GHz of separation.

Narrowing the spacing to 12.5 GHz makes it possible to support 320 channels on a single fiber, further increasing transmission capacity. Such systems with a large number of channels and little space between each channel are referred to as **DWDM** (**Dense WDM**). DWDM systems tend to be more expensive because they must maintain stable wavelengths and frequencies, due to the close spacing of each channel. As a result, these systems closely regulate their temperature to ensure that frequencies are accurate.

One of the drivers of WDM technology is the development of all-optical components. Previously, every 100 km it was necessary to split up all the channels and convert each one to an electrical signal for amplification separately before reconverting them to optical signals and combining them. Nowadays, all-optical amplifiers can regenerate the entire signal once every 1000 km without the need for multiple opto-electrical conversions.

In the example of Fig. 2-23, we have a fixed-wavelength system. Bits from input fiber 1 go to output fiber 3, bits from input fiber 2 go to output fiber 1, etc. However, it is also possible to build WDM systems that are switched in the optical domain. In such a device, the output filters are tunable using Fabry-Perot or Mach-Zehnder interferometers. These devices allow the selected frequencies to be changed dynamically by a control computer. This ability provides a large amount of flexibility to provision many different wavelength paths through the telephone network from a fixed set of fibers. For more information about optical networks and WDM, see Grobe and Eiselt (2013).

2.5 THE PUBLIC SWITCHED TELEPHONE NETWORK

When two computers that are physically close to each other need to communicate, it is often easiest just to run a cable between them. Local Area Networks (LANs) work this way. However, when the distances are large or there are many computers or the cables have to pass through a public road or other public right of way, the costs of running private cables are usually prohibitive. Furthermore, in just about every country in the world, stringing private transmission lines across (or underneath) public property is illegal. Consequently, the network designers must rely on the existing telecommunication facilities, such as the telephone network, the cellular network, or the cable television network.

The limiting factor for data networking has long been the ''last mile'' over which customers connect, which might rely on any one of these physical technologies, as opposed to the so-called ''backbone'' infrastructure for the rest of the access network. Over the past decade, this situation has changed dramatically, with speeds of 1 Gbps to the home becoming increasingly commonplace. Although one contributor to faster last-mile speeds is the continued rollout of fiber at the edge of the network, perhaps an even more significant contributor in some countries is the sophisticated engineering of the *existing* telephone and cable networks to squeeze increasingly more bandwidth out of the existing infrastructure. It turns out that engineering the existing physical infrastructure to increase transmission speeds is a lot less expensive than putting new (fiber) cables in the ground to everyone's homes. We now explore the architectures and characteristics of each of these physical communications infrastructures.

These existing facilities, especially the **PSTN** (**Public Switched Telephone Network**), were usually designed many years ago, with a completely different goal in mind: transmitting the human voice in a more-or-less recognizable form. A cable running between two computers can transfer data at 10 Gbps or more; the phone network thus has its work cut out for it in terms of transmitting bits at high rates. Early Digital Subscriber Line (DSL) technologies could only transmit data at rates of a few Mbps; now, more modern versions of DSL, can achieve rates approaching 1 Gbps. In the following sections, we will describe the telephone system and show how it works. For additional information about the innards of the telephone system, see Laino (2017).

2.5.1 Structure of the Telephone System

Soon after Alexander Graham Bell patented the telephone in 1876 (just a few hours ahead of his rival, Elisha Gray), there was an enormous demand for his new invention. The initial market was for the sale of telephones, which came in pairs. It was up to the customer to string a single wire between them. If a telephone owner wanted to talk to n other telephone owners, separate wires had to be strung to all n houses. Within a year, the cities were covered with wires passing over

houses and trees in a wild jumble. It became immediately obvious that the model of connecting every telephone to every other telephone, as shown in Fig. 2-24(a), was not going to work.

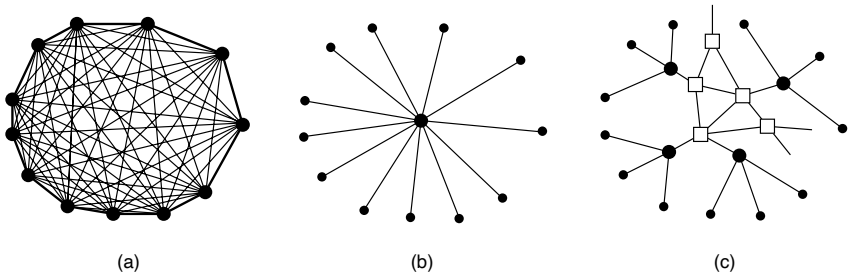


Figure 2-24. (a) Fully interconnected network. (b) Centralized switch. (c) Two-level hierarchy.

To his credit, Bell saw this problem early on and formed the Bell Telephone Company, which opened its first switching office (in New Haven, Connecticut) in 1878. The company ran a wire to each customer's house or office. To make a call, the customer would crank the phone to make a ringing sound in the telephone company office to attract the attention of an operator, who would then manually connect the caller to the callee by using a short jumper cable. The model of a single switching office is illustrated in Fig. 2-24(b).

Pretty soon, Bell System switching offices were springing up everywhere and people wanted to make long-distance calls between cities, so the Bell System began to connect the switching offices. The original problem soon returned: to connect every switching office to every other switching office by means of a wire between them quickly became unmanageable, so second-level switching offices were invented. After a while, multiple second-level offices were needed, as illustrated in Fig. 2-24(c). Eventually, the hierarchy grew to five levels.

By 1890, the three major parts of the telephone system were in place: the switching offices, the wires between the customers and the switching offices (by now balanced, insulated, twisted pairs instead of open wires with an earth return), and the long-distance connections between the switching offices. For a short technical history of the telephone system, see Hawley (1991).

While there have been improvements in all three areas since then, the basic Bell System model has remained essentially intact for over 100 years. The following description is highly simplified but gives the essential flavor nevertheless. Each telephone has two copper wires coming out of it that go directly to the telephone company's nearest **end office** (also called a **local central office**). The distance is typically around 1 to 10 km, being shorter in cities than in rural areas. In

the United States alone there are about 22,000 end offices. The two-wire connections between each subscriber's telephone and the end office are known in the trade as the **local loop**. If the world's local loops were stretched out end to end, they would extend to the moon and back 1000 times.

At one time, 80% of AT&T's capital value was the copper in the local loops. AT&T was then, in effect, the world's largest copper mine. Fortunately, this fact was not well known in the investment community. Had it been known, some corporate raider might have bought AT&T, ended all telephone service in the United States, ripped out all the wire, and sold it to a copper refiner for a quick payback.

If a subscriber attached to a given end office calls another subscriber attached to the same end office, the switching mechanism within the office sets up a direct electrical connection between the two local loops. This connection remains intact for the duration of the call.

If the called telephone is attached to another end office, a different procedure has to be used. Each end office has a number of outgoing lines to one or more nearby switching centers, called **toll offices** (or, if they are within the same local area, **tandem offices**). These lines are called **toll connecting trunks**. The number of different kinds of switching centers and their topology varies from country to country depending on the country's telephone density.

If both the caller's and callee's end offices happen to have a toll connecting trunk to the same toll office (a likely occurrence if they are relatively close by), the connection may be established within the toll office. A telephone network consisting only of telephones (the small dots), end offices (the large dots), and toll offices (the squares) is shown in Fig. 2-24(c).

If the caller and callee do not have a toll office in common, a path will have to be established between two toll offices. The toll offices communicate with each other via high-bandwidth **intertoll trunks** (also called **interoffice trunks**). Prior to the 1984 breakup of AT&T, the U.S. telephone system used hierarchical routing to find a path, going to higher levels of the hierarchy until there was a switching office in common. This was then replaced with more flexible, non-hierarchical routing. Figure 2-25 shows how a long-distance connection might be routed.

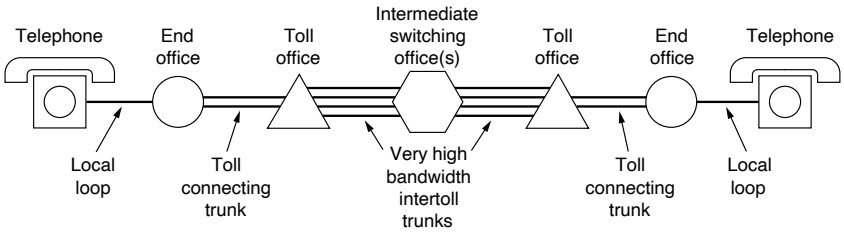


Figure 2-25. A typical circuit route for a long-distance call.

A variety of transmission media are used for telecommunication. Unlike modern office buildings, where the wiring is commonly Category 5 or Category 6, local loops to homes mostly consist of Category 3 twisted pairs, although some local loops are now fiber, as well. Coaxial cables, microwaves, and especially fiber optics are widely used between switching offices.

In the past, transmission throughout the telephone system was analog, with the actual voice signal being transmitted as an electrical voltage from source to destination. With the advent of fiber optics, digital electronics, and computers, all the trunks and switches are now digital, leaving the local loop as the last piece of analog technology in the system. Digital transmission is preferred because it is not necessary to accurately reproduce an analog waveform after it has passed through many amplifiers on a long call. Being able to correctly distinguish a 0 from a 1 is enough. This property makes digital transmission more reliable than analog. It is also cheaper and easier to maintain.

In summary, the telephone system consists of three major components:

1. Local loops (analog twisted pairs between end offices and local houses and businesses).
2. Trunks (very high-bandwidth digital fiber-optic links connecting the switching offices).
3. Switching offices (where calls are moved from one trunk to another either electrically or optically).

The local loops provide everyone access to the whole system, so they are critical. Unfortunately, they are also the weakest link in the system. The main challenge for long-haul trunks involves collecting multiple calls and sending them out over the same fiber, which is done using wavelength division multiplexing (WDM). Finally, there are two fundamentally different ways of doing switching: circuit switching and packet switching. We will look at both.

2.5.2 The Local Loop: Telephone Modems, ADSL, and Fiber

In this section, we will study the local loop, both old and new. We will cover telephone modems, ADSL, and fiber to the home. In some places, the local loop has been modernized by installing optical fiber to (or at least very close to) the home. These installations support computer networks from the ground up, with the local loop having ample bandwidth for data services. Unfortunately, the cost of laying fiber to homes is substantial. Sometimes, it is done when local city streets are dug up for other purposes; some municipalities, especially in densely populated urban areas, have fiber local loops. By and large, however, fiber local loops are the exception, but they are clearly the future.

Telephone Modems

Most people are familiar with the two-wire local loop coming from a telephone company end office into houses. The local loop is also frequently referred to as the ''last mile,'' although the length can be up to several miles. Much effort has been devoted to squeezing data networking out of the copper local loops that are already deployed. Telephone modems send digital data between computers over the narrow channel the telephone network provides for a voice call. They were once widely used, but have been largely displaced by broadband technologies such as ADSL that reuse the local loop to send digital data from a customer to the end office, where they are siphoned off to the Internet. Both modems and ADSL must deal with the limitations of old local loops: relatively narrow bandwidth, attenuation and distortion of signals, and susceptibility to electrical noise such as crosstalk.

To send bits over the local loop, or any other physical channel for that matter, they must be converted to analog signals that can be transmitted over the channel. This conversion is accomplished using the methods for digital modulation that we studied in the previous section. At the other end of the channel, the analog signal is converted back to bits.

A device that converts between a stream of digital bits and an analog signal that represents the bits is called a **modem**, which is short for ''*mo*dulator *dem*odulator.'' Modems come in many varieties, including telephone modems, DSL modems, cable modems, and wireless modems. In the case of a cable or DSL modem, the device is typically a separate piece of hardware that sits in between the physical line coming into the house and the rest of the network inside the home. Wireless devices typically have their own built-in modems. Logically, the modem is inserted between the (digital) computer and the (analog) telephone system, as seen in Fig. 2-26.

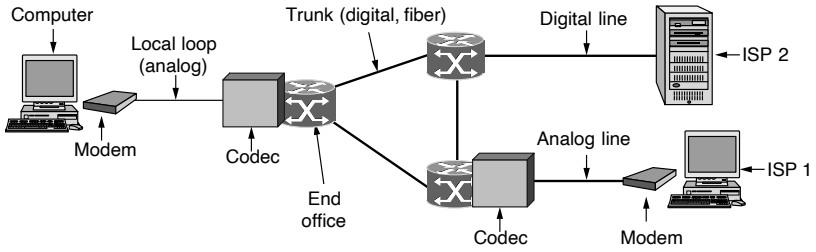


Figure 2-26. The use of both analog and digital transmission for a computer-to-computer call. Conversion is done by the modems and codecs.

Telephone modems are used to send bits between two computers over a voice-grade telephone line, in place of the conversation that usually fills the line. The

main difficulty in doing so is that a voice-grade telephone line is limited to only 3100 Hz, about what is sufficient to carry a conversation. This bandwidth is more than four orders of magnitude less than the bandwidth that is used for Ethernet or 802.11 (WiFi). Unsurprisingly, the data rates of telephone modems are also four orders of magnitude less than that of Ethernet and 802.11.

Let us run the numbers to see why this is the case. The Nyquist theorem tells us that even with a perfect 3000-Hz line (which a telephone line is decidedly not), there is no point in sending symbols at a rate faster than 6000 baud. Let us consider, for example, an older modem sending at a rate of 2400 symbols/sec, (2400 baud) and focus on getting multiple bits per symbol while allowing traffic in both directions at the same time (by using different frequencies for different directions).

The humble 2400-bps modem uses 0 volts for a logical 0 and 1 volt for a logical 1, with 1 bit per symbol. One step up, it can use four different symbols, as in the four phases of QPSK, so with 2 bits/symbol it can get a data rate of 4800 bps.

A long progression of higher rates has been achieved as technology has improved. Higher rates require a larger set of symbols (see Fig. 2-17). With many symbols, even a small amount of noise in the detected amplitude or phase can result in an error. To reduce the chance of errors, standards for the higher-speed modems use some of the symbols for error correction. The schemes are known as **TCM** (**Trellis Coded Modulation**). Some common modem standards are shown in Fig. 2-27.

Modem standard	Baud	Bits/symbol	Bps
V.32	2400	4	9600
V.32 bis	2400	6	14,400
V.34	2400	12	28,800
V.34 bis	2400	14	33,600

Figure 2-27. Some modem standards and their bit rate.

Why does it stop at 33,600 bps? The reason is that the Shannon limit for the telephone system is about 35 kbps based on the average length and quality of local loops. Going faster than this would violate the laws of physics (department of thermodynamics) or require new local loops (which is gradually being done).

However, there is one way we can change the situation. At the telephone company end office, the data are converted to digital form for transmission within the telephone network (the core of the telephone network converted from analog to digital long ago). The 35-kbps limit is for the situation in which there are two local loops, one at each end. Each of these adds noise to the signal. If we could get rid of one of these local loops, we would increase the SNR and the maximum rate would be doubled.

This approach is how 56-kbps modems are made to work. One end, typically an ISP (Internet Service Provider), gets a high-quality digital feed from the nearest

end office. Thus, when one end of the connection is a high-quality signal, as it is with most ISPs now, the maximum data rate can be as high as 70 kbps. Between two home users with modems and analog lines, the maximum is still 33.6 kbps.

The reason that 56-kbps modems (rather than 70-kbps modems) are in use has to do with the Nyquist theorem. A telephone channel is carried inside the telephone system as digital samples. Each telephone channel is 4000 Hz wide when the guard bands are included. The number of samples per second needed to reconstruct it is thus 8000. The number of bits per sample in North America is 8, of which one is used for control purposes, allowing 56,000 bits/sec of user data. In Europe, all 8 bits are available to users, so 64,000-bit/sec modems could have been used, but to get international agreement on a standard, 56,000 was chosen.

The end result is the **V.90** and **V.92** modem standards. They provide for a 56-kbps downstream channel (ISP to user) and a 33.6-kbps and 48-kbps upstream channel (user to ISP), respectively. The asymmetry is because there is usually more data transported from the ISP to the user than the other way. It also means that more of the limited bandwidth can be allocated to the downstream channel to increase the chances of it actually working at 56 kbps.

Digital Subscriber Lines (DSL)

When the telephone industry finally got to 56 kbps, it patted itself on the back for a job well done. Meanwhile, the cable TV industry was offering speeds up to 10 Mbps on shared cables. As Internet access became an increasingly important part of their business, the local telephone companies began to realize they needed a more competitive product. Their answer was to offer new digital services over the local loop.

Initially, there were many overlapping high-speed offerings, all under the general name of **xDSL** (**Digital Subscriber Line**), for various *x*. Services with more bandwidth than standard telephone service are sometimes referred to as **broadband**, although the term really is more of a marketing concept than a specific technical concept. Later, we will discuss what has become the most popular of these services, **ADSL** (**Asymmetric DSL**). We will also use the term DSL or xDSL as shorthand for all flavors.

The reason that modems are so slow is that telephones were invented for carrying the human voice, and the entire system has been carefully optimized for this purpose. Data have always been stepchildren. At the point where each local loop terminates in the end office, the wire runs through a filter that attenuates all frequencies below 300 Hz and above 3400 Hz. The cutoff is not sharp—300 Hz and 3400 Hz are the 3-dB points—so the bandwidth is usually quoted as 4000 Hz even though the distance between the 3 dB points is 3100 Hz. Data on the wire are thus also restricted to this narrow band.

The trick that makes xDSL work is that when a customer subscribes to it, the incoming line is connected to a different kind of switch that does not have this

filter, thus making the entire capacity of the local loop available. The limiting factor then becomes the physics of the local loop, which supports roughly 1 MHz, not the artificial 3100 Hz bandwidth created by the filter.

Unfortunately, the capacity of the local loop falls rather quickly with distance from the end office as the signal is increasingly degraded along the wire. It also depends on the thickness and general quality of the twisted pair. A plot of the potential bandwidth as a function of distance is given in Fig. 2-28. This figure assumes that all the other factors are optimal (new wires, modest bundles, etc.).

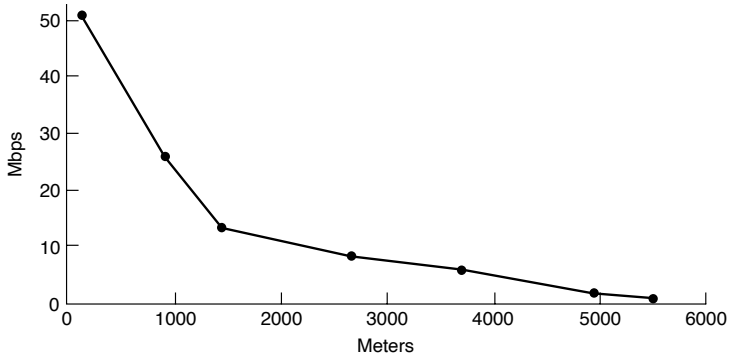


Figure 2-28. Bandwidth versus distance over Category 3 UTP for DSL.

The implication of this figure creates a problem for the telephone company. When it picks a speed to offer, it is simultaneously picking a radius from its end offices beyond which the service cannot be offered. This means that when distant customers try to sign up for the service, they may be told ''Thanks a lot for your interest, but you live 100 meters too far from the nearest end office to get this service. Could you please move?'' The lower the chosen speed is, the larger the radius and the more customers are covered. But the lower the speed, the less attractive the service is and the fewer the people who will be willing to pay for it. This is where business meets technology.

The xDSL services have all been designed with certain goals in mind. First, the services must work over the existing Category 3 twisted-pair local loops. Second, they must not affect customers' existing telephones and fax machines. Third, they must be much faster than 56 kbps. Fourth, they should be always on, with just a monthly charge and no per-minute charge.

To meet the technical goals, the available 1.1-MHz spectrum on the local loop is divided into 256 independent channels of 4312.5 Hz each. This arrangement is shown in Fig. 2-29. The OFDM scheme, which we saw in the previous section, is used to send data over these channels, though it is often called **DMT** (**Discrete MultiTone**) in the context of ADSL. Channel 0 is used for **POTS** (**Plain Old**

Telephone Service). Channels 1–5 are not used, to keep the voice and data signals from interfering with each other. Of the remaining 250 channels, one is used for upstream control and one is used for downstream control. The rest are available for user data.

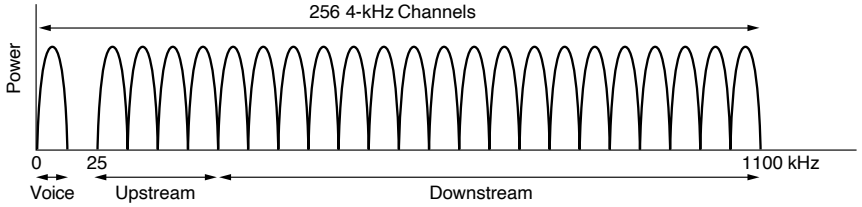


Figure 2-29. Operation of ADSL using discrete multitone modulation.

In principle, each of the remaining channels can be used for a full-duplex data stream, but harmonics, crosstalk, and other effects keep practical systems well below the theoretical limit. It is up to the provider to determine how many channels are available for upstream and how many for downstream. A 50/50 mix of upstream and downstream is technically possible, but most providers allocate something like 80–90% of the bandwidth to the downstream channel since most users download more data than they upload. This choice gives rise to the "A" in ADSL. A common split is 32 channels for upstream and the rest downstream. It is also possible to have a few of the highest upstream channels be bidirectional for increased bandwidth, although making this optimization requires adding a special circuit to cancel echoes.

The international ADSL standard, known as **G.dmt**, was approved in 1999. It allows speeds of as much as 8 Mbps downstream and 1 Mbps upstream. It was superseded by a second generation in 2002, called ADSL2, with various improvements to allow speeds of as much as 12 Mbps downstream and 1 Mbps upstream. ADSL2+ doubles the downstream throughput to 24 Mbps by doubling the bandwidth to use 2.2 MHz over the twisted pair.

The next improvement (in 2006) was **VDSL**, which pushed the data rate over the shorter local loops to 52 Mbps downstream and 3 Mbps upstream. Then, a series of new standards from 2007 to 2011, going under the name of **VDSL2**, on high-quality local loops managed to use 12-MHz bandwidth and achieve data rates of 200 Mbps downstream and 100 Mbps upstream. In 2015, **Vplus** was proposed for local loops shorter than 250 m. In principle, it can achieve 300 Mbps downstream and 100 Mbps upstream, but making it work in practice is not easy. We may be near the end of the line here for existing Category 3 wiring, except maybe for even shorter distances.

Within each channel, QAM modulation is used at a rate of roughly 4000 symbols/sec. The line quality in each channel is constantly monitored and the data rate

is adjusted by using a larger or smaller constellation, like those in Fig. 2-17. Different channels may have different data rates, with up to 15 bits per symbol sent on a channel with a high SNR, and down to 2, 1, or no bits per symbol sent on a channel with a low SNR depending on the standard.

A typical ADSL arrangement is shown in Fig. 2-30. In this scheme, a telephone company technician must install a **NID** (**Network Interface Device**) on the customer's premises. This small plastic box marks the end of the telephone company's property and the start of the customer's property. Close to the NID (or sometimes combined with it) is a **splitter**, an analog filter that separates the 0–4000-Hz band used by POTS from the data. The POTS signal is routed to the existing telephone or fax machine. The data signal is routed to an ADSL modem, which uses digital signal processing to implement OFDM. Since most ADSL modems are external, the computer must be connected to them at high speed. Usually, this is done using Ethernet, a USB cable, or 802.11.

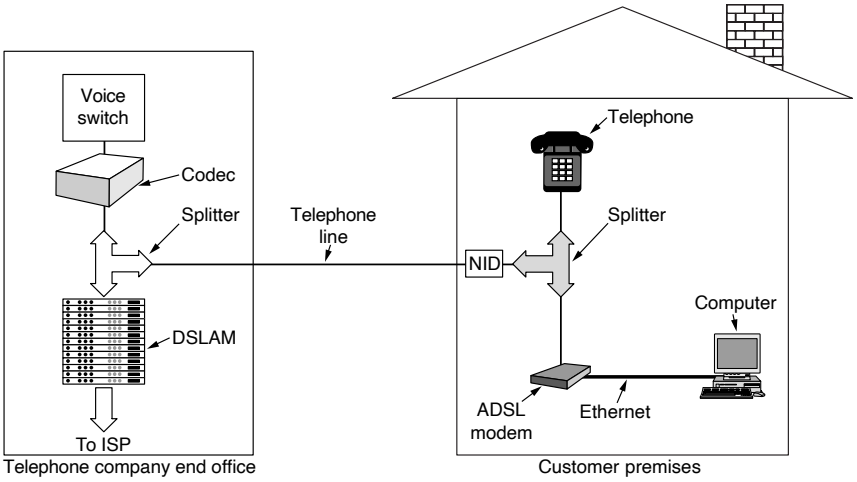


Figure 2-30. A typical ADSL equipment configuration.

At the other end of the wire, on the end office side, a corresponding splitter is installed. Here, the voice portion of the signal is filtered out and sent to the normal voice switch. The signal above 26 kHz is routed to a new kind of device called a **DSLAM** (**Digital Subscriber Line Access Multiplexer**), which contains the same kind of digital signal processor as the ADSL modem. The DSLAM converts the signal to bits and sends packets to the Internet service provider's data network.

This complete separation between the voice system and ADSL makes it relatively easy for a telephone company to deploy ADSL. All that is needed is buying a DSLAM and splitter and attaching the ADSL subscribers to the splitter.

Other high-bandwidth services delivered over the telephone network (e.g., ISDN) require the telephone company to make much greater changes to the existing switching equipment.

The next frontier for DSL deployments is to reach transmission speeds of 1 Gbps and higher. These efforts are focusing on a variety of complementary techniques, including a technique called **bonding**, which creates a single virtual DSL connection by combining two or more physical DSL connections. Obviously, if one combines two twisted pairs, one should be able to double the bandwidth. In some places, the telephone wires entering houses use a cable that in fact has two twisted pairs. The original idea was to allow two separate telephone lines and numbers in the house, but by using pair bonding, a single higher-speed Internet connection can be achieved. Increasing numbers of ISPs in Europe, Australia, Canada, and the United States are already deploying a technology called **G.fast** that uses pair bonding. As with other forms of DSL, the performance of G.fast depends on the distance of the transmission; recent tests have seen symmetric speeds approaching 1 Gbps at distances of 100 meters. When coupled with a fiber deployment known as **FTTdp** (**Fiber to the Distribution Point**), which brings fiber to a distribution point of several hundred subscribers and uses copper to transmit data the rest of the way to the home (in VDSL2, this may be up to 1 kilometer, although at lower speeds). FTTdp is just one type of fiber deployment that takes fiber from the core of the network to some point close to the network edge. The next section describes various modes of fiber deployment.

Fiber To The X (FTTX)

The speed of last-mile networks is often constrained by the copper cables used in conventional telephone networks, which cannot transmit data at high rates over as long a distance as fiber. Thus, an ultimate goal, where it is cost effective, is to bring fiber all the way to a customer home, sometimes called **FTTH** (**Fiber to the Home**). Telephone companies continue to try to improve the performance of the local loop, often by deploying fiber as far as they can to the home. If not directly to the home itself, the company may provide **FTTN** (**Fiber to the Node**) (or neighborhood), whereby fiber is terminated in a cabinet on a street sometimes several miles from the customer home. Fiber to the Distribution Point (FTTdp), as mentioned above, moves fiber one step closer to the customer home, often bringing fiber to within a few meters of the customer premises. In between these options is **FTTC** (**Fiber to the Curb**). All of these **FTTX** (**Fiber to the X**) designs are sometimes also called ''fiber in the loop'' because some amount of fiber is used in the local loop.

Several variations of the form ''FTTX'' (where *X* stands for the basement, curb, or neighborhood) exist. They are used to note that the fiber deployment may reach close to the house. In this case, copper (twisted pair or coaxial cable) provides fast enough speeds over the last short distance. The choice of how far to lay

the fiber is an economic one, balancing cost with expected revenue. In any case, the point is that optical fiber has crossed the traditional barrier of the ''last mile.'' We will focus on FTTH in our discussion.

Like the copper wires before it, the fiber local loop is passive, which means no powered equipment is required to amplify or otherwise process signals. The fiber simply carries signals between the home and the end office. This, in turn, reduces cost and improves reliability. Usually, the fibers from the houses are joined together so that only a single fiber reaches the end office per group of up to 100 houses. In the downstream direction, optical splitters divide the signal from the end office so that it reaches all the houses. Encryption is needed for security if only one house should be able to decode the signal. In the upstream direction, optical combiners merge the signals from the houses into a single signal that is received at the end office.

This architecture is called a **PON** (**Passive Optical Network**), and it is shown in Fig. 2-31. It is common to use one wavelength shared between all the houses for downstream transmission, and another wavelength for upstream transmission.

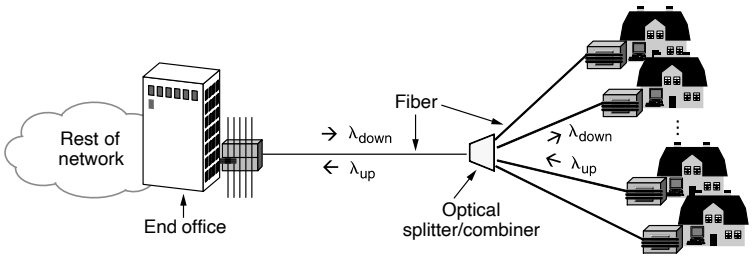


Figure 2-31. Passive optical network for Fiber To The Home.

Even with the splitting, the tremendous bandwidth and low attenuation of fiber mean that PONs can provide high rates to users over distances of up to 20 km. The actual data rates and other details depend on the type of PON. Two kinds are common. **GPONs** (**Gigabit-capable PONs**) come from the world of telecommunications, so they are defined by an ITU standard. **EPONs** (**Ethernet PONs**) are more in tune with the world of networking, so they are defined by an IEEE standard. Both run at around a gigabit and can carry traffic for different services, including Internet, video, and voice. For example, GPONs provide 2.4 Gbps downstream and 1.2 or 2.4 Gbps upstream.

Additional protocols are needed to share the capacity of the single fiber at the end office between the different houses. The downstream direction is quite easy. The end office can send messages to each different house in whatever order it likes. In the upstream direction, however, messages from different houses cannot be sent at the same time, or different signals would collide. The houses also cannot hear each other's transmissions so they cannot listen before transmitting. The solution

is that equipment at the houses requests and is granted time slots to use by equipment in the end office. For this to work, there is a ranging process to adjust the transmission times from the houses so that all the signals received at the end office are synchronized. The design is similar to cable modems, which we cover later in this chapter. For more information on PONs, see Grobe and Elbers (2008) or Andrade et al. (2014).

2.5.3 Trunks and Multiplexing

Trunks in the telephone network are not only much faster than the local loops, they are different in two other respects. The core of the telephone network carries digital information, not analog information; that is, bits not voice. This necessitates a conversion at the end office to digital form for transmission over the long-haul trunks. The trunks carry thousands, even millions, of calls simultaneously. This sharing is important for achieving economies of scale, since it costs essentially the same amount of money to install and maintain a high-bandwidth trunk as a low-bandwidth trunk between two switching offices. It is accomplished with versions of TDM and FDM.

Below, we will briefly examine how voice signals are digitized so that they can be transported by the telephone network. After that, we will see how TDM is used to carry bits on trunks, including the TDM system used for fiber optics (SONET). Then, we will turn to FDM as it is applied to fiber optics, which is called wavelength division multiplexing.

Digitizing Voice Signals

Early in the development of the telephone network, the core handled voice calls as analog information. FDM techniques were used for many years to multiplex 4000-Hz voice channels (each comprising 3100 Hz plus guard bands) into larger and larger units. For example, 12 calls in the 60 kHz–to–108 kHz band are known as a **group**, five groups (a total of 60 calls) are known as a **supergroup**, and so on. These FDM methods are still used over some copper wires and microwave channels. However, FDM requires analog circuitry and is not amenable to being done by a computer. In contrast, TDM can be handled entirely by digital electronics, so it has become far more widespread in recent years. Since TDM can only be used for digital data and the local loops produce analog signals, a conversion is needed from analog to digital in the end office, where all the individual local loops come together to be combined onto outgoing trunks.

The analog signals are digitized in the end office by a device called a **codec** (short for "*co*der-*dec*oder") using a technique is called **PCM** (**Pulse Code Modulation**), which forms the heart of the modern telephone system. The codec makes 8000 samples per second (125 μsec/sample) because the Nyquist theorem says that this is sufficient to capture all the information from the 4-kHz telephone channel

bandwidth. At a lower sampling rate, information would be lost; at a higher one, no extra information would be gained. Almost all time intervals within the telephone system are multiples of 125 μsec. The standard uncompressed data rate for a voice-grade telephone call is thus 8 bits every 125 μsec, or 64 kbps.

Each sample of the amplitude of the signal is quantized to an 8-bit number. To reduce the error due to quantization, the quantization levels are unevenly spaced. A logarithmic scale is used that gives relatively more bits to smaller signal amplitudes and relatively fewer bits to large signal amplitudes. In this way, the error is proportional to the signal amplitude. Two versions of quantization are widely used: **μ-law**, used in North America and Japan, and **A-law**, used in Europe and the rest of the world. Both versions are specified in standard ITU G.711. An equivalent way to think about this process is to imagine that the dynamic range of the signal (or the ratio between the largest and smallest possible values) is compressed before it is (evenly) quantized, and then expanded when the analog signal is recreated. For this reason, it is called **companding**. It is also possible to compress the samples after they are digitized so that they require much less than 64 kbps. However, we will leave this topic for when we explore audio applications such as voice over IP.

At the other end of the call, an analog signal is recreated from the quantized samples by playing them out (and smoothing them) over time. It will not be exactly the same as the original analog signal, even though we sampled at the Nyquist rate, because the samples were quantized.

T-Carrier: Multiplexing Digital Signals on the Phone Network

The **T-Carrier** is a specification for transmitting multiple TDM channels over a single circuit. TDM with PCM is used to carry multiple voice calls over trunks by sending a sample from each call every 125 μsec. When digital transmission began emerging as a feasible technology, ITU (then called CCITT) was unable to reach agreement on an international standard for PCM. Consequently, a variety of incompatible schemes are now in use in different countries around the world.

The method used in North America and Japan is the **T1** carrier, depicted in Fig. 2-32. (Technically speaking, the format is called DS1 and the carrier is called T1, but following widespread industry tradition, we will not make that subtle distinction here.) The T1 carrier consists of 24 voice channels multiplexed together. Each of the 24 channels, in turn, gets to insert 8 bits into the output stream. The T1 carrier was introduced in 1962.

A frame consists of $24 \times 8 = 192$ bits plus one extra bit for control purposes, yielding 193 bits every 125 μsec. This gives a gross data rate of 1.544 Mbps, of which 8 kbps is for signaling. The 193rd bit is used for frame synchronization and signaling. In one variation, the 193rd bit is used across a group of 24 frames called an **extended superframe**. Six of the bits, in the 4th, 8th, 12th, 16th, 20th, and 24th positions, take on the alternating pattern 001011 . . . . Normally, the receiver

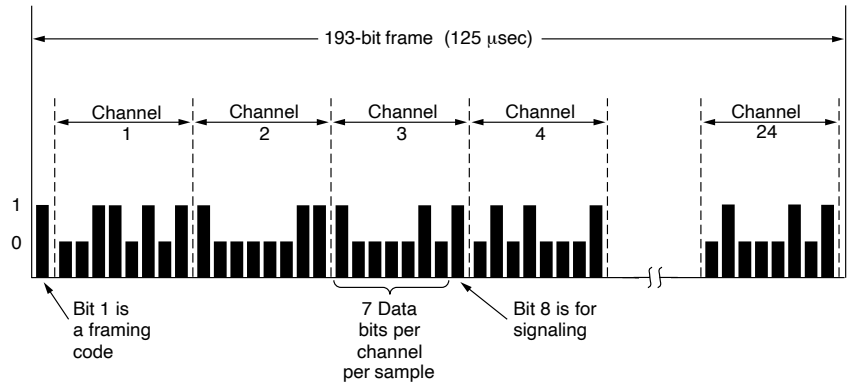


Figure 2-32. The T1 carrier (1.544 Mbps).

keeps checking for this pattern to make sure that it has not lost synchronization. Six more bits are used to send an error check code to help the receiver confirm that it is synchronized. If it does get out of sync, the receiver can scan for the pattern and validate the error check code to get resynchronized. The remaining 12 bits are used for control information for operating and maintaining the network, such as performance reporting from the remote end.

The T1 format has several variations. The earlier versions sent signaling information **in-band**, meaning in the same channel as the data, by using some of the data bits. This design is one form of **channel-associated signaling**, because each channel has its own private signaling subchannel. In one arrangement, the least significant bit out of an 8-bit sample on each channel is used in every sixth frame. It has the colorful name of **robbed-bit signaling**. The idea is that a few stolen bits will not matter for voice calls. No one will hear the difference.

For data, however, it is another story. Delivering the wrong bits is unhelpful, to say the least. If older versions of T1 are used to carry data, only 7 of 8 bits, or 56 kbps, can be used in each of the 24 channels. Instead, newer versions of T1 provide clear channels in which all of the bits may be used to send data. Clear channels are what businesses who lease a T1 line want when they send data across the telephone network in place of voice samples. Signaling for any voice calls is then handled **out-of-band**, meaning in a separate channel from the data. Often, the signaling is done with **common-channel signaling** in which there is a shared signaling channel. One of the 24 channels may be used for this purpose.

Outside of North America and Japan, the 2.048-Mbps **E1** carrier is used instead of T1. This carrier has 32 8-bit data samples packed into the basic 125-μsec frame. Thirty of the channels are used for information and up to two are used for signaling. Each group of four frames provides 64 signaling bits, half of which are

used for signaling (whether channel-associated or common-channel) and half of which are used for frame synchronization or are reserved for each country to use as it wishes.

Time division multiplexing allows multiple T1 carriers to be multiplexed into higher-order carriers. Figure 2-33 shows how this can be done. At the left, we see four T1 channels being multiplexed into one T2 channel. The multiplexing at T2 and above is done bit for bit, rather than byte for byte with the 24 voice channels that make up a T1 frame. Four T1 streams at 1.544 Mbps really ought to generate 6.176 Mbps, but T2 is actually 6.312 Mbps. The extra bits are used for framing and recovery in case the carrier slips.

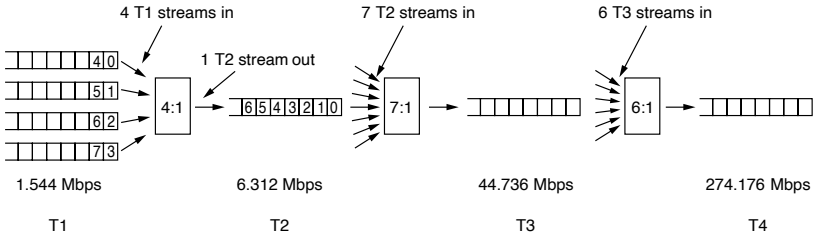


Figure 2-33. Multiplexing T1 streams into higher carriers.

At the next level, seven T2 streams are combined bitwise to form a T3 stream. Then, six T3 streams are joined to form a T4 stream. At each step, a small amount of overhead is added for framing and recovery in case the synchronization between sender and receiver is lost. T1 and T3 are widely used by customers, whereas T2 and T4 are only used within the telephone system itself, so they are not well-known.

Just as there is little agreement on the basic carrier between the United States and the rest of the world, there is equally little agreement on how it is to be multiplexed into higher-bandwidth carriers. The U.S. scheme of stepping up by 4, 7, and 6 did not strike everyone else as the way to go, so the ITU standard calls for multiplexing four streams into one stream at each level. Also, the framing and recovery data are different in the U.S. and ITU standards. The ITU hierarchy for 32, 128, 512, 2048, and 8192 channels runs at speeds of 2.048, 8.848, 34.304, 139.264, and 565.148 Mbps.

Multiplexing Optical Networks: SONET/SDH

In the early days of fiber optics, every telephone company had its own proprietary optical TDM system. After the U.S. government broke up AT&T in 1984, local telephone companies had to connect to multiple long-distance carriers, all

with optical TDM systems from different vendors and suppliers, so the need for standardization became obvious. In 1985, Bellcore, the research arm of the Regional Bell Operating Companies (RBOCs), began working on a standard, called **SONET** (**Synchronous Optical NETwork**).

Later, ITU joined the effort, which resulted in a SONET standard and a set of parallel ITU recommendations (G.707, G.708, and G.709) in 1989. The ITU recommendations are called **SDH** (**Synchronous Digital Hierarchy**) but differ from SONET only in minor ways. Virtually all of the long-distance telephone traffic in the United States, and much of it elsewhere, now uses trunks running SONET in the physical layer. For additional information about SONET, see Perros (2005).

The SONET design had four major goals:

1. Carrier interoperability: SONET had to make it possible for different carriers to interoperate. Achieving this goal required defining a common signaling standard with respect to wavelength, timing, framing structure, and other issues.
2. Unification across regions: some means was needed to unify the U.S., European, and Japanese digital systems, all of which were based on 64-kbps PCM channels but combined them in different (and incompatible) ways.
3. Multiplexing digital channels: SONET had to provide a way to multiplex multiple digital channels. At the time SONET was devised, the highest-speed digital carrier actually used widely in the United States was T3, at 44.736 Mbps. T4 was defined, but not used much, and nothing was even defined above T4 speed. Part of SONET's mission was to continue the hierarchy to gigabits/sec and beyond. A standard way to multiplex slower channels into one SONET channel was also needed.
4. Management support: SONET had to provide support for operations, administration, and maintenance (OAM), which are needed to manage the network. Previous systems did not do this very well.

An early decision was to make SONET a conventional TDM system, with the entire bandwidth of the fiber devoted to one channel containing time slots for the various subchannels. As such, SONET is a synchronous system. Each sender and receiver is tied to a common clock. The master clock that controls the system has an accuracy of about 1 part in 10^9. Bits on a SONET line are sent out at extremely precise intervals, controlled by the master clock.

The basic SONET frame is a block of 810 bytes put out every 125 μsec. Since SONET is synchronous, frames are emitted whether or not there are any useful data to send. Having 8000 frames/sec exactly matches the sampling rate of the PCM channels used in all digital telephony systems.

The 810-byte SONET frames are best thought of as a rectangle of bytes, 90 columns wide by 9 rows high. Thus, $8 \times 810 = 6480$ bits are transmitted 8000 times per second, for a gross data rate of 51.84 Mbps. This layout is the basic SONET channel, called **STS-1** (**Synchronous Transport Signal-1**). All SONET trunks are multiples of STS-1.

The first three columns of each frame are reserved for system management information, as illustrated in Fig. 2-34. In this block, the first three rows contain the section overhead; the next six contain the line overhead. The section overhead is generated and checked at the start and end of each section, whereas the line overhead is generated and checked at the start and end of each line.

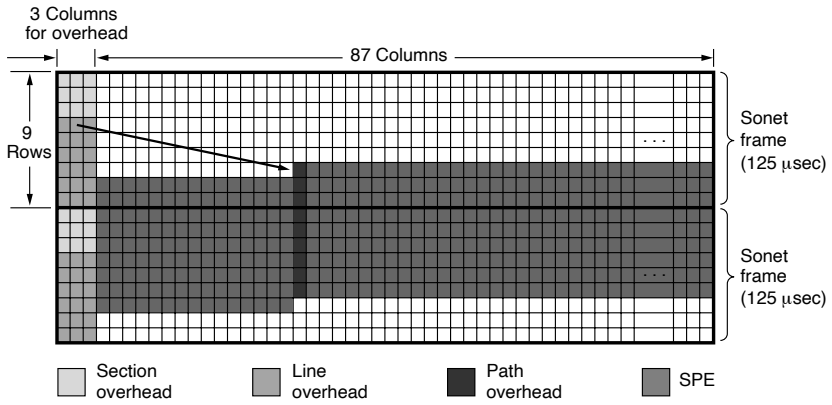


Figure 2-34. Two back-to-back SONET frames.

A SONET transmitter sends back-to-back 810-byte frames, without gaps between them, even when there are no data (in which case it sends dummy data). From the receiver's point of view, all it sees is a continuous bit stream, so how does it know where each frame begins? The answer is that the first 2 bytes of each frame contain a fixed pattern that the receiver searches for. If it finds this pattern in the same place in a large number of consecutive frames, it assumes that it is in sync with the sender. In theory, a user could insert this pattern into the payload in a regular way, but in practice, it cannot be done due to the multiplexing of multiple users into the same frame and other reasons.

The final 87 columns of each frame hold $87 \times 9 \times 8 \times 8000 = 50.112$ Mbps of user data. This user data could be voice samples, T1 and other carriers, or packets. SONET is simply a container for transporting bits. The **SPE** (**Synchronous Payload Envelope**), which carries the user data does not always begin in row 1, column 4. The SPE can begin anywhere within the frame. A pointer to the first byte is contained in the first row of the line overhead. The first column of the SPE is the path overhead (i.e., the header for the end-to-end path sublayer protocol).

The ability to allow the SPE to begin anywhere within the SONET frame and even to span two frames, as shown in Fig. 2-34, gives added flexibility to the system. For example, if a payload arrives at the source while a dummy SONET frame is being constructed, it can be inserted into the current frame instead of being held until the start of the next one.

The SONET/SDH multiplexing hierarchy is shown in Fig. 2-35. Rates from STS-1 to STS-768 have been defined, ranging from roughly a T3 line to 40 Gbps. Even higher rates will surely be defined over time, with OC-3072 at 160 Gbps being the next in line if and when it becomes technologically feasible. The optical carrier corresponding to STS-*n* is called OC-*n* but is bit for bit the same except for a certain bit reordering needed for synchronization. The SDH names are different, and they start at OC-3 because ITU-based systems do not have a rate near 51.84 Mbps. We have shown the common rates, which proceed from OC-3 in multiples of four. The gross data rate includes all the overhead. The SPE data rate excludes the line and section overhead. The user data rate excludes all three kinds of overhead and counts only the 86 payload columns.

SONET		SDH	Data rate (Mbps)		
Electrical	**Optical**	**Optical**	**Gross**	**SPE**	**User**
STS-1	OC-1		51.84	50.112	49.536
STS-3	OC-3	STM-1	155.52	150.336	148.608
STS-12	OC-12	STM-4	622.08	601.344	594.432
STS-48	OC-48	STM-16	2488.32	2405.376	2377.728
STS-192	OC-192	STM-64	9953.28	9621.504	9510.912
STS-768	OC-768	STM-256	39813.12	38486.016	38043.648

Figure 2-35. SONET and SDH multiplex rates.

As an aside, when a carrier, such as OC-3, is not multiplexed, but carries the data from only a single source, the letter *c* (for concatenated) is appended to the designation, so OC-3 indicates a 155.52-Mbps carrier consisting of three separate OC-1 carriers, but OC-3c indicates a data stream from a single source at 155.52 Mbps. The three OC-1 streams within an OC-3c stream are interleaved by column—first column 1 from stream 1, then column 1 from stream 2, then column 1 from stream 3, followed by column 2 from stream 1, and so on—leading to a frame 270 columns wide and 9 rows deep.

2.5.4 Switching

From the point of view of the average telephone engineer, the phone system has two principal parts: outside plant (the local loops and trunks, since they are physically outside the switching offices) and inside plant (the switches, which are

inside the switching offices). We have just looked at the outside plant. Now, it is time to examine the inside plant.

Two different switching techniques are used by the network nowadays: circuit switching and packet switching. The traditional telephone system is based on circuit switching, although voice over IP technology relies on packet switching. We will go into circuit switching in some detail and contrast it with packet switching. Both kinds of switching are important enough that we will come back to them when we get to the network layer.

Circuit Switching

Traditionally, when you or your computer placed a telephone call, the switching equipment within the telephone system sought out a physical path all the way from your telephone to the receiver's telephone and maintained it for the duration of the call. This technique is called **circuit switching**. It is shown schematically in Fig. 2-36(a). Each of the six rectangles represents a carrier switching office (end office, toll office, etc.). In this example, each office has three incoming lines and three outgoing lines. When a call passes through a switching office, a physical connection is established between the line on which the call came in and one of the output lines, as shown by the dotted lines.

In the early days of the telephone, the connection was made by the operator plugging a jumper cable into the input and output sockets. In fact, a surprising little story is associated with the invention of automatic circuit-switching equipment. It was invented by a 19th-century Missouri undertaker named Almon B. Strowger. Shortly after the telephone was invented, when someone died, one of the survivors would call the town operator and say ''Please connect me to an undertaker.'' Unfortunately for Mr. Strowger, there were two undertakers in his town, and the other one's wife was the town telephone operator. He quickly saw that either he was going to have to invent automatic telephone switching equipment or he was going to go out of business. He chose the first option. For nearly 100 years, the circuit-switching equipment used worldwide was known as **Strowger gear**. (History does not record whether the now-unemployed switchboard operator got a job as an information operator, answering questions such as ''What is the phone number of an undertaker?'')

The model shown in Fig. 2-36(a) is highly simplified, of course, because parts of the physical path between the two telephones may, in fact, be microwave or fiber links onto which thousands of calls are multiplexed. Nevertheless, the basic idea is valid: once a call has been set up, a dedicated path between both ends exists and will continue to exist until the call is finished.

An important property of circuit switching is the need to set up an end-to-end path *before* any data can be sent. The elapsed time between the end of dialing and the start of ringing can sometimes be 10 seconds, more on long-distance or international calls. During this time interval, the telephone system is hunting for a path,

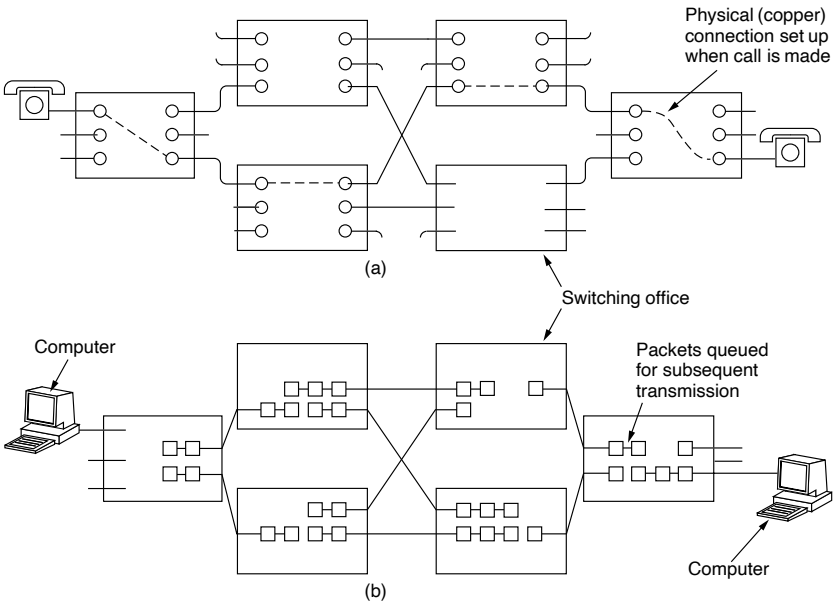


Figure 2-36. (a) Circuit switching. (b) Packet switching.

as shown in Fig. 2-37(a). Note that before data transmission can even begin, the call request signal must propagate all the way to the destination and be acknowledged. For many computer applications (e.g., point-of-sale credit verification), long setup times are undesirable.

As a consequence of the reserved path between the calling parties, once the setup has been completed, the only delay for data is the propagation time for the electromagnetic signal: about 5 milliseconds per 1000 km. Also, as a consequence of the established path, there is no danger of congestion—that is, once the call has been put through, you never get busy signals. Of course, you might get one before the connection has been established due to lack of switching or trunk capacity.

Packet Switching

The alternative to circuit switching is **packet switching**, shown in Fig. 2-36(b) and described in Chap. 1. With this technology, packets are sent as soon as they are available. In contrast to circuit switching, there is no need to set up a dedicated path in advance. Packet switching is analogous to sending a series of letters using the postal system: each one travels independently of the others. It is up to routers

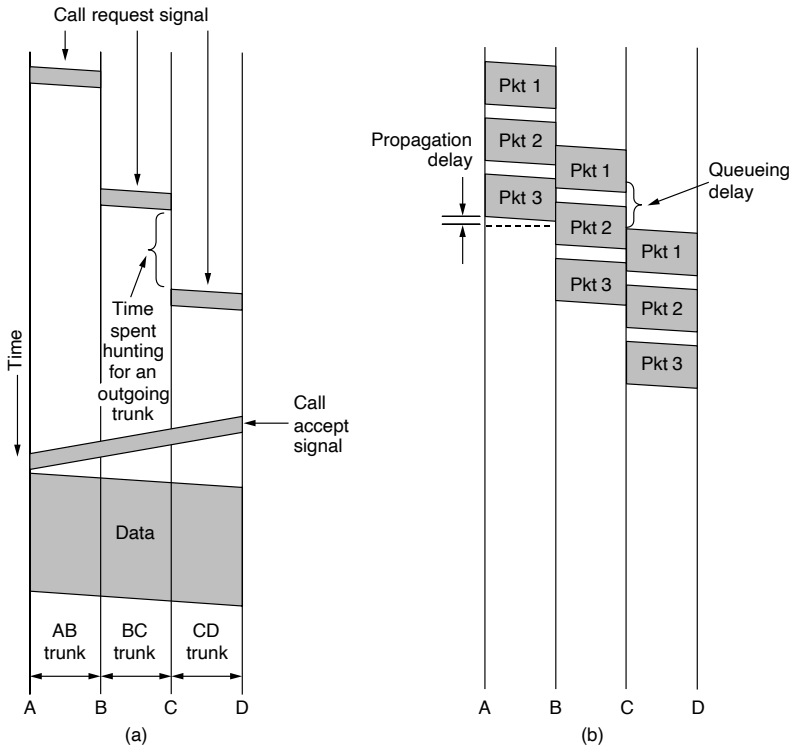


Figure 2-37. Timing of events in (a) circuit switching, (b) packet switching.

to use store-and-forward transmission to send each packet on its way toward the destination on its own. This procedure is unlike circuit switching, where the result of the connection setup is the reservation of bandwidth all the way from the sender to the receiver and all data on the circuit follows this path. In circuit switching, having all the data follow the same path means that it cannot arrive out of order. With packet switching, there is no fixed path, so different packets can follow different paths, depending on network conditions at the time they are sent, and they may arrive out of order.

Packet-switching networks place a tight upper limit on the size of packets. This ensures that no user can monopolize any transmission line for very long (e.g., many milliseconds), so that packet-switched networks can handle interactive traffic. It also reduces delay since the first packet of a long message can be forwarded before the second one has fully arrived. However, the store-and-forward delay of accumulating a packet in the router's memory before it is sent on to the next router

exceeds that of circuit switching. With circuit switching, the bits just flow through the wire continuously. Nothing is ever stored and forwarded later.

Packet and circuit switching also differ in other ways. Because no bandwidth is reserved with packet switching, packets may have to wait to be forwarded. This introduces **queueing delay** and congestion if many packets are sent at the same time. On the other hand, there is no danger of getting a busy signal and being unable to use the network. Thus, congestion occurs at different times with circuit switching (at setup time) and packet switching (when packets are sent).

If a circuit has been reserved for a particular user and there is no traffic, its bandwidth is wasted. It cannot be used for other traffic. Packet switching does not waste bandwidth and thus is more efficient from a system perspective. Understanding this trade-off is crucial for comprehending the difference between circuit switching and packet switching. The trade-off is between guaranteed service and wasting resources versus not guaranteeing service and not wasting resources.

Packet switching is more fault tolerant than circuit switching. In fact, that is why it was invented. If a switch goes down, all of the circuits using it are terminated and no more traffic can be sent on any of them. With packet switching, packets can be routed around dead switches.

Another difference between circuit and packet switching is how traffic is billed. With circuit switching (i.e., for voice telephone calls over the PSTN), billing has historically been based on distance and time. For mobile voice, distance usually does not play a role, except for international calls, and time plays only a coarse role (e.g., a calling plan with 2000 free minutes costs more than one with 1000 free minutes and sometimes nights or weekends are cheap). With packet-switched networks, including both fixed-line and mobile networks, time connected is not an issue, but the volume of traffic is. For home users in the United States and Europe, ISPs usually charge a flat monthly rate because it is less work for them and their customers can understand this model. In some developing countries, billing is often still volume-based: users may purchase a ''data bundle'' of a certain size and use that data over the course of a billing cycle. Certain times of day, or even certain destinations, may be free of charge or not count against the data cap or quota; these services are sometimes called **zero-rated services**. Generally, carrier Internet service providers in the Internet backbone charge based on traffic volumes. A typical billing model is based on the 95th percentile of five-minute samples: on a given link, an ISP will measure the volume of traffic that has passed over the link in the last five minutes. A 30-day billing cycle will have 8640 such five-minute intervals, and the ISP will bill based on the 95th percentile of these samples. This technique is often called **95th percentile billing**.

The differences between circuit switching and packet switching are summarized in Fig. 2-38. Traditionally, telephone networks have used circuit switching to provide high-quality telephone calls, and computer networks have used packet switching for simplicity and efficiency. However, there are notable exceptions. Some older computer networks have been circuit switched under the covers (e.g.,

X.25) and some newer telephone networks use packet switching with voice over IP technology. This looks just like a standard telephone call on the outside to users, but inside the network packets of voice data are switched. This approach has let upstarts market cheap international calls via calling cards, though perhaps with lower call quality than the incumbents.

Item	Circuit switched	Packet switched
Call setup	Required	Not needed
Dedicated physical path	Yes	No
Each packet follows the same route	Yes	No
Packets arrive in order	Yes	No
Is a switch crash fatal	Yes	No
Bandwidth available	Fixed	Dynamic
Time of possible congestion	At setup time	On every packet
Potentially wasted bandwidth	Yes	No
Store-and-forward transmission	No	Yes
Charging	Per minute	Per byte

Figure 2-38. A comparison of circuit-switched and packet-switched networks.

2.6 CELLULAR NETWORKS

Even if the conventional telephone system someday gets multigigabit end-to-end fiber, people now expect to make phone calls and to use their phones to check email and surf the Web from airplanes, cars, swimming pools, and while jogging in the park. Consequently, there is a tremendous amount of interest (and investment) in wireless telephony.

The mobile phone system is used for wide area voice and data communication. **Mobile phones** (sometimes called **cell phones**) have gone through five distinct generations, widely called 1G, 2G, 3G, 4G, and 5G. The initial three generations provided analog voice, digital voice, and both digital voice and data (Internet, email, etc.), respectively. 4G technology adds additional capabilities, including additional physical layer transmission techniques (e.g., OFDM uplink transmissions), and IP-based femtocells (home cellular nodes that are connected to fixed-line Internet infrastructure). 4G does not support circuit-switched telephony, unlike its predecessors; it is based on packet switching only. 5G is being rolled out now, but it will take years before it completely replaces the earlier generations everywhere. 5G technology will support up to 20 Gbps transmissions, as well as denser deployments. There is also some focus on reducing network latency to support a wider range of applications, for example, highly interactive gaming.

2.6.1 Common Concepts: Cells, Handoff, Paging

In all mobile phone systems, a geographic region is divided up into **cells**, which is why the handsets are sometimes called cell phones. Each cell uses some set of frequencies not used by any of its neighbors. The key idea that gives cellular systems far more capacity than previous systems is the use of relatively small cells and the reuse of transmission frequencies in nearby (but not adjacent) cells. The cellular design increases the system capacity as the cells get smaller. Furthermore, smaller cells mean that less power is needed, which leads to smaller and cheaper transmitters and handsets.

Cells allow for frequency reuse, which is illustrated in Fig. 2-39(a). The cells are normally roughly circular, but they are easier to model as hexagons. In Fig. 2-39(a), the cells are all the same size. They are grouped in units of seven cells. Each letter indicates a group of frequencies. Notice that for each frequency set, there is a buffer about two cells wide where that frequency is not reused, providing for good separation and low interference.

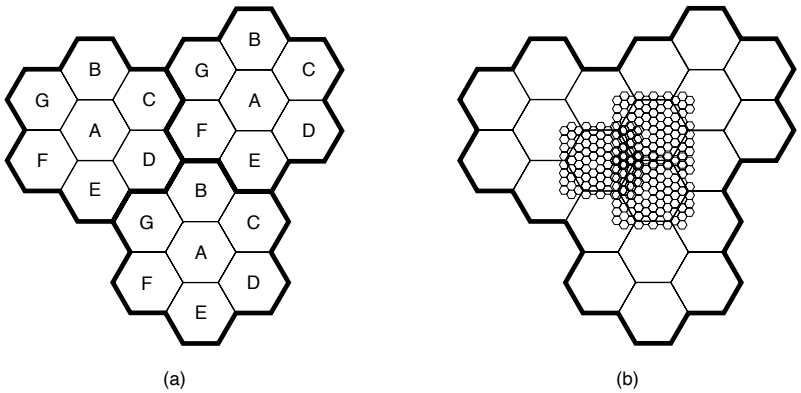


Figure 2-39. (a) Frequencies are not reused in adjacent cells. (b) To add more users, smaller cells can be used.

In an area where the number of users has grown to the point that the system is overloaded, the power can be reduced and the overloaded cells split into smaller **microcells** to permit more frequency reuse, as shown in Fig. 2-39(b). Telephone companies sometimes create temporary microcells, using portable towers with satellite links at sporting events, rock concerts, and other places where large numbers of mobile users congregate for a few hours.

At the center of each cell is a base station to which all the telephones in the cell transmit. The base station consists of a computer and transmitter/receiver connected to an antenna. In a small system, all the base stations are connected to a

single device called an **MSC** (**Mobile Switching Center**) or **MTSO** (**Mobile Telephone Switching Office**). In a larger one, several MSCs may be needed, all of which are connected to a second-level MSC, and so on. The MSCs are essentially end offices as in the telephone system, and are in fact connected to at least one telephone system end office. The MSCs communicate with the base stations, each other, and the PSTN using a packet-switching network.

At any instant, each mobile telephone is logically in one specific cell and under the control of that cell's base station. When a mobile telephone physically leaves a cell, its base station notices the telephone's signal fading away and then asks all the surrounding base stations how much power they are getting from it. When the answers come back, the base station then transfers ownership to the cell getting the strongest signal; under most conditions that is the cell where the telephone is now located. The telephone is then informed of its new boss, and if a call is in progress, it is asked to switch to a new channel (because the old one is not reused in any of the adjacent cells). This process, called **handoff**, takes about 300 milliseconds. Channel assignment is done by the MSC, the nerve center of the system. The base stations are really just dumb radio relays.

Finding locations high in the air to place base station antennas is a major issue. This problem has led some telecommunication carriers to forge alliances with the Roman Catholic Church, since the latter owns a substantial number of exalted potential antenna sites worldwide, all conveniently under a single management.

Cellular networks typically have four types of **channels**. **Control channels** (base to mobile) are used to manage the system. **Paging channels** (base to mobile) alert mobile users to calls for them. **Access channels** (bidirectional) are used for call setup and channel assignment. Finally, **data channels** (bidirectional) carry voice, fax, or data.

2.6.2 First-Generation (1G) Technology: Analog Voice

Let us look at cellular network technology, starting with the earliest system. Mobile radiotelephones were used sporadically for maritime and military communication during the early decades of the 20th century. In 1946, the first system for car-based telephones was set up in St. Louis. This system used a single large transmitter on top of a tall building and had a single channel, used for both sending and receiving. To talk, the user had to push a button that enabled the transmitter and disabled the receiver. Such systems, known as **push-to-talk systems**, were installed beginning in the 1950s. Taxis and police cars often used this technology.

In the 1960s, **IMTS** (**Improved Mobile Telephone System**) was installed. It, too, used a high-powered (200-watt) transmitter on top of a hill but it had two frequencies, one for sending and one for receiving, so the push-to-talk button was no longer needed. Since all communication from the mobile telephones went inbound on a different channel than the outbound signals, the mobile users could not hear each other (unlike the push-to-talk system used in older taxis).

IMTS supported 23 channels spread out from 150 MHz to 450 MHz. Due to the small number of channels, users often had to wait a long time before getting a dial tone. Also, due to the large power of the hilltop transmitters, adjacent systems had to be several hundred kilometers apart to avoid interference. All in all, the limited capacity made the system impractical.

AMPS (**Advanced Mobile Phone System**), an analog mobile phone system invented by Bell Labs and first deployed in the United States in 1983, significantly increased the capacity of the cellular network. It was also used in England, where it was called TACS, and in Japan, where it was called MCS-L1. AMPS was formally retired in 2008, but we will look at it to understand the context for the 2G and 3G systems that improved on it. In AMPS, cells are typically 10 to 20 km across; in digital systems, the cells are smaller. Whereas an IMTS system 100 km across can have only one call on each frequency, an AMPS system might have 100 10-km cells in the same area and be able to have 10 to 15 calls on each frequency, in widely separated cells.

AMPS uses FDM to separate the channels. The system uses 832 full-duplex channels, each consisting of a pair of simplex channels. This arrangement is known as **FDD** (**Frequency Division Duplex**). The 832 simplex channels from 824 to 849 MHz are used for mobile to base station transmission, and 832 simplex channels from 869 to 894 MHz are used for base station to mobile transmission. Each of these simplex channels is 30 kHz wide.

The 832 channels in AMPS are divided into four categories. Since the same frequencies cannot be reused in nearby cells and 21 channels are reserved in each cell for control, the actual number of voice channels available per cell is much smaller than 832, typically about 45.

Call Management

Each mobile telephone in AMPS has a 32-bit serial number and a 10-digit telephone number in its programmable read-only memory. The telephone number in many countries is represented as a 3-digit area code in 10 bits and a 7-digit subscriber number in 24 bits. When a phone is switched on, it scans a preprogrammed list of 21 control channels to find the most powerful signal. The phone then broadcasts its 32-bit serial number and 34-bit telephone number. Like all the control information in AMPS, this packet is sent in digital form, multiple times, and with an error-correcting code, even though the voice channels themselves are analog.

When the base station hears the announcement, it tells the MSC, which records the existence of its new customer and also informs the customer's home MSC of his current location. During normal operation, the mobile telephone reregisters about once every 15 minutes.

To make a call, a mobile user switches on the phone, (at least conceptually) enters the number to be called on the keypad, and hits the CALL button. The phone then transmits the number to be called and its own identity on the access

channel. If a collision occurs there, it tries again later. When the base station gets the request, it informs the MSC. If the caller is a customer of the MSC's company (or one of its partners), the MSC looks for an idle channel for the call. If one is found, the channel number is sent back on the control channel. The mobile phone then automatically switches to the selected voice channel and waits until the called party picks up the phone.

Incoming calls work differently. To start with, all idle phones continuously listen to the paging channel to detect messages directed at them. When a call is placed to a mobile phone (either from a fixed phone or another mobile phone), a packet is sent to the callee's home MSC to find out where it is. A packet is then sent to the base station in its current cell, which sends a broadcast on the paging channel of the form ''Unit 14, are you there?'' The called phone responds with a ''Yes'' on the access channel. The base then says something like: ''Unit 14, call for you on channel 3.'' At this point, the called phone switches to channel 3 and starts making ringing sounds (or playing some melody the owner was given as a birthday present).

2.6.3 Second-Generation (2G) Technology: Digital Voice

The first generation of mobile phones was analog; the second generation is digital. Switching to digital has several advantages. It provides capacity gains by allowing voice signals to be digitized and compressed. It improves security by allowing voice and control signals to be encrypted. This, in turn, deters fraud and eavesdropping, whether from intentional scanning or echoes of other calls due to RF propagation. Finally, it enables new services such as text messaging.

Just as there was no worldwide standardization during the first generation, there was also no worldwide standardization during the second, either. Several different systems were developed, and three have been widely deployed. **D-AMPS** (**Digital Advanced Mobile Phone System**) is a digital version of AMPS that coexists with AMPS and uses TDM to place multiple calls on the same frequency channel. It is described in International Standard IS-54 and its successor IS-136. **GSM** (**Global System for Mobile communications**) has emerged as the dominant system, and while it was slow to catch on in the United States it is now used virtually everywhere in the world. Like D-AMPS, GSM is based on a mix of FDM and TDM. **CDMA** (**Code Division Multiple Access**), described in **International Standard IS-95**, is a completely different kind of system and is based on neither FDM nor TDM. While CDMA has not become the dominant 2G system, its technology has become the basis for 3G systems.

Also, the name **PCS** (**Personal Communications Services**) is sometimes used in the marketing literature to indicate a second-generation (i.e., digital) system. Originally it meant a mobile phone using the 1900 MHz band, but that distinction is rarely made now. The dominant 2G system in most of the world is GSM which we now describe in detail.

2.6.4 GSM: The Global System for Mobile Communications

GSM started life in the 1980s as an effort to produce a single European 2G standard. The task was assigned to a telecommunications group called (in French) Groupe Specialé Mobile. The first GSM systems were deployed starting in 1991 and were a quick success. It soon became clear that GSM was going to be more than a European success, with the uptake stretching to countries as far away as Australia, so GSM was renamed to have a more worldwide appeal.

GSM and the other mobile phone systems we will study retain from 1G systems a design based on cells, frequency reuse across cells, and mobility with handoffs as subscribers move. It is the details that differ. Here, we will briefly discuss some of the main properties of GSM. However, the printed GSM standard is over 5000 [sic] pages long. A large fraction of this material relates to engineering aspects of the system, especially the design of receivers to handle multipath signal propagation, and synchronizing transmitters and receivers. None of this will be even mentioned here.

Fig. 2-40 shows that the GSM architecture is similar to the AMPS architecture, though the components have different names. The mobile itself is now divided into the handset and a removable chip with subscriber and account information called a **SIM card**, short for **Subscriber Identity Module**. It is the SIM card that activates the handset and contains secrets that let the mobile and the network identify each other and encrypt conversations. A SIM card can be removed and plugged into a different handset to turn that handset into your mobile as far as the network is concerned.

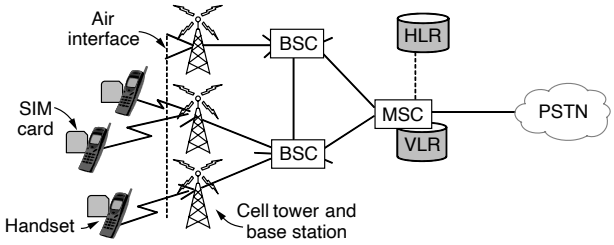


Figure 2-40. GSM mobile network architecture.

The mobile talks to cell base stations over an **air interface** that we will describe in a moment. The cell base stations are each connected to a **BSC** (**Base Station Controller**) that controls the radio resources of cells and handles handoff. The BSC in turn is connected to an MSC (as in AMPS) that routes calls and connects to the PSTN (Public Switched Telephone Network).

To be able to route calls, the MSC needs to know where mobiles can currently be found. It maintains a database of nearby mobiles that are associated with the

cells it manages. This database is called the **VLR** (**Visitor Location Register**). There is also a database in the mobile network that gives the last known location of each mobile. It is called the **HLR** (**Home Location Register**). This database is used to route incoming calls to the right locations. Both databases must be kept up to date as mobiles move from cell to cell.

We will now describe the air interface in some detail. GSM runs on a range of frequencies worldwide, including 900, 1800, and 1900 MHz. More spectrum is allocated than for AMPS in order to support a much larger number of users. GSM is a frequency division duplex cellular system, like AMPS. That is, each mobile transmits on one frequency and receives on another, higher frequency (55 MHz higher for GSM versus 80 MHz higher for AMPS). However, unlike with AMPS, with GSM a single frequency pair is split by time division multiplexing into time slots. In this way, it is shared by multiple mobiles.

To handle multiple mobiles, GSM channels are much wider than the AMPS channels (200 kHz versus 30 kHz). One 200-kHz channel is shown in Fig. 2-41. A GSM system operating in the 900-MHz region has 124 pairs of simplex channels. Each simplex channel is 200 kHz wide and supports eight separate connections on it, using time division multiplexing. Each currently active station is assigned one time slot on one channel pair. Theoretically, 992 channels can be supported in each cell, but many of them are not available, to avoid frequency conflicts with neighboring cells. In Fig. 2-41, the eight shaded time slots all belong to the same connection, four of them in each direction. Transmitting and receiving does not happen in the same time slot because the GSM radios cannot transmit and receive at the same time and it takes time to switch from one to the other. If the mobile device assigned to 890.4/935.4 MHz and time slot 2 wanted to transmit to the base station, it would use the lower four shaded slots (and the ones following them in time), putting some data in each slot until all the data had been sent.

The TDM slots shown in Fig. 2-41 are part of a complex framing hierarchy. Each TDM slot has a specific structure, and groups of TDM slots form multiframes, also with a specific structure. A simplified version of this hierarchy is shown in Fig. 2-42. Here we can see that each TDM slot consists of a 148-bit data frame that occupies the channel for 577 μsec (including a 30-μsec guard time after each slot). Each data frame starts and ends with three 0 bits, for frame delineation purposes. It also contains two 57-bit *Information* fields, each one having a control bit that indicates whether the following *Information* field is for voice or data. Between the *Information* fields is a 26-bit *Sync* (training) field that is used by the receiver to synchronize to the sender's frame boundaries.

A data frame is transmitted in 547 μsec, but a transmitter is only allowed to send one data frame every 4.615 msec, since it is sharing the channel with seven other stations. The gross rate of each channel is 270,833 bps, divided among eight users. However, as with AMPS, the overhead eats up a large fraction of the bandwidth, ultimately leaving 24.7 kbps worth of payload per user before error correction is applied. After error correction, 13 kbps is left for speech. While this is

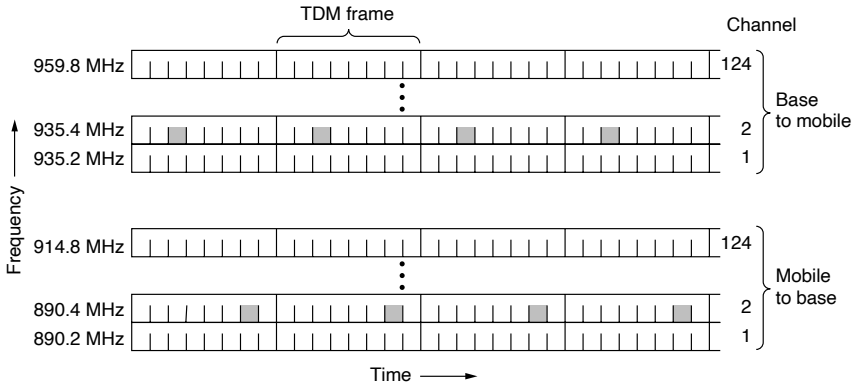


Figure 2-41. GSM uses 124 frequency channels, each of which uses an eight-slot TDM system.

substantially less than 64 kbps PCM for uncompressed voice signals in the fixed telephone network, compression on the mobile device can reach these levels with little loss of quality.

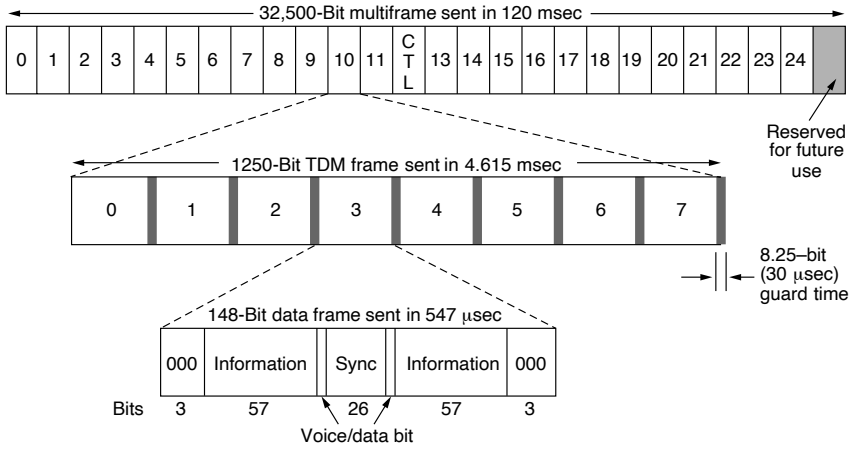


Figure 2-42. A portion of the GSM framing structure.

As can be seen from Fig. 2-42, eight data frames make up a TDM frame and 26 TDM frames make up a 120-msec multiframe. Of the 26 TDM frames in a

multiframe, slot 12 is used for control and slot 25 is reserved for future use, so only 24 are available for user traffic.

However, in addition to the 26-slot multiframe shown in Fig. 2-42, a 51-slot multiframe (not shown) is also used. Some of these slots are used to hold several control channels used to manage the system. The **broadcast control channel** is a continuous stream of output from the base station containing the base station's identity and the channel status. All mobile stations monitor their signal strength to see when they have moved into a new cell.

The **dedicated control channel** is used for location updating, registration, and call setup. In particular, each BSC maintains a database of mobile stations currently under its jurisdiction, the VLR. Information needed to maintain the VLR is sent on the dedicated control channel.

The system also has a **common control channel**, which is split up into three logical subchannels. The first of these subchannels is the **paging channel**, which the base station uses to announce incoming calls. Each mobile station monitors it continuously to watch for calls it should answer. The second is the **random access channel**, which allows users to request a slot on the dedicated control channel. If two requests collide, they are garbled and have to be retried later. Using the dedicated control channel slot, the station can set up a call. The assigned slot is announced on the third subchannel, the **access grant channel**.

Finally, GSM differs from AMPS in how handoff is handled. In AMPS, the MSC manages it completely without help from the mobile devices. With time slots in GSM, the mobile is neither sending nor receiving most of the time. The idle slots are an opportunity for the mobile to measure signal quality to other nearby base stations. It does so and sends this information to the BSC. The BSC can use it to determine when a mobile is leaving one cell and entering another so it can perform the handoff. This design is called **MAHO** (**Mobile Assisted HandOff**).

2.6.5 Third-Generation (3G) Technology: Digital Voice and Data

The first generation of mobile phones was analog voice, and the second generation was digital voice. The third generation of mobile phones, or **3G** as it is called, is all about digital voice *and* data. A number of factors drove the industry to 3G technology. First, around the time of 3G, data traffic began to exceed voice traffic on the fixed network; similar trends began to emerge for mobile devices. Second, phone, Internet, and video services began to converge. The rise of smartphones, starting with Apple's iPhone, which was first released in 2007, accelerated the shift to mobile data. Data volumes are rising steeply with the popularity of iPhones. When the iPhone was first released, it used a **2.5G** network (essentially an enhanced 2G network) that did not have enough data capacity. Data-hungry iPhone users further drove the transition to 3G technologies, to support higher data transmission rates. A year later, in 2008, Apple released an updated version of its iPhone that could use the 3G data network.

Operators initially took small steps in the direction of 3G by going to what is sometimes called **2.5G**. One such system is **EDGE** (**Enhanced Data rates for GSM Evolution**), which is essentially GSM with more bits per symbol. The trouble is, more bits per symbol also means more errors per symbol, so EDGE has nine different schemes for modulation and error correction, differing in terms of how much of the bandwidth is devoted to fixing the errors introduced by the higher speed. EDGE is one step along an evolutionary path that is defined from GSM to other 3G technologies that we discuss in this section.

ITU tried to get a bit more specific about the 3G vision starting back around 1992. It issued a blueprint for getting there called **IMT-2000**, where IMT stood for **International Mobile Telecommunications**. The basic services that the IMT-2000 network was supposed to provide to its users are:

1. High-quality voice transmission.
2. Messaging (replacing email, fax, SMS, chat, etc.).
3. Multimedia (playing music, viewing videos, films, television, etc.).
4. Internet access (Web surfing, including pages with audio and video).

Additional services might be video conferencing, telepresence, group game playing, and m-commerce (waving your telephone at the cashier to pay in a store). Furthermore, all these services are supposed to be available worldwide (with automatic connection via a satellite when no terrestrial network can be located), instantly (always on), and with quality of service guarantees. In other words, pie in the sky.

ITU envisioned a single worldwide technology for IMT-2000, so manufacturers could build a single device that could be sold and used anywhere in the world. Having a single technology would also make life much simpler for network operators and would encourage more people to use the services.

As it turned out, this was more than a bit optimistic. The number 2000 stood for three things: (1) the year it was supposed to go into service, (2) the frequency it was supposed to operate at (in MHz), and (3) the bandwidth the service should have (in kbps). It did not make it on any of the three counts. Nothing was implemented by 2000. ITU recommended that all governments reserve spectrum at 2 GHz so devices could roam seamlessly from country to country. China reserved the required bandwidth but nobody else did. Finally, it was recognized that 2 Mbps is not currently feasible for users who are *too* mobile (due to the difficulty of performing handoffs quickly enough). More realistic is 2 Mbps for stationary indoor users, 384 kbps for people walking, and 144 kbps for connections in cars.

Despite these initial setbacks, a great deal has been accomplished since then. Several IMT-2000 proposals were made and, after some winnowing, it came down to two primary ones: (1) **WCDMA** (**Wideband CDMA**), proposed by Ericsson

and pushed by the European Union, which called it **UMTS** (**Universal Mobile Telecommunications System**) and (2) **CDMA2000**, proposed by Qualcomm in the United States

Both of these systems are more similar than different; both are based on broadband CDMA. WCDMA uses 5-MHz channels and CDMA2000 uses 1.25-MHz channels. If the Ericsson and Qualcomm engineers were put in a room and told to come to a common design, they probably could find one in an hour. The trouble is that the real problem is not engineering, but politics (as usual). Europe wanted a system that interworked with GSM, whereas the United States wanted a system that was compatible with one already widely deployed in the United States (IS-95). Each side (naturally) also supported its local company (Ericsson is based in Sweden; Qualcomm is in California). Finally, Ericsson and Qualcomm were involved in numerous lawsuits over their respective CDMA patents. To add to the confusion, UMTS became a single 3G standard with multiple incompatible options, including CDMA2000. This change was an effort to unify the various camps, but it just papers over the technical differences and obscures the focus of ongoing efforts. We will use UMTS to mean WCDMA, as distinct from CDMA2000.

Another improvement of WCDMA over the simplified CDMA scheme we described earlier is to allow different users to send data at different rates, independent of each other. This trick is accomplished naturally in CDMA by fixing the rate at which chips are transmitted and assigning different users chip sequences of different lengths. For example, in WCDMA, the chip rate is 3.84 Mchips/sec and the spreading codes vary from 4 to 256 chips. With a 256-chip code, around 12 kbps is left after error correction, and this capacity is sufficient for a voice call. With a 4-chip code, the user data rate is close to 1 Mbps. Intermediate-length codes give intermediate rates; in order to get to multiple Mbps, the mobile must use more than one 5-MHz channel at once.

We will focus our discussion on the use of CDMA in cellular networks, as it is the distinguishing feature of both systems. CDMA is neither FDM nor TDM but a kind of mix in which each user sends on the same frequency band at the same time. When it was first proposed for cellular systems, the industry gave it approximately the same reaction that Columbus first got from Queen Isabella when he proposed reaching India by sailing in the wrong direction. However, through the persistence of a single company, Qualcomm, CDMA succeeded as a 2G system (IS-95) and matured to the point that it became the technical basis for 3G.

To make CDMA work in the mobile phone setting requires more than the basic CDMA technique that we described in Sec. 2.4. Specifically, we described a system called **synchronous CDMA**, in which the chip sequences are exactly orthogonal. This design works when all users are synchronized on the start time of their chip sequences, as in the case of the base station transmitting to mobiles. The base station can transmit the chip sequences starting at the same time so that the signals will be orthogonal and able to be separated. However, it is difficult to synchronize the transmissions of independent mobile phones. Without some special efforts,

their transmissions would arrive at the base station at different times, with no guarantee of orthogonality. To let mobiles send to the base station without synchronization, we want code sequences that are orthogonal to each other at all possible offsets, not simply when they are aligned at the start.

While it is not possible to find sequences that are exactly orthogonal for this general case, long pseudorandom sequences come close enough. They have the property that, with high probability, they have a low **cross-correlation** with each other at all offsets. This means that when one sequence is multiplied by another sequence and summed up to compute the inner product, the result will be small; it would be zero if they were orthogonal. (Intuitively, random sequences should always look different from each other. Multiplying them together should then produce a random signal, which will sum to a small result.) This lets a receiver filter unwanted transmissions out of the received signal. Also, the **auto-correlation** of pseudorandom sequences is also small, with high probability, except at a zero offset. This means that when one sequence is multiplied by a delayed copy of itself and summed, the result will be small, except when the delay is zero. (Intuitively, a delayed random sequence looks like a different random sequence, and we are back to the cross-correlation case.) This lets a receiver lock onto the beginning of the wanted transmission in the received signal.

The use of pseudorandom sequences lets the base station receive CDMA messages from unsynchronized mobiles. However, an implicit assumption in our discussion of CDMA is that the power levels of all mobiles are the same at the receiver. If they are not, a small cross-correlation with a powerful signal might overwhelm a large auto-correlation with a weak signal. Thus, the transmit power on mobiles must be controlled to minimize interference between competing signals. It is this interference that limits the capacity of CDMA systems.

The power levels received at a base station depend on how far away the transmitters are as well as how much power they transmit. There may be many mobile stations at varying distances from the base station. A good heuristic to equalize the received power is for each mobile station to transmit to the base station at the inverse of the power level it receives from the base station. In other words, a mobile station receiving a weak signal from the base station will use more power than one getting a strong signal. For more accuracy, the base station also gives each mobile feedback to increase, decrease, or hold steady its transmit power. The feedback is frequent (1500 times per second) because good power control is important to minimize interference.

Now let us describe the advantages of CDMA. First, CDMA can improve capacity by taking advantage of small periods when some transmitters are silent. In polite voice calls, one party is silent while the other talks. On average, the line is busy only 40% of the time. However, the pauses may be small and are difficult to predict. With TDM or FDM systems, it is not possible to reassign time slots or frequency channels quickly enough to benefit from these small silences. However, in CDMA, by simply not transmitting one user lowers the interference for other users,

and it is likely that some fraction of users will not be transmitting in a busy cell at any given time. Thus CDMA takes advantage of expected silences to allow a larger number of simultaneous calls.

Second, with CDMA each cell uses the same set of frequencies. Unlike GSM and AMPS, FDM is not needed to separate the transmissions of different users. This eliminates complicated frequency planning tasks and improves capacity. It also makes it easy for a base station to use multiple directional antennas, or **sectored antennas**, instead of an omnidirectional antenna. Directional antennas concentrate a signal in the intended direction and reduce the signal (and interference) in other directions. This, in turn, increases capacity. Three-sector designs are common. The base station must track the mobile as it moves from sector to sector. This tracking is easy with CDMA because all frequencies are used in all sectors.

Third, CDMA facilitates **soft handoff**, in which the mobile is acquired by the new base station before the previous one signs off. In this way, there is no loss of continuity. Soft handoff is shown in Fig. 2-43. It is easy with CDMA because all frequencies are used in each cell. The alternative is a **hard handoff**, in which the old base station drops the call before the new one acquires it. If the new one is unable to acquire it (e.g., because there is no available frequency), the call is disconnected abruptly. Users tend to notice this, but it is inevitable occasionally with the current design. Hard handoff is the norm with FDM designs to avoid the cost of having the mobile transmit or receive on two frequencies simultaneously.

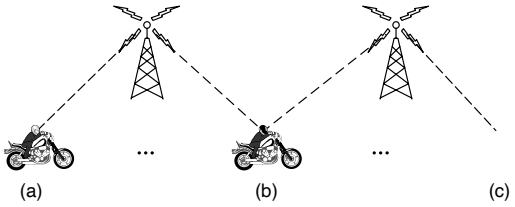


Figure 2-43. Soft handoff (a) before, (b) during, and (c) after.

2.6.6 Fourth-Generation (4G) Technology: Packet Switching

In 2008, the ITU specified a set of standards for 4G systems. **4G**, which is sometimes also called **IMT Advanced** is based completely on packet-switched network technology, including to its predecessors. Its immediate predecessor was a technology often referred to as **LTE** (**Long Term Evolution**). Another precursor and related technology to 4G was 3GPP LTE, sometimes called ''4G LTE.'' The terminology is a bit confusing, as ''4G'' effectively refers to a generation of mobile communications, where any generation may, in fact, have multiple standards. For example, ITU considers IMT Advanced as a 4G standard, although it also accepts LTE as a 4G standard. Other technologies such as the doomed WiMAX (IEEE

802.16) are also considered 4G technologies. Technically, LTE and ''true'' 4G are different releases of the 3GPP standard (releases 8 and 10, respectively).

The main innovation of 4G over previous 3G systems is that 4G networks use packet switching, as opposed to circuit switching. The innovation that allows packet switching is called an **EPC** (**Evolved Packet Core**), which is essentially a simplified IP network that separates voice traffic from the data network. The EPC network carries both voice and data in IP packets. It is thus a (**VoIP**) **Voice over IP** network, with resources allocated using the statistical multiplexing approaches described earlier. As such, the EPC must manage resources in such a way that voice quality remains high in the face of network resources that are shared among many users. The performance requirements for LTE include, among other things, peak throughput of 100 Mbps upload and 50 Mbps download. To achieve these higher rates, 4G networks use a collection of additional frequencies, including 700 MHz, 850 MHz, 800 MHz, and others. Another aspect of the 4G standard is ''spectral efficiency,'' or how many bits can be transmitted per second for a given frequency; for 4G technologies, peak spectral efficiency should be 15 bps/Hz for a downlink and 6.75 bps/Ghz for uplink.

The LTE architecture includes the following elements as part of the Evolved Packet Core, as shown in Chap. 1 as Fig. 1-19.

1. **Serving Gateway** (**S-GW**). The SGW forwards data packets to ensure that packets continue to be forwarded to the user's device when switching from one eNodeB to another.
2. **MME** (**Mobility Management Entity**). The MME tracks and pages the user device and chooses the SGW for a device when it first connects to the network, as well as during handoffs. It also authenticates the user's device.
3. **Packet Data Network Gateway** (**P-GW**). The PDN GW interfaces between the user device and a packet data network (i.e., a packet-switched network), and can perform such functions such as address allocation for that network (e.g., via DHCP), rate limiting, filtering, deep packet inspection, and lawful interception of traffic. User devices establish connection-oriented service with the packet gateway using a so-called **EPS bearer**, which is established when the user device attaches to the network.
4. **HSS** (**Home Subscriber Server**), The MME queries the HSS to determine that the user device corresponds to a valid subscriber.

The 4G network also has an evolved **Radio Access Network** (**RAN**). The radio access network for LTE introduces an access node called an **eNodeB**, which performs operations at the physical layer (as we focus on in this chapter), as well as the **MAC** (**Medium Access Control**), **RLC** (**Radio Link Control**), and **PDCP**

(**Packet Data Control Protocol**) layers, many of which are specific to the cellular network architecture. The eNodeB performs resource management, admission control, scheduling, and other control-plane functions.

On 4G networks, voice traffic can be carried over the EPC using a technology called **VoLTE** (**Voice over LTE**), making it possible for carriers to transmit voice traffic over the packet-switched network and removing any dependency on the legacy circuit-switched voice network.

2.6.7 Fifth-Generation (5G) Technology

Around 2014, the LTE system reached maturity, and people began to start thinking about what would come next. Obviously, after 4G comes 5G. The real question, of course, is ''What Will 5G Be?'' which Andrews et al. (2014) discuss at length. Years later, 5G came to mean many different things, depending on the audience and who is using the term. Essentially, the next generation of mobile cellular network technology boils down to two main factors: higher data rates and lower latency than 4G technologies. There are specific technologies that enable faster speed and lower latency, of course, which we discuss below.

Cellular network performance is often measured in terms of **aggregate data rate** or **area capacity**, which is the total amount of data that the network can serve in bits per unit area. One goal of 5G is to improve the area capacity of the network by three orders of magnitude (more than 1000 times that of 4G), using a combination of technologies:

1. Ultra-densification and offloading. One of the most straightforward ways to improve network capacity is by adding more cells per area. Whereas 1G cell sizes were on the order of hundreds of square kilometers, 5G aims for smaller cell sizes, including **picocells** (cells that are less than 100 meters in diameter) and even **femtocells** (cells that have WiFi-like range of tens of meters). One of the most important benefits of the shrinking of the cell size is the ability to reuse spectrum in a given geographic area, thus reducing the number of users that are competing for resources at any given base station. Of course, shrinking the cell size comes with its own set of complications, including more complicated mobility management and handoff.
2. Increased bandwidth with millimeter waves. Most spectrum from previous technologies has been in the range of several hundred MHz to a few GHz, corresponding to wavelengths that are in range of centimeters to about a meter. This spectrum has become increasingly crowded, especially in major markets during peak hours. There are considerable amounts of unused spectrum in the millimeter wave range of 20–300 GHz, with wavelengths of less than 10 millimeters. Until recently, this spectrum was not considered suitable for wireless

communication because shorter wavelengths do not propagate as well. One of the ways that propagation challenges are being tackled is by using large arrays of directional antennas, which is a significant architectural shift from previous generations of cellular networks: everything from interference properties to the process of associating a user to a base station is different.

3. Increased spectral efficiency through advances in massive **MIMO** (**Multiple-Input Multiple-Output**) technology. MIMO improves the capacity of a radio link by using multiple transmit and receive antennas to take advantage of multipath propagation, whereby the transmitted radio signal reaches the receiver via two or more paths. MIMO was introduced into WiFi communication and 3G cellular technologies around 2006. MIMO has quite a few variations; earlier cellular standards take advantage of **MU-MIMO** (**Multi-User MIMO**). Generally, these technologies take advantage of the spatial diversity of users to cancel out interference that may occur at either end of the wireless transmission. **Massive MIMO** is a type of MU-MIMO that increases the number of base station antennas so that there are many more antennas than endpoints. There is even the possibility of using a three-dimensional antenna array, in a so-called **FD-MIMO** (**Full-Dimension MIMO**).

Another capability that will accompany 5G is **network slicing**, which will let cellular carriers create multiple virtual networks on top of the same shared physical infrastructure, devoting portions of their network to specific customer use cases. Distinct fractions of the network (and its resources) may be dedicated to different application providers, where different applications may have different requirements. For example, applications that require high throughput may be allocated to a different network slice than those that do not require high throughput. **SDN** (**Software-Defined Networking**) and **NFV** (**Network Functions Virtualization**) are emerging technologies that will help support slicing. We will discuss these technologies in later chapters.

2.7 CABLE NETWORKS

The fixed and wireless phone systems will clearly play a role in future networks, but the cable networks will also factor heavily into future broadband access networks. Many people nowadays get their television, telephone, and Internet service over cable. In the following sections, we will look at cable television as a network in more detail, contrasting it with the telephone systems we have just studied. For more information see Harte (2017). The 2018 DOCSIS standard also provides helpful information, particularly related to modern cable network architectures.

2.7.1 A History of Cable Networks: Community Antenna Television

Cable television was conceived in the late 1940s as a way to provide better television reception to people living in rural or mountainous areas. The system initially consisted of a big antenna on top of a hill to pluck the television signal out of the air, an amplifier, called the **headend**, to strengthen it, and a coaxial cable to deliver it to people's houses, as illustrated in Fig. 2-44.

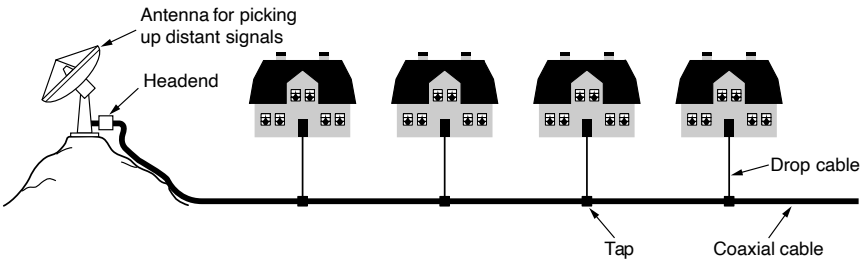


Figure 2-44. An early cable television system.

In the early years, cable television was called **CATV** (**Community Antenna Television**). It was very much a mom-and-pop operation; anyone handy with electronics could set up a service for his town, and the users would chip in to pay the costs. As the number of subscribers grew, additional cables were spliced onto the original cable and amplifiers were added as needed. Transmission was one way, from the headend to the users. By 1970, thousands of independent systems existed.

In 1974, Time Inc. started a new channel, Home Box Office, with new content (movies) distributed only on cable. Other cable-only channels followed, focusing on news, sports, cooking, history, movies, science, kids, and many other topics. This development gave rise to two changes in the industry. First, large corporations began buying up existing cable systems and laying new cable to acquire new subscribers. Second, there was now a need to connect multiple systems, often in distant cities, in order to distribute the new cable channels. The cable companies began to lay cable between the cities to connect them all into a single system. This pattern was analogous to what happened in the telephone industry 80 years earlier with the connection of previously isolated end offices to make long-distance calling possible.

2.7.2 Broadband Internet Access Over Cable: HFC Networks

Over the course of the years the cable system grew and the cables between the various cities were replaced by high-bandwidth fiber, similar to what happened in the telephone system. A system with fiber for the long-haul runs and coaxial cable

to the houses is called an **HFC** (**Hybrid Fiber Coax**) system and is the predominant architecture for today's cable networks. The trend of moving fiber closer to the subscriber home continues, as described in the earlier section on FTTX. The electro-optical converters that interface between the optical and electrical parts of the network are called **fiber nodes**. Because the bandwidth of fiber is so much greater than that of coax, a single fiber node can feed multiple coaxial cables. Part of a modern HFC system is shown in Fig. 2-45(a).

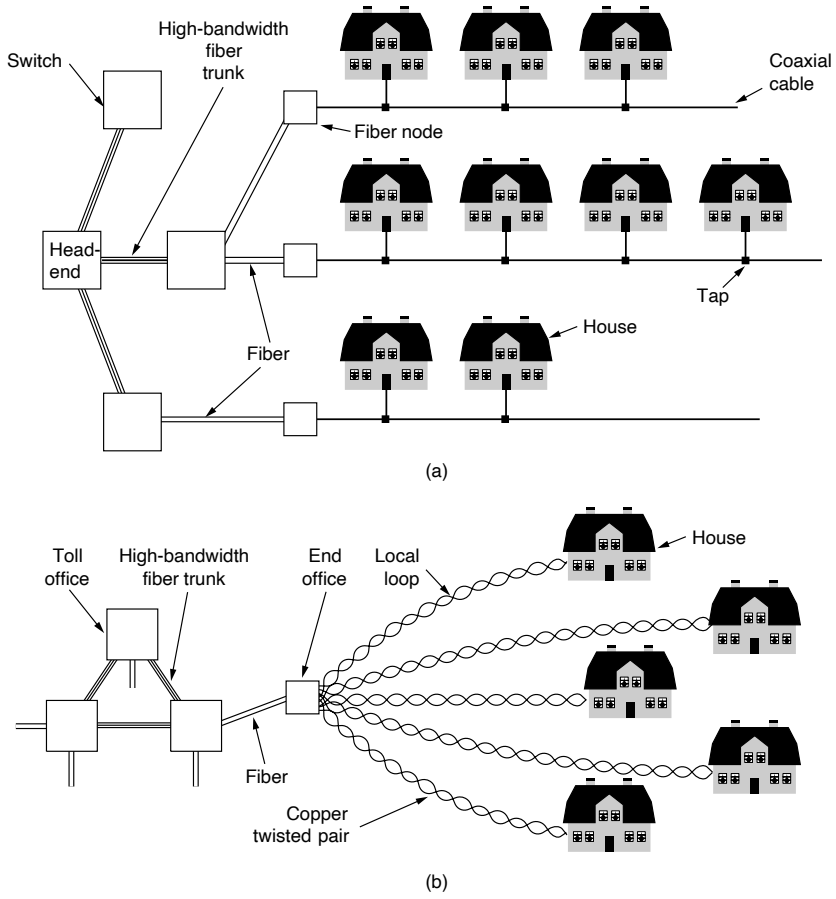


Figure 2-45. (a) Hybrid Fiber-Coax cable network. (b) The fixed phone system.

In the late 1990s, many cable operators began to enter the Internet access business as well as the telephony business. Technical differences between the cable

plant and telephone plant had an effect on what had to be done to achieve these goals. For one thing, all the one-way amplifiers in the system had to be replaced by two-way amplifiers to support upstream as well as downstream transmissions. While this was happening, early Internet over cable systems used the cable television network for downstream transmissions and a dial-up connection via the telephone network for upstream transmissions. It was a kludge if ever there was one, but it sort of worked.

Throwing off all the TV channels and using the cable infrastructure strictly for Internet access would probably generate a fair number of irate customers (mostly older customers, since many younger ones have already cut the cord), so cable companies are hesitant to do this. Furthermore, most cities heavily regulate what is on the cable, so the cable operators would not be allowed to do this even if they really wanted to. As a consequence, they needed to find a way to have television and Internet peacefully coexist on the same cable.

The solution is to build on frequency division multiplexing. Cable television channels in North America occupy the 54–550 MHz region (except for FM radio, from 88 to 108 MHz). These channels are 6-MHz wide, including guard bands, and can carry one traditional analog television channel or several digital television channels. In Europe, the low end is usually around 65 MHz and the channels are 6–8 MHz wide for the higher resolution required by PAL and SECAM, but otherwise the allocation scheme is similar. The low part of the band is not used. Modern cables can also operate well above 550 MHz, often at up to 750 MHz or more. The solution chosen was to introduce upstream channels in the 5–42-MHz band (slightly higher in Europe) and use the frequencies at the high end for the downstream signals. The cable spectrum is illustrated in Fig. 2-46.

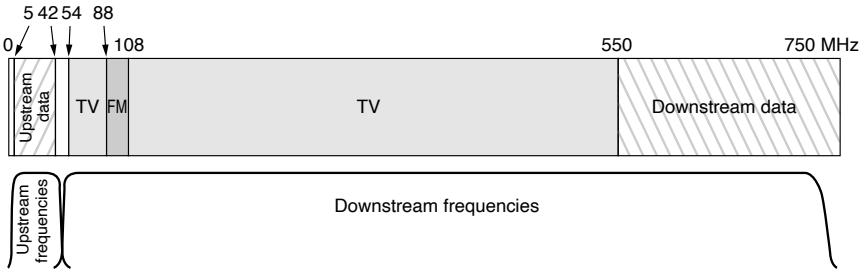


Figure 2-46. Frequency allocation in a typical cable TV system used for Internet access.

Because the television signals are all downstream, it is possible to use upstream amplifiers that work only in the 5–42-MHz region and downstream amplifiers that work only at 54 MHz and up, as shown in the figure. Thus, we get an asymmetry in the upstream and downstream bandwidths because more spectrum

is available above television than below it. On the other hand, most users want more downstream traffic, so cable operators are not unhappy with this fact of life. As we saw earlier, telephone companies usually offer an asymmetric DSL service, even though they have no technical reason for doing so. In addition to upgrading the amplifiers, the operator has to upgrade the headend, too, from a dumb amplifier to an intelligent digital computer system with a high-bandwidth fiber interface to an ISP. This upgraded headend is now sometimes called a **CMTS** (**Cable Modem Termination System**). The CMTS and headend refer to the same component.

2.7.3 DOCSIS

Cable companies operate networks that include HFC physical-layer technology for last-mile connectivity, as well as fiber and wireless last-mile connections. The HFC part of those networks is widely deployed across the United States, Canada, Europe, and other markets, and use the CableLabs **DOCSIS** (**Data Over Cable Service Interface Specification**) standards.

DOCSIS version 1.0 was released in 1997. DOCSIS 1.0 and 1.1 had a working limit of 38 Mbps downstream and 9 Mbps upstream DOCSIS 2.0 in 2001 resulted in a tripling of upstream bandwidth. Later, DOCSIS 3.0 (2006) introduced support for IPv6 and enabled channel bonding for downstream and upstream communications, dramatically increasing the potential capacity for each home served to hundreds of megabits per second. DOCSIS 3.1 (2013), which introduced Orthogonal Frequency Division Multiplexing (OFDM), wider channel bandwidth and higher efficiency, enabled over 1 Gbps of downstream capacity per home. Extensions to DOCSIS 3.1 have been added via updates to the DOCSIS 3.1 standard, including Full Duplex operation (2017), which will enable multigigabit symmetric downstream and upstream capacity, as well as DOCSIS Low Latency (2018) and other features to reduce latency.

At the hybrid fiber coaxial (HFC) layer, the network is highly dynamic, with cable network operators performing fiber node splits on a regular basis, which pushes fiber closer to the home and reduces the number of homes served by each node, thereby making more capacity available for each home served. In some cases the HFC last mile is replaced with fiber to the home, and many new builds are fiber to the home as well.

Cable Internet subscribers require a DOCSIS cable modem to serve as the interface between the home network and the ISP network. Each cable modem sends data on one upstream and one downstream channel. Each channel is allocated using FDM. DOCSIS 3.0 uses multiple channels. The usual scheme is to take each 6 or 8 MHz downstream channel and modulate it with QAM-64 or, if the cable quality is exceptionally good, QAM-256; a 6-MHz channel and QAM-64 yields about 36 Mbps. Accounting for signaling overhead, the net bandwidth is about 27 Mbps. With QAM-256, the net payload is about 39 Mbps. The European values are 1/3 larger due to the larger bandwidth available.

The modem-to-home network interface is straightforward: it is typically an Ethernet connection. These days, many home Internet users connect the cable modem to a WiFi access point to set up a home wireless network. In some cases, the user's Internet service provider (ISP) provides a single hardware device that combines the cable modem and wireless access point. The interface between the cable modem and the rest of the ISP network is more complicated, as it involves coordinating resource sharing among many cable subscribers who may be connected to the same headend. This resource sharing technically occurs at the link layer, not the physical layer, although we will cover it in this chapter for the sake of continuity.

2.7.4 Resource Sharing in DOCSIS Networks: Nodes and Minislots

There is one important fundamental difference between the HFC system of Fig. 2-45(a) and the telephone system of Fig. 2-45(b). In a given residential neighborhood, a single cable is shared by many houses, whereas in the telephone system, every house has its own private local loop. When these cables are used for television broadcasting, sharing is natural. All the programs are broadcast on the cable and it does not matter whether there are 10 viewers or 10,000 viewers. When the same cable is used for Internet access, however, it matters a lot if there are 10 users or 10,000. If one user decides to download a very large file or stream an 8K movie, that bandwidth is not available to other users. More users sharing a single cable creates more competition for the bandwidth of the cable. The telephone system does not have this particular property: downloading a large file over an ADSL line does not reduce your neighbor's bandwidth. On the other hand, the bandwidth of coax is much higher than that of twisted pairs. In essence, the bandwidth that a given subscriber receives at any given moment depends quite a bit on the usage of subscribers who happen to be sharing the same cable, as we describe in more detail below.

Cable ISPs have tackled this problem by splitting up long cables and connecting each one directly to a fiber node. The bandwidth from the headend to each fiber node is significant, so as long as there are not too many subscribers on each cable segment, the amount of traffic is manageable. A typical node size about ten or fifteen years ago was 500–2000 homes, although the number of homes per node continues to decrease as buildout to the edge continues in an effort to increase speeds to subscribers. Increases in cable Internet subscribers over the past decade, coupled with increasing traffic demand from subscribers, has created the need to increasingly split these cables and add more fiber nodes. By 2019, a typical node size was about 300–500 homes, although in some areas, ISPs are building N+0 HFC (a.k.a. ''Fiber Deep'') architectures, which can reduce this number to as low as 70, which eliminates the need for cascading signal amplifiers and runs fiber direct from network headends to nodes at the last segment of coaxial cable.

When a cable modem is plugged in and powered up, it scans the downstream channels looking for a special packet that the headend periodically sends, providing system parameters to modems that have just come online. Upon receiving this packet, the new modem announces its presence on one of the upstream channels. The headend responds by assigning the modem an upstream and a downstream channel. These assignments can be changed later if the headend deems it necessary to balance the load.

There is more RF noise in the upstream direction because the system was not originally designed for data, and noise from multiple subscribers is funneled to the headend, so the modem transmits using a more conservative approach. This ranges from QPSK to QAM-128, where some of the symbols are used for error protection with trellis coded modulation. With fewer bits per symbol on the upstream, the asymmetry between upstream and downstream rates is much more than suggested by Fig. 2-46.

Today's DOCSIS modems request a time to transmit, and then the CMTS grants one or more timeslots that the modem can transmit, based on availability; simultaneous users all contend for upstream and downstream access. The network uses TDM to share upstream bandwidth across multiple subscribers. Time is divided into **minislots**; each subscriber sends in a different minislot. The headend announces the start of a new round of minislots periodically, but the announcement for the start of each minislot is not heard at all modems simultaneously due to signal propagation time down the cable. By knowing how far it is from the headend, each modem can compute how long ago the first minislot really started.

It is important for the modem to know its distance to the headend to get the timing right. The modem first determines its distance from the headend by sending it a special packet and seeing how long it takes to get the response. This process is called **ranging**. Each upstream packet must fit in one or more consecutive minislots at the headend when it is received. Minislot length is network dependent. A typical payload is 8 bytes.

During initialization, the headend assigns each modem to a minislot to use for requesting upstream bandwidth. When a computer wants to send a packet, it transfers the packet to the modem, which then requests the necessary number of minislots for it. If the request is accepted, the headend puts an acknowledgement on the downstream channel telling the modem which minislots have been reserved for its packet. The packet is then sent, starting in the minislot allocated to it. Additional packets can be requested using a field in the header.

As a rule, multiple modems will be assigned the same minislot, which leads to contention (multiple modems attempting to send upstream data at the same time). CDMA can allow multiple subscribers to share the same minislot, although it reduces the rate per subscriber. Another alternative is to not use CDMA, in which case there may be no acknowledgement to the request because of a collision. When collisions occur in this case, the modem just waits a random time and tries again. After each successive failure, the randomization time is doubled. (For readers

already somewhat familiar with networking, this algorithm is just slotted ALOHA with binary exponential backoff. Ethernet cannot be used on cable because stations cannot sense the medium. We will come back to these issues in Chap. 4.)

The downstream channels are managed differently from the upstream channels. For starters, there is only one sender (the headend), so there is no contention and no need for minislots. For another, the amount of traffic downstream is usually much larger than upstream, so a fixed packet size of 204 bytes is used. Part of that is a Reed-Solomon error-correcting code and some other overhead, leaving a user payload of 184 bytes. These numbers were chosen for compatibility with digital television using MPEG-2, so the TV and downstream data channels are formatted the same way. Logically, the connections are as depicted in Fig. 2-47.

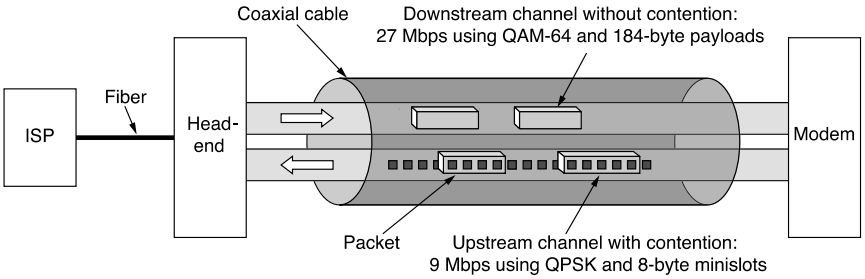


Figure 2-47. Typical details of the upstream and downstream channels in North America.

2.8 COMMUNICATION SATELLITES

In the 1950s and early 1960s, people tried to set up communication systems by bouncing signals off metallized weather balloons. Unfortunately, the received signals were too weak to be of any practical use. Then, the U.S. Navy noticed a kind of permanent weather balloon in the sky—the moon—and built an operational system for ship-to-shore communication by bouncing signals off it.

Further progress in the celestial communication field had to wait until the first communication satellite was launched. The key difference between an artificial satellite and a real one is that the artificial one can amplify the signals before sending them back, turning a strange curiosity into a powerful communication system.

Communication satellites have some interesting properties that make them attractive for many applications. In its simplest form, a communication satellite can be thought of as a big microwave repeater in the sky. It contains several **transponders**, each of which listens to some portion of the spectrum, amplifies the

incoming signal, and then rebroadcasts it at another frequency to avoid interference with the incoming signal. This mode of operation is known as a **bent pipe**. Digital processing can be added to separately manipulate or redirect data streams in the overall band, or digital information can even be received by the satellite and rebroadcast. Regenerating signals in this way improves performance compared to a bent pipe because the satellite does not amplify noise in the upward signal. The downward beams can be broad, covering a substantial fraction of the earth's surface, or narrow, covering an area only hundreds of kilometers in diameter.

According to Kepler's law, the orbital period of a satellite varies as the radius of the orbit to the 3/2 power. The higher the satellite, the longer the period. Near the surface of the earth, the period is about 90 minutes. Consequently, low-orbit satellites pass out of view fairly quickly (due to the satellites' motion), so many of them are needed to provide continuous coverage and ground antennas must track them. At an altitude of about 35,800 km, the period is 24 hours. At an altitude of 384,000 km, the period is about 1 month, as anyone who has observed the moon regularly can testify.

A satellite's period is important, but it is not the only issue in determining where to place it. Another issue is the presence of the Van Allen belts, layers of highly charged particles trapped by the earth's magnetic field. Any satellite flying within them would be destroyed fairly quickly by the particles. These factors lead to three regions in which satellites can be placed safely. These regions and some of their properties are illustrated in Fig. 2-48. Below, we will briefly describe the satellites that inhabit each of these regions.

2.8.1 Geostationary Satellites

In 1945, the science fiction writer Arthur C. Clarke calculated that a satellite at an altitude of 35,800 km in a circular equatorial orbit would appear to remain motionless in the sky, so it would not need to be tracked (Clarke, 1945). He went on to describe a complete communication system that used these (manned) **geostationary satellites**, including the orbits, solar panels, radio frequencies, and launch procedures. Unfortunately, he concluded that satellites were impractical due to the impossibility of putting power-hungry, fragile vacuum tube amplifiers into orbit, so he never pursued this idea further, although he wrote some science fiction stories about it.

The invention of the transistor changed all that, and the first artificial communication satellite, Telstar, was launched in July 1962. Since then, communication satellites have become a multibillion dollar business and the only aspect of outer space that has become highly profitable. These high-flying satellites are often called **GEO** (**Geostationary Earth Orbit**) satellites.

With current technology, it is technologically unwise to have geostationary satellites spaced much closer than 2 degrees in the 360-degree equatorial plane, to

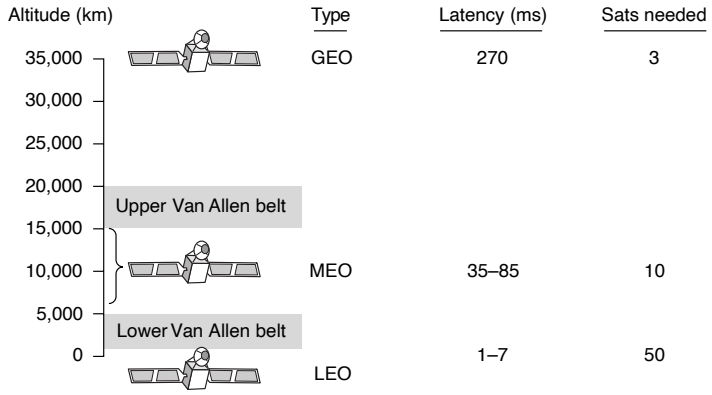


Figure 2-48. Communication satellites and some of their properties, including altitude above the earth, round-trip delay time, and number of satellites needed for global coverage.

avoid interference. With a spacing of 2 degrees, there can only be 360/2 = 180 of these satellites in the sky at once. However, each transponder can use multiple frequencies and polarizations to increase the available bandwidth.

To prevent total chaos in the sky, orbit slot allocation is done by ITU. This process is highly political, with countries barely out of the stone age demanding ''their'' orbit slots (for the purpose of leasing them to the highest bidder). Other countries, however, maintain that national property rights do not extend up to the moon and that no country has a legal right to the orbit slots above its territory. To add to the fight, commercial telecommunication is not the only application. Television broadcasters, governments, and the military also want a piece of the orbiting pie.

Modern satellites can be quite large, weighing over 5000 kg and consuming several kilowatts of electric power produced by the solar panels. The effects of solar, lunar, and planetary gravity tend to move them away from their assigned orbit slots and orientations, an effect countered by on-board rocket motors. This fine-tuning activity is called **station keeping**. However, when the fuel for the motors has been exhausted (typically after about 10 years), the satellite drifts and tumbles helplessly, so it has to be turned off. Eventually, the orbit decays and the satellite reenters the atmosphere and burns up or (very rarely) crashes to earth.

Orbit slots are not the only bone of contention. Frequencies are an issue, too, because the downlink transmissions interfere with existing microwave users. Consequently, ITU has allocated certain frequency bands to satellite users. The main ones are listed in Fig. 2-49. The C band was the first to be made available for commercial satellite traffic. Two frequency ranges are assigned in it, the lower one for

downlink traffic (from the satellite) and the upper one for uplink traffic (to the satellite). To allow traffic to go both ways at the same time, two channels are required. These channels are already overcrowded because they are also used by the common carriers for terrestrial microwave links. The L and S bands were added by international agreement in 2000. However, they are narrow and also crowded.

Band	Downlink	Uplink	Bandwidth	Problems
L	1.5 GHz	1.6 GHz	15 MHz	Low bandwidth; crowded
S	1.9 GHz	2.2 GHz	70 MHz	Low bandwidth; crowded
C	4.0 GHz	6.0 GHz	500 MHz	Terrestrial interference
Ku	11 GHz	14 GHz	500 MHz	Rain
Ka	20 GHz	30 GHz	3500 MHz	Rain, equipment cost

Figure 2-49. The principal satellite bands.

The next-highest band available to commercial telecommunication carriers is the Ku (K under) band. This band is not (yet) congested, and at its higher frequencies, satellites can be spaced as close as 1 degree; transmission speeds in this band can reach more than 500 Mbps. However, another problem exists: rain. Water absorbs these short microwaves well. Fortunately, heavy storms are usually localized, so using several widely separated ground stations instead of just one circumvents the problem, but at the price of extra antennas, extra cables, and extra electronics to enable rapid switching between stations. Bandwidth has also been allocated in the Ka (K above) band for commercial satellite traffic, but the equipment needed to use it is expensive. In addition to these commercial bands, many government and military bands also exist.

A modern satellite has around 40 transponders, most often with a 36-MHz bandwidth. Usually, each transponder operates as a bent pipe, but recent satellites have some on-board processing capacity, allowing more sophisticated operation. In the earliest satellites, the division of the transponders into channels was static: the bandwidth was simply split up into fixed frequency bands. Nowadays, each transponder beam is divided into time slots, with various users taking turns. Once again, we see how TDM and FDM are used in many contexts.

The first geostationary satellites had a single spatial beam that illuminated about 1/3 of the earth's surface, called its **footprint**. With the enormous decline in the price, size, and power requirements of microelectronics, a much more sophisticated broadcasting strategy has become possible. Each satellite is equipped with multiple antennas and multiple transponders. Each downward beam can be focused on a small geographical area, so multiple upward and downward transmissions can take place simultaneously. Typically, these so-called **spot beams** are elliptically shaped, and can be as small as a few hundred km in diameter. A communication satellite for the United States typically has one wide beam for the contiguous 48 states, plus spot beams for Alaska and Hawaii.

One important development in the communication satellite world are low-cost microstations, sometimes called **VSATs** (**Very Small Aperture Terminals**) (Abramson, 2000). These tiny terminals have 1-meter or smaller antennas (versus 10 m for a standard GEO antenna) and can put out about 1 watt of power. The uplink is generally good for up to 1 Mbps, but the downlink is often up to several megabits/sec. Direct broadcast satellite television uses this technology for one-way transmission.

In many VSAT systems, the microstations do not have enough power to communicate directly with one another (via the satellite, of course). Instead, a special ground station, the **hub**, with a large, high-gain antenna is needed to relay traffic between VSATs, as shown in Fig. 2-50. In this mode of operation, either the sender or the receiver has a large antenna and a powerful amplifier. The trade-off is a longer delay in return for having cheaper end-user stations.

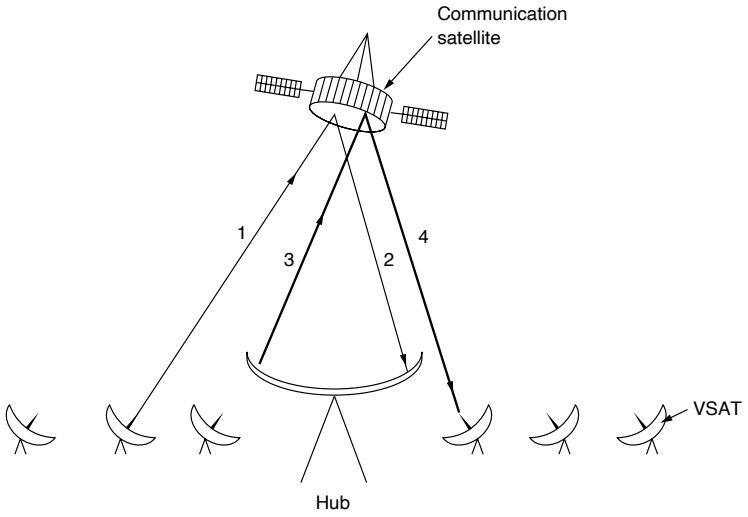


Figure 2-50. VSATs using a hub.

VSATs have great potential in rural areas, especially in developing countries. In much of the world, there are no landlines or cell towers. Stringing telephone wires to thousands of small villages is far beyond the budgets of most developing-country governments. Erecting cell towers is easier, but the cell towers need wired connections to the national telephone network. However, installing 1-meter VSAT dishes powered by solar cells is often feasible. VSATs provide the technology that can finish wiring the world. They can also provide Internet access to smartphone users in areas where there is no terrestrial infrastructure, which is true in much of the developing world.

Communication satellites have several properties that are radically different from terrestrial point-to-point links. To begin with, even though signals to and from a satellite travel at the speed of light (nearly 300,000 km/sec), the long round-trip distance introduces a substantial delay for GEO satellites. Depending on the distance between the user and the ground station and the elevation of the satellite above the horizon, the end-to-end latency is between 250 and 300 msec. A typical roundtrip value is 270 msec (540 msec for a VSAT system with a hub).

For comparison purposes, terrestrial microwave links have a propagation delay of roughly 3 μsec/km, and coaxial cable or fiber-optic links have a delay of approximately 5 μsec/km. The latter are slower than the former because electromagnetic signals travel faster in air than in solid materials.

Another important property of satellites is that they are inherently broadcast media. It does not cost any more to send a message to thousands of stations within a transponder's footprint than it does to send to only one. For some applications, this property is very useful. For example, one could imagine a satellite broadcasting popular Web pages to the caches of a large number of computers spread over a wide area. Even when broadcasting can be simulated with point-to-point lines, satellite broadcasting may be much cheaper. On the other hand, from a privacy point of view, satellites are a complete disaster: everybody can hear everything. Encryption is essential for confidentiality.

Satellites also have the property that the cost of transmitting a message is independent of the distance traversed. A call across the ocean costs no more to service than a call across the street. Satellites also have excellent error rates and can be deployed almost instantly, a major consideration for disaster response and military communication.

2.8.2 Medium-Earth Orbit Satellites

At much lower altitudes, between the two Van Allen belts, we find the **MEO** (**Medium-Earth Orbit**) satellites. As viewed from the earth, these drift slowly in longitude, taking something like 6 hours to circle the earth. Accordingly, they must be tracked as they move through the sky. Because they are lower than the GEOs, they have a smaller footprint on the ground and require less powerful transmitters to reach them. Currently, they are used for navigation systems rather than telecommunications, so we will not examine them further here. The constellation of roughly 30 **GPS** (**Global Positioning System**) satellites orbiting at about 20,200 km are examples of MEO satellites.

2.8.3 Low-Earth Orbit Satellites

Moving down in altitude, we come to the **LEO** (**Low-Earth Orbit**) satellites. Due to their rapid motion, large numbers of them are needed for a complete system. On the other hand, because the satellites are so close to the earth, the ground

stations do not need much power, and the round-trip delay is much less: deployments see round-trip latencies of anywhere between around 40 and 150 milliseconds. The launch cost is substantially cheaper too. In this section, we will examine two examples of satellite constellations used for voice service: Iridium and Globalstar.

For the first 30 years of the satellite era, low-orbit satellites were rarely used because they zip into and out of view so quickly. In 1990, Motorola broke new ground by filing an application with the FCC asking for permission to launch 77 low-orbit satellites for the **Iridium** project (element 77 is iridium). The plan was later revised to use only 66 satellites, so the project should have been renamed Dysprosium (element 66), but that probably sounded too much like a disease. The idea was that as soon as one satellite went out of view, another would replace it. This proposal set off a feeding frenzy among other communication companies. All of a sudden, everyone wanted to launch a chain of low-orbit satellites.

After seven years of cobbling together partners and financing, communication service began in November 1998. Unfortunately, the commercial demand for large, heavy satellite telephones was negligible because the mobile phone network had grown in a spectacular way since 1990. As a consequence, Iridium was not profitable and was forced into bankruptcy in August 1999 in one of the most spectacular corporate fiascos in history. The satellites and other assets (worth $5 billion) were later purchased by an investor for $25 million at a kind of extraterrestrial garage sale. Other satellite business ventures promptly followed suit.

The Iridium service restarted in March 2001 and has been growing ever since. It provides voice, data, paging, fax, and navigation service everywhere on land, air, and sea, via hand-held devices that communicate directly with the Iridium satellites. Customers include the maritime, aviation, and oil exploration industries, as well as people traveling in parts of the world lacking a telecom infrastructure (e.g., deserts, mountains, the South Pole, and some developing countries).

The Iridium satellites are positioned at an altitude of 670 km, in circular polar orbits. They are arranged in north-south necklaces, with one satellite every 32 degrees of latitude, as shown in Fig. 2-51. Each satellite has a maximum of 48 cells (spot beams) and a capacity of 3840 channels, some of which are used for paging and navigation, while others are used for data and voice.

With six satellite necklaces, the entire earth is covered, as suggested by Fig. 2-51. An interesting property of Iridium is that communication between distant customers takes place in space, as shown in Fig. 2-52(a). Here we see a caller at the North Pole contacting a satellite directly overhead. Each satellite has four neighbors with which it can communicate, two in the same necklace (shown) and two in adjacent necklaces (not shown). The satellites relay the call across this grid until it is finally sent down to the callee at the South Pole.

An alternative design to Iridium is **Globalstar**. It is based on 48 LEO satellites but uses a different switching scheme than the one used by Iridium. Whereas Iridium relays calls from satellite to satellite, which requires complex switching

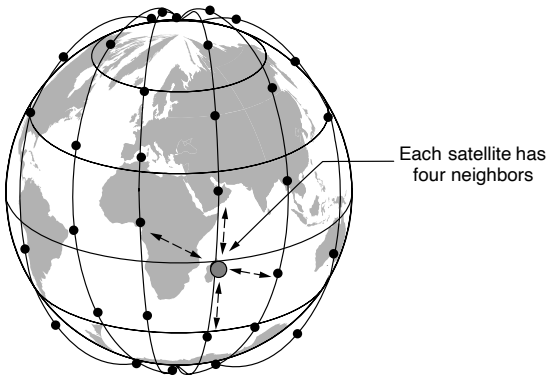


Figure 2-51. The Iridium satellites form six necklaces around the earth.

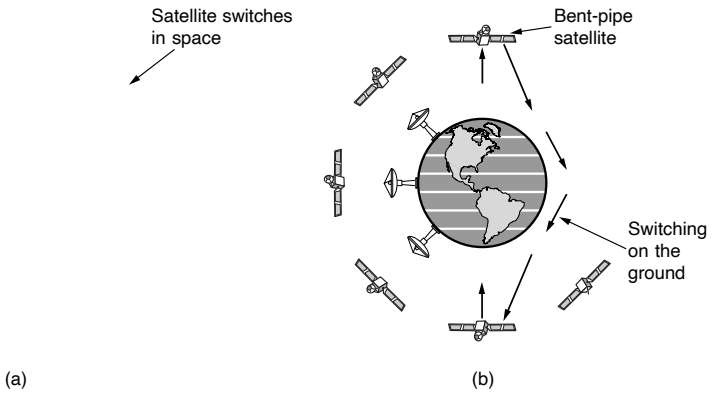


Figure 2-52. (a) Relaying in space. (b) Relaying on the ground.

equipment in the satellites, Globalstar uses a traditional bent-pipe design. The call originating at the North Pole in Fig. 2-52(b) is sent back to earth and picked up by the large ground station at Santa's Workshop. The call is then routed via a terrestrial network to the ground station nearest the callee and delivered by a bent-pipe connection as shown. The advantage of this scheme is that it puts much of the complexity on the ground, where it is much easier to manage. Also, the use of large ground station antennas that can put out a powerful signal and receive a weak one means that lower-powered telephones can be used. After all, the telephone puts out only a few milliwatts of power, so the signal that gets back to the ground station is fairly weak, even after having been amplified by the satellite.

Satellites continue to be launched at a rate of around 20 satellites per year, including ever-larger satellites that now weigh over 5000 kilograms. But there are also very small satellites for the more budget-conscious organization. To make space research more accessible, academic researchers from California Polytechnic University and Stanford got together in 1999 to define a standard for miniature satellites and an associated launcher that would greatly lower launch costs (Nugent et al., 2008). These **cubesats** are satellites in units of 10 cm × 10 cm × 10 cm cubes, each weighing no more than 1 kilogram, that can be launched for a price as little as $40,000 each. The launcher flies as a secondary payload on commercial space missions. It is basically a tube that takes up to three units of cubesats and uses springs to release them into orbit. Roughly 20 cubesats have launched so far, with many more in the works. Most of them communicate with ground stations on the UHF and VHF bands.

Another deployment of LEO satellites is an attempted satellite-based Internet backbone network, OneWeb's deployment will initially involve a constellation of several hundred satellites. If successful, the project promises to bring high-speed Internet access to places which may not currently have it. The satellites will operate in the Ku band and will use a technique called ''progressive pitch,'' whereby the satellites are turned slightly to avoid interference with geostationary satellites that are transmitting in the same band.

2.9 COMPARING DIFFERENT ACCESS NETWORKS

Let's now compare the properties of the different types of access networks that we have surveyed.

2.9.1 Terrestrial Access Networks: Cable, Fiber, and ADSL

Cable, FTTH, and ADSL are much more similar than they are different. They offer comparable service and, as competition between them heats up, probably comparable prices. All access network technologies, including cable, ADSL, and Fiber to the Home, now use fiber in the backbone; they differ on the last-mile access technology at the physical and link layers. Fiber and ADSL providers tend to deliver more consistent bandwidth to each subscriber because each user has dedicated capacity. Ongoing and recent reports in the United States, such as the FCC's Measuring Broadband America (MBA) initiative (which is released annually), report that access ISPs typically meet their advertised rates.

As an ADSL or FTTH access network acquires more users, their increasing numbers have little effect on existing users, since each user has a dedicated connection all the way to the home. On the other hand, cable subscribers share the capacity of a single node; as a result, when one or more users on a node increase their usage, other users may experience congestion. Consequently, cable providers now

tend to over-provision the capacity that they sell to each subscriber. More modern DOCSIS standards such as DOCSIS 3.0 require that cable modems be capable of bonding at least four channels, to achieve approximately 170 Mbps downstream and 120 Mbps upstream (with about 10% of that throughput dedicated to signaling overhead).

Ultimately, the maximum speeds that a cable subscriber can achieve are limited by the capacity of the coaxial cable, the amount of usable spectrum in fiber is far greater by comparison. With cable, as more subscribers sign up for Internet service, the performance of other users in the same node will suffer. In response, cable ISPs split busy cables, connecting each one to a fiber node directly (this practice is sometimes called a **node split**. As previously discussed, the number of homes per node continues to steadily decrease as cable ISPs continue to build fiber closer to the edge of the network.

Cable, fiber, and ADSL are available in different regions, and performance of these networks differs according to both the technology itself, and how each respective technology is deployed. Most home users in developed countries can have a telephone line if they want it, but not all users are close enough to their end offices to get ADSL. Some are stuck with 56-kbps dial-up lines, especially in rural areas. In fact, even in the United States, there are large areas in which a 1.544-Mbps T1 line is an unobtainable luxury. In Europe, with its higher population density, 500 Mbps fiber-optic Internet is common in big cities. Some even have 1-Gbps service available.

Also, not everyone has cable. If you do have cable and the company provides Internet access, you can get it; distance to the fiber node or headend is not an issue. Availability of cable and fiber in certain regions, particularly sparsely populated regions, remains a concern though. Ultimately, high-speed Internet access today still depends on the deployment of fiber or cable to homes. In the case of cable networks, increasing node splits require the deployment of fiber further into the neighborhood, as opposed to relying on existing coaxial cable infrastructure. Even in the case of ADSL, speed drops off significantly beyond a few kilometers from a central office, so even ADSL requires some kind of fiber buildout at the edge (e.g., FTTN) to offer high speed to sparsely populated areas. All of these are expensive propositions.

Historically, the telephone infrastructure (and DSL networks) have generally been more reliable than cable, although data from the FCC's MBA project show that gap has narrowed, with most cable and DSL service achieving at least ''two nines'' of reliability (i.e., 99% uptime, or tens of hours of downtime a year). Satellite and metropolitan-area wireless networks perform less reliably. By comparison, the conventional phone network achieves ''five nines'' of reliability, which corresponds to only a few minutes of unavailability each year (Bischof et al., 2018).

Being a point-to-point medium, ADSL is inherently more secure than cable. Any cable user can easily read all the packets going down the cable, no matter for whom they are intended. For this reason, any decent cable provider will encrypt all

traffic in both directions. Nevertheless, having your neighbor get your encrypted messages is still less secure than having him not get anything at all.

2.9.2 Satellites Versus Terrestrial Networks

A comparison between satellite and terrestrial communication networks is instructive. Some time ago, it seemed that communication satellites might have been the future of communication. After all, the telephone system had changed little in the previous 100 years and showed no signs of changing in the next 100 years. This glacial movement was caused in no small part by the regulatory environment in which the telephone companies were expected to provide good voice service at reasonable prices (which they did), and in return got a guaranteed profit on their investment. For people with data to transmit, 1200-bps modems were available. That was pretty much all there was.

The introduction of competition in telecommunications in 1984 in the United States and somewhat later in Europe radically changed this situation. Telephone companies began replacing their long-haul networks with fiber and introduced high-bandwidth services like ADSL. They also stopped their long-time practice of charging artificially high prices to long-distance users to subsidize local service. All of a sudden, terrestrial fiber looked like the winner.

Nevertheless, communication satellites have some niche markets that fiber cannot address. First, when rapid deployment is critical, satellites win easily. A quick response is useful for military communication systems in times of war and disaster response in times of peace. Following the massive December 2004 Sumatra earthquake and subsequent tsunami, for example, communications satellites were able to restore communications to first responders within 24 hours. This rapid response was possible because there is a developed market in which large players, such as Intelsat with over 50 satellites, can rent out capacity pretty much anywhere it is needed. For customers served by existing satellite networks, a solar-powered VSAT can be set up easily and quickly to provide a megabit/sec link.

A second niche is for communication in places where the terrestrial infrastructure is poorly developed. Many people nowadays want to communicate everywhere they go. Mobile phone networks cover those locations with good population density, but do not do an adequate job in other places (e.g., at sea or in the desert). Conversely, Iridium provides voice service everywhere on earth, even at the South Pole. Terrestrial infrastructure can also be expensive to install, depending on the terrain and necessary rights of way. Indonesia, for example, has its own satellite for domestic telephone traffic. Launching one satellite was cheaper than stringing thousands of undersea cables among the 13,677 islands in the archipelago.

A third niche is when broadcasting is essential. A message sent by satellite can be received by thousands of ground stations at once. Satellites are used to distribute much network TV programming to local stations for this reason. There is now a large market for satellite broadcasts of digital TV and radio directly to end

users with satellite receivers in their homes and cars. All sorts of other content can be broadcast, too. For example, an organization transmitting a stream of stock, bond, or commodity prices to thousands of dealers might find a satellite system to be much cheaper than simulating broadcasting on the ground.

The United States has some competing satellite-based Internet providers, including Hughes (often marketed as DISH, previously EchoStar) and Viasat, which operate satellites mostly in geostationary or MEO, with some providers moving to LEO. In 2016, the FCC's Measuring Broadband America project reported that these satellite-based providers were among the few Internet Service Providers who were seeing decreased performance over time, likely because of increased subscribership and limited bandwidth. The report found that these providers were unable to offer speeds more than about 10 Mbps.

Nonetheless, in recent years, satellite Internet access has seen growing interest, particularly in niche markets such as in-flight Internet access. Some in-flight Internet access involves direct communication with mobile broadband towers, but for flights over oceans, this does not work. Another method that helps cope with limited bandwidth on airplanes involves transmission of data to a collection of satellites in geostationary orbit. Other companies including OneWeb, as discussed above, and Boeing are working on building a satellite-based Internet backbone using LEO satellites. The markets will still be somewhat niche, as the throughput will be approximately 50 Mbps, much lower than terrestrial Internet.

In short, it looks like the mainstream communication of the future will be terrestrial fiber optics combined with cellular networks, but for some specialized uses, satellites are better. However, one caveat applies to all of this: economics. Although fiber offers more bandwidth, it is conceivable that terrestrial and satellite communication may be able to compete aggressively on price in some markets. If advances in technology radically cut the cost of deploying a satellite (e.g., if some future space vehicle can toss out dozens of satellites on one launch) or low-orbit satellites catch on in a big way, it is not certain that fiber will win all markets.

2.10 POLICY AT THE PHYSICAL LAYER

Various aspects of the physical layer involve regulatory and policy decisions that ultimately affect how these technologies are used and developed. We briefly discuss ongoing policy activity in both terrestrial networks (i.e., the telephone and cable networks) and wireless networks.

2.10.1 Spectrum Allocation

The biggest challenge concerning the electromagnetic spectrum concerns performing **spectrum allocation** efficiently and fairly. If multiple parties can transmit data in the same part of the spectrum in the same geographic region, there is

significant potential for the communicating parties to interfere with one another. To prevent total chaos, there are national and international agreements about who gets to use which frequencies. Because everyone wants a higher data rate, everyone wants more spectrum. National governments allocate spectrum for AM and FM radio, television, and mobile phones, as well as for telephone companies, police, maritime, navigation, military, government, and many other competing users. Worldwide, an agency of ITU-R (WRC) tries to coordinate this allocation so devices that work in multiple countries can be manufactured. However, countries are not bound by ITU-R's recommendations, and the FCC which does the allocation for the United States, has occasionally rejected ITU-R's recommendations (usually because they required some politically powerful group to give up some piece of the spectrum).

Even when a portion of spectrum has been allocated to a specific use, such as mobile phones, there is the additional issue of which company is allowed to use which frequencies. Three algorithms were widely used in the past. The oldest algorithm, often called the **beauty contest**, requires each carrier to explain why its proposal serves the public interest best. Government officials then decide which of the nice stories they enjoy most. Having a government official award property worth billions of dollars to his favorite company often leads to bribery, corruption, nepotism, and worse. Furthermore, even a scrupulously honest government official who thought that a foreign company could do a better job than any of the national companies would have a lot of explaining to do.

This observation led to the second algorithm: holding a **lottery** among the interested companies. The problem with lotteries is that companies with no interest in using the spectrum can enter the lottery. If, say, a hamburger restaurant or shoe store chain wins, it can resell the spectrum to a carrier at a huge profit and with no risk.

Bestowing huge windfalls on alert but otherwise random companies has been severely criticized by many, which led to the third approach: **auction** the spectrum to the highest bidder. When the British government auctioned off the frequencies needed for 3G mobile systems in 2000, it expected to get about $4 billion. It actually received about $40 billion because the carriers got into a feeding frenzy, scared to death of missing the mobile boat. This event switched on other governments' greedy bits and inspired them to hold their own auctions. It worked, but it also left some of the carriers with so much debt that they are close to bankruptcy. Even in the best cases, it will take many years to recoup these licensing fees.

A completely different approach to allocating frequencies is to not allocate them at all. Instead, let everyone transmit at will, but regulate the power used so that stations have such a short range that they do not interfere with each other. Accordingly, most governments have set aside some frequency bands, called the **ISM** (**Industrial, Scientific, Medical**) bands for unlicensed usage. Garage door openers, cordless phones, radio-controlled toys, wireless mice, and numerous other wireless household devices use the ISM bands. To minimize interference between

these uncoordinated devices, the FCC mandates that all devices in the ISM bands limit their transmit power (e.g., to 1 Watt) and use techniques to spread their signals over a range of frequencies. Devices may also need to take care to avoid interference with radar installations.

The location of these bands varies somewhat from country to country. In the United States, for example, the bands that networking devices use in practice without requiring a FCC license are shown in Fig. 2-53. The 900-MHz band was used for early versions of 802.11, but it is crowded. The 2.4-GHz band is available in most countries and widely used for 802.11b/g and Bluetooth, though it is subject to interference from microwave ovens and radar installations. The 5-GHz part of the spectrum includes **U-NII** (**Unlicensed National Information Infrastructure**) bands. The 5-GHz bands are relatively undeveloped but, since they have the most bandwidth and are used by WiFi specifications such as 802.11ac, they have become massively popular and crowded, as well.

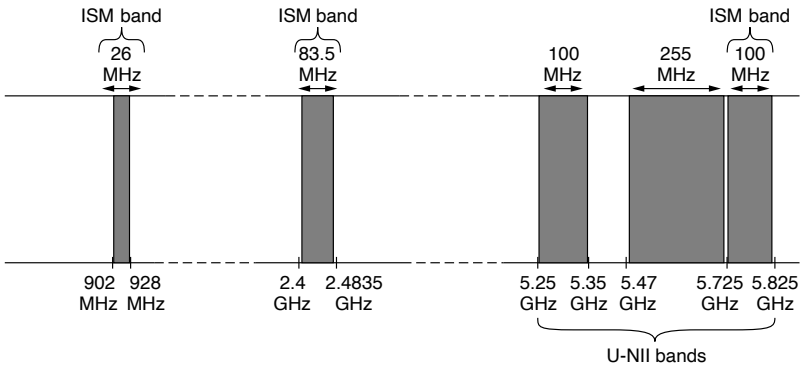


Figure 2-53. ISM and U-NII bands used in the United States by wireless devices.

The unlicensed bands have been a roaring success over the past several decades. The ability to use the spectrum freely has unleashed a huge amount of innovation in wireless LANs and PANs, evidenced by the widespread deployment of technologies including 802.11 and Bluetooth. Even some ISPs are now getting into the game with technologies such as LTE-U, which involves a deployment of an LTE cellular network in the unlicensed spectrum. Such technology could allow mobile devices to operate in this unlicensed spectrum, in addition to the portions of spectrum that are explicitly allocated to operating cellular networks. LTE-U might allow fixed-line ISPs who are deploying WiFi access points in hundreds of millions of homes to turn their network of access points into a network of cellular base stations. Of course, allowing cellular phones to use the unlicensed spectrum comes with its own set of complications. For example, devices that operate in the unlicensed spectrum must respect other devices that are using the same spectrum and

attempt not to interfere with so-called ''incumbent'' devices. LTE-U may also face its own reliability and performance challenges as it must back off to interact nicely with other devices that use the unlicensed spectrum, from other WiFi devices to baby monitors.

Various developments in policy over the past 10 years continue to enable more innovation in wireless technologies. One development in the United States is the potential future allocation of more unlicensed spectrum. In 2009, the FCC decided to allow unlicensed use of **white spaces** around 700 MHz. White spaces are frequency bands that have been allocated but are not being used locally. The transition from analog to all-digital television broadcasts in the United States in 2010 freed up white spaces around 700 MHz. One challenge is that to use the white spaces, unlicensed devices must be able to detect any nearby licensed transmitters, including wireless microphones, that have first rights to use the frequency band. The FCC also opened 57 GHz to 64 GHz for unlicensed operation in 2001. This range is an enormous portion of spectrum, more than all the other ISM bands combined, so it can support the kind of high-speed networks that would be needed to stream high-definition TV through the air across your living room. At 60 GHz, radio waves are absorbed by oxygen. This means that signals do not propagate far, making them well suited to short-range networks. The high frequencies (60 GHz is in the Extremely High Frequency or ''millimeter'' band, just below infrared radiation) posed an initial challenge for equipment makers, but products are now on the market.

In the United States, other spectrum bands are also being repurposed and auctioned off to carriers, including 2.5 and 2.9 GHz, the C-Band (previously used for satellite communications) in the 3.7–4.2 GHz range, as well as others, including 3.5, 6, 24, 28, 37, and 49 GHz. The FCC is also considering the use of certain very high bands for short-range communication, such as the 95 GHz range. In late 2018, the FCC launched its first 5G auction, with more auctions are planned for future years. These auctions will open up a significant amount of spectrum to for mobile broadband, enabling the higher bandwidths that would be required for streaming video and Internet of Things applications. The 24 and 28 GHz spectrum each have approximately 3,000 licenses up for sale. The FCC is also giving discounts to small business and rural providers. Auctions for pieces of the 37, 39, and 49 GHz spectrum bands are scheduled as well. In other countries, some of these spectrum bands may operate as unlicensed spectrum. For example, the automotive industry in Germany successfully lobbied to allow the 3.5 GHz band for private enterprise use; other European countries are likely to follow suit.

2.10.2 The Cellular Network

It is interesting how political and tiny marketing decisions can have a huge impact on the deployment of cellular networks in the United States and Europe. The first mobile system was devised in the U.S. by AT&T and later mandated for

the whole country by the FCC. As a result, the entire U.S. had a single (analog) system and a mobile phone purchased in California also worked in New York. In contrast, when mobile phones came to Europe, every country devised its own system, which resulted in a fiasco.

Europe learned from its mistake and when digital came around, the government-run PTTs got together and standardized on a single system (GSM), so any European mobile phone would work anywhere in Europe. By then, the U.S. had decided that government should not be in the standardization business, so it left digital to the marketplace. This decision resulted in different equipment manufacturers producing different kinds of mobile phones. As a consequence, in the U.S. two major—and completely incompatible—digital mobile phone systems were deployed, as well as other minor systems.

Despite an initial lead by the U.S., mobile phone ownership and usage in Europe is now far greater than in the U.S. Having a single system that works anywhere in Europe and with any provider is part of the reason, but there is more. A second area where the U.S. and Europe differed is in the humble matter of phone numbers. In the U.S., mobile phones are mixed in with regular (fixed) telephones. Thus, there is no way for a caller to see if, say, (212) 234-5678 is a fixed telephone (cheap or free call) or a mobile phone (expensive call). To keep people from getting nervous about placing calls, the telephone companies decided to make the mobile phone owner pay for incoming calls. As a consequence, many people hesitated buying a mobile phone for fear of running up a big bill by just receiving calls. In Europe, mobile phone numbers have a special area code (analogous to 800 and 900 numbers) so they are instantly recognizable. Consequently, the usual rule of ''caller pays'' also applies to mobile phones in Europe (except for international calls, where costs are split).

A third issue that has had a large impact on adoption is the widespread use of prepaid mobile phones in Europe (up to 75% in some areas), which can be purchased in many stores, and even online. These cards are preloaded with a balance of, for example, 20 or 50 euros and can be recharged (using a secret PIN code) when the balance drops to zero. As a consequence, practically every teenager and many small children in Europe have (usually prepaid) mobile phones so their parents can locate them, without the danger of the child running up a huge bill. If the mobile phone is used only occasionally, its use is essentially free since there is no monthly charge or charge for incoming calls.

The auctioning of coveted spectrum bands for 5G, coupled with many technological advances previously discussed in this chapter, is poised to shake up the cellular network edge in the next several years. Already, we are seeing the rise of **MVNO**s (**Mobile Virtual Network Operators**) which are wireless carriers which do not own the network infrastructure over which they provide service to their customers. As cell sizes continue to shrink with higher frequencies and hardware for small cells continues to be commoditized, MVNOs pay to share capacity on an infrastructure that is operated by another carrier. They have the choice whether to

operate their own components of an LTE architecture or use the infrastructure that is owned by the underlying carrier. MVNOs that operate their own core network are sometimes called ''full'' MVNOs. Companies including Qualcomm and Intel are putting together reference design for small cell hardware that could result in the complete disaggregation of the network edge, especially when coupled with the use of unlicensed spectrum. Industry is also beginning to move towards infrastructure with ''whitebox'' eNodeBs that connect to a central office that has virtual EPC services; the Open Networking Foundation's M-CORD project has implemented such an architecture.

2.10.3 The Telephone Network

For decades prior to 1984, the Bell System provided both local and long-distance service throughout most of the United States. In the 1970s, the U.S. federal government came to believe that this was an illegal monopoly and sued to break it up. The government won, and on January 1, 1984, AT&T was broken up into AT&T Long Lines, 23 **BOCs** (**Bell Operating Companies**), and a few other pieces. The 23 BOCs were grouped into seven regional BOCs (RBOCs) to make them economically viable. The entire nature of telecommunication in the United States was changed overnight by court order (*not* by an act of Congress).

The exact specifications of the divestiture were described in the so-called **MFJ** (**Modification of Final Judgment**), an oxymoron if ever there was one. This event led to increased competition, better service, and lower long-distance rates for consumers and businesses. However, prices for local service rose as the cross subsidies from long-distance calling were eliminated and local service had to become self supporting. Many other countries have now introduced competition along similar lines.

Of direct relevance to our studies is that the brand new competitive framework caused a key technical feature to be added to the architecture of the telephone network. To make it clear who could do what, the United States was divided up into 164 **LATAs** (**Local Access and Transport Areas**). Very roughly, a LATA is about as big as the area covered by one area code. Within each LATA, there was one **LEC** (**Local Exchange Carrier**) with a monopoly on traditional telephone service within its area. The most important LECs were the BOCs, although some LATAs contained one or more of the 1500 independent telephone companies operating as LECs.

The new feature was that all inter-LATA traffic was handled by a different kind of company, an **IXC** (**IntereXchange Carrier**). Originally, AT&T Long Lines was the only serious IXC, but now there are well-established competitors such as Verizon and Sprint in the IXC business. One of the concerns at the breakup was to ensure that all the IXCs would be treated equally in terms of line quality, tariffs, and the number of digits their customers would have to dial to use them. The way this is handled is illustrated in Fig. 2-54. Here we see three example LATAs, each

with several end offices. LATAs 2 and 3 also have a small hierarchy with tandem offices (intra-LATA toll offices).

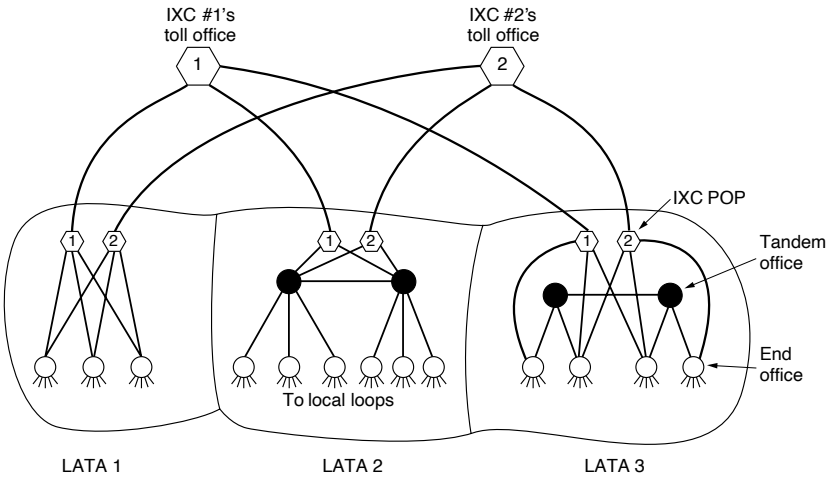


Figure 2-54. The relationship of LATAs, LECs, and IXCs. All the circles are LEC switching offices. Each hexagon belongs to the IXC whose number is in it.

Any IXC that wishes to handle calls originating in a LATA can build a switching office called a **POP** (**Point of Presence**) there. The LEC is required to connect each IXC to every end office, either directly, as in LATAs 1 and 3, or indirectly, as in LATA 2. Furthermore, the terms of the connection, both technical and financial, must be identical for all IXCs. This requirement enables, a subscriber in, say, LATA 1, to choose which IXC to use for calling subscribers in LATA 3.

As part of the MFJ, the IXCs were forbidden to offer local telephone service and the LECs were forbidden to offer inter-LATA telephone service, although both were free to enter any other business, such as operating fried chicken restaurants. In 1984, that was a fairly unambiguous statement. Unfortunately, technology has a funny way of making the law obsolete. Neither cable television nor mobile phones were covered by the agreement. As cable television went from one way to two way and mobile phones exploded in popularity, both LECs and IXCs began buying up or merging with cable and mobile operators.

By 1995, Congress saw that trying to maintain a distinction between the various kinds of companies was no longer tenable and drafted a bill to preserve accessibility for competition but allow cable TV companies, local telephone companies, long-distance carriers, and mobile operators to enter one another's businesses. The idea was that any company could then offer its customers a single integrated package containing cable TV, telephone, and information services and

that different companies would compete on service and price. The bill was enacted into law in February 1996 as a major overhaul of telecommunications regulation. As a result, some BOCs became IXCs and some other companies, such as cable television operators, began offering local telephone service in competition with the LECs.

One interesting property of the 1996 law is the requirement that LECs implement **local number portability**. This means that a customer can change local telephone companies without having to get a new telephone number. Portability for mobile phone numbers (and between fixed and mobile lines) followed suit in 2003. These provisions removed a huge hurdle for many people, making them much more inclined to switch LECs. As a result, the U.S. telecommunications landscape became much more competitive, and other countries have followed suit. Often other countries wait to see how this kind of experiment works out in the U.S. If it works well, they do the same thing; if it works badly, they try something else.

In recent years, telecommunications policy has been relatively quiet, as it pertains to telephone companies, with most of the action and activity shifting to Internet service providers. Two recent developments, however, involve policy activity surrounding the insecurities of a signaling protocol called **SS7** (**Signaling System 7**), which is the protocol that allows cellular networks to talk to one another. The protocol is insecure, and Congress has asked the FCC to take action to address some of these insecurities. Another interesting development related to the 1996 Telecommunications Act is how text messages are classified; unlike voice traffic over the telephone network, which is classified as a communications service (like phone calls), SMS messages (''text messages'') are classified as an information service (akin to instant messages or other Internet communications services), which subjects them to very different sets of regulations concerning everything from how they can be billed to the privacy rules that govern these messages.

2.11 SUMMARY

The physical layer is the basis of all networks. Nature imposes two fundamental limits on all channels, and these determine their bandwidth. These limits are the Nyquist limit, which deals with noiseless channels, and the Shannon limit, which deals with noisy channels.

Transmission media can be guided or unguided. The principal guided media are twisted pair, coaxial cable, and fiber optics. Unguided media include terrestrial radio, microwaves, infrared, lasers through the air, and satellites.

Digital modulation methods send bits over guided and unguided media as analog signals. Line codes operate at baseband, and signals can be placed in a passband by modulating the amplitude, frequency, and phase of a carrier. Channels can be shared between users with time, frequency, and code division multiplexing.

A key element in many wide area networks is the telephone system. Its main components are the local loops, trunks, and switches. ADSL offers speeds up to 40 Mbps over the local loop by dividing it into many subcarriers that run in parallel. This far exceeds the rates of telephone modems. PONs bring fiber to the home for even greater access rates than ADSL. Trunks carry digital information. They are multiplexed with WDM to provision many high capacity links over individual fibers, as well as with TDM to share each high rate link between users. Both circuit switching and packet switching play a role.

Another system for network access is the cable infrastructure, which has gradually evolved from coaxial cable to hybrid fiber coax, where many cable Internet service providers now offer subscribers up to 1 Gbps (and, within a few years, likely 10 Gbps). The architecture of these networks is quite different, however, in that the capacity of the network is shared among subscribers in the same service node.

For mobile devices applications, the fixed telephone system is not suitable. Mobile phones are currently in widespread use for voice and data; since 4G, all voice is, in fact, carried over a packet-switched network. The first generation, 1G, was analog and dominated by AMPS. 2G was digital, with GSM presently the most widely deployed mobile phone system in the world. 3G is digital and based on broadband CDMA. 4G's main innovation was to shift to a packet-switched core. 5G is defined by smaller cell sizes, massive MIMO, and the use of significantly more spectrum.

Many aspects of the physical layer are ultimately determined not only by the technologies themselves, but also by policy organizations, such as standards bodies and regulatory agencies. One area of the physical layer that is fairly dynamic in the policy arena is wireless spectrum, much of which is highly regulated. As the need for more bandwidth for data communications grows, regulatory agencies are actively searching for ways to use existing spectrum more efficiently, such as reappropriating and auctioning portions of previously allocated spectrum.

PROBLEMS

1. Is an oil pipeline a simplex system, a half-duplex system, a full-duplex system, or none of the above? What about a river or a walkie-talkie-style communication?

2. What are the advantages of fiber optics over copper as a transmission medium? Is there any downside of using fiber optics over copper?

3. How much bandwidth is there in 0.1 microns of spectrum at a wavelength of 1 micron?

4. It is desired to send a sequence of computer screen images over an optical fiber. The screen is 3840 × 2160 pixels, each pixel being 24 bits. There are 50 screen images per second. What data rate is needed is needed?

5. In Fig. 2-5, the left-hand band is narrower than the others. Why?

6. Radio antennas often work best when the diameter of the antenna is equal to the wavelength of the radio wave. Reasonable antennas range from 1 cm to 1 meter in diameter. What frequency range does this cover?

7. Multipath fading is maximized when the two beams arrive 180 degrees out of phase. How much of a path difference is required to maximize the fading for a 100-km-long 1-GHz microwave link?

8. A laser beam 1 mm wide is aimed at a detector 1 mm wide 100 m away on the roof of a building. How much of an angular diversion (in degrees) does the laser have to have before it misses the detector?

9. Compute the Fourier coefficients for the function $f(t) = t \quad (0 \le t \le 1)$.

10. Identify three physical properties that limit the maximum data rate of digital communication channels used in practice. Explain your answers.

11. A noiseless 10-kHz channel is sampled every 1 msec. What is the maximum data rate?

12. Is the Nyquist theorem true for high-quality single-mode optical fiber or only for copper wire?

13. Television channels are 6 MHz wide. How many bits/sec can be sent if four-level digital signals are used? Assume a noiseless channel.

14. If a binary signal is sent over a 3-kHz channel whose signal-to-noise ratio is 20 dB, what is the maximum achievable data rate?

15. You need to select a line code that will only be used to send the bit sequences 10101010 and 00111100. Which of the lines codes shown in Fig. 2-14 is not a good candidate? Consider both bandwidth efficiency and clock recovery.

16. What is the minimum bandwidth needed to achieve a data rate of B bits/sec if the signal is transmitted using NRZ, MLT-3, and Manchester encoding? Explain.

17. Prove that in 4B/5B mapped data with the NRZI encoding, a signal transition will occur at least every four bit times.

18. A modem constellation diagram similar to Fig. 2-17 has data points at (0, 1) and (0, 2). Does the modem use phase modulation or amplitude modulation?

19. In a constellation diagram, all the points lie on a circle centered on the origin. What kind of modulation is being used?

20. Ten signals, each requiring 4000 Hz, are multiplexed onto a single channel using FDM. What is the minimum bandwidth required for the multiplexed channel? Assume that the guard bands are 400 Hz wide.

21. Suppose that A, B, and C are simultaneously transmitting 0 bits, using a CDMA system with the chip sequences of Fig. 2-22(a). What is the resulting chip sequence?

22. In the discussion about orthogonality of CDMA chip sequences, it was stated that if $\mathbf{S} \cdot \mathbf{T} = 0$ then $\mathbf{S} \cdot \bar{\mathbf{T}}$ is also 0. Prove this.

23. Consider a different way of looking at the orthogonality property of CDMA chip se-

quences. Each bit in a pair of sequences can match or not match. Express the orthogonality property in terms of matches and mismatches.

24. A CDMA receiver gets the following chips: (−1 +1 −3 +1 −1 −3 +1 +1). Assuming the chip sequences defined in Fig. 2-22(a), which stations transmitted, and which bits did each one send?

25. In Fig. 2-22, there are four stations that can transmit. Suppose four more stations are added. Provide the chip sequences of these stations.

26. A base station schedules a single slot for devices A and B to send data using their corresponding chip sequences from Fig. 2-22. During this time, other stations remain silent. Due to noise, some of the chips are lost. The base station receives the following sequence: (0, 0, ?, 2, ?, ?, 0,-2). What are the bit values transmitted by stations A and B?

27. How many end office codes were there pre-1984, when each end office was named by its three-digit area code and the first three digits of the local number? Area codes started with a digit in the range 2–9, had a 0 or 1 as the second digit, and ended with any digit. The first two digits of a local number were always in the range 2–9. The third digit could be any digit.

28. A simple telephone system consists of two end offices and a single toll office to which each end office is connected by a 1-MHz full-duplex trunk. The average telephone is used to make four calls per 8-hour workday. The mean call duration is 6 min. Ten percent of the calls are long distance (i.e., pass through the toll office). What is the maximum number of telephones an end office can support? (Assume 4 kHz per circuit.) Explain why a telephone company may decide to support a lesser number of telephones than this maximum number at the end office.

29. A regional telephone company has 15 million subscribers. Each of their telephones is connected to a central office by a copper twisted pair. The average length of these twisted pairs is 10 km. How much is the copper in the local loops worth? Assume that the cross section of each strand is a circle 1 mm in diameter, the density of copper is 9.0 grams/cm^3, and that copper sells for $6 per kilogram.

30. What is the maximum bit rate achievable in a V.32 standard modem if the baud rate is 9600 and no error correction is used?

31. The cost of a fast microprocessor has dropped to the point where it is now possible to put one in each modem. How does that affect the handling of telephone line errors? Does it negate the need for error checking/correction in layer 2?

32. An ADSL system using DMT allocates 3/4 of the available data channels to the downstream link. It uses QAM-64 modulation on each channel. What is the capacity of the downstream link?

33. Why has the PCM sampling time been set at 125 μsec?

34. What signal-to-noise ratio is needed to put a T1 carrier on a 200-kHz line?

35. Compare the maximum data rate of a noiseless 4-kHz channel using
(a) Analog encoding (e.g., QPSK) with 2 bits per sample.
(b) The T1 PCM system.

36. If a T1 carrier system slips and loses track of where it is, it tries to resynchronize using the first bit in each frame. How many frames will have to be inspected on average to resynchronize with a probability of 0.001 of being wrong?

37. What is the percent overhead on a T1 carrier? That is, what percent of the 1.544 Mbps are not delivered to the end user? How does it relate to the percent overhead in OC-1 or OC-768 lines?

38. SONET clocks have a drift rate of about 1 part in 10^9. How long does it take for the drift to equal the width of 1 bit? Do you see any practical implications of this calculation? If so, what?

39. In Fig. 2-35, the user data rate for OC-3 is stated to be 148.608 Mbps. Show how this number can be derived from the SONET OC-3 parameters. What will be the gross, SPE, and user data rates of an OC-3072 line?

40. To accommodate lower data rates than STS-1, SONET has a system of virtual tributaries (VTs). A VT is a partial payload that can be inserted into an STS-1 frame and combined with other partial payloads to fill the data frame. VT1.5 uses 3 columns, VT2 uses 4 columns, VT3 uses 6 columns, and VT6 uses 12 columns of an STS-1 frame. Which VT can accommodate

(a) A DS-1 service (1.544 Mbps)?
(b) European CEPT-1 service (2.048 Mbps)?
(c) A DS-2 service (6.312 Mbps)?

41. What is the available user bandwidth in an OC-12c connection?

42. What is the difference, if any, between the demodulator part of a modem and the coder part of a codec? (After all, both convert analog signals to digital ones.)

43. Three packet-switching networks each contain n nodes. The first network has a star topology with a central switch, the second is a (bidirectional) ring, and the third is fully interconnected, with a wire from every node to every other node. What are the best-, average-, and worst-case transmission paths in hops?

44. Compare the delay in sending an x-bit message over a k-hop path in a circuit-switched network and in a (lightly loaded) packet-switched network. The circuit setup time is s sec, the propagation delay is d sec per hop, the packet size is p bits, and the data rate is b bps. Under what conditions does the packet network have a lower delay? Also, explain the conditions under which a packet-switched network is preferable to a circuit-switched network.

45. Suppose that x bits of user data are to be transmitted over a k-hop path in a packet-switched network as a series of packets, each containing p data bits and h header bits, with $x \gg p + h$. The bit rate of the lines is b bps and the propagation delay is negligible. What value of p minimizes the total delay?

46. In a typical mobile phone system with hexagonal cells, it is forbidden to reuse a frequency band in an adjacent cell. If 840 frequencies are available, how many can be used in a given cell?

47. The actual layout of cells is seldom as regular that as shown in Fig. 2-39. Even the

shapes of individual cells are typically irregular. Give a possible reason why this might be. How do these irregular shapes affect frequency assignment to each cell?

48. Make a rough estimate of the number of PCS microcells 100 m in diameter it would take to cover San Francisco (120 square km).

49. Sometimes when a mobile user crosses the boundary from one cell to another, the current call is abruptly terminated, even though all transmitters and receivers are functioning perfectly. Why?

50. At the low end, the telephone system is star shaped, with all the local loops in a neighborhood converging on an end office. In contrast, cable television consists of a single long cable snaking its way past all the houses in the same neighborhood. Suppose that a future TV cable were 10-Gbps fiber instead of copper. Could it be used to simulate the telephone model of everybody having their own private line to the end office? If so, how many one-telephone houses could be hooked up to a single fiber?

51. A cable company decides to provide Internet access over cable in a neighborhood consisting of 5000 houses. The company uses a coaxial cable and spectrum allocation allowing 100 Mbps downstream bandwidth per cable. To attract customers, the company decides to guarantee at least 2 Mbps downstream bandwidth to each house at any time. Describe what the cable company needs to do to provide this guarantee.

52. Using the spectral allocation of Fig. 2-46 and the information given in the text, how many Mbps does a cable system allocate to upstream and how many to downstream?

53. How fast can a cable user receive data if the network is otherwise idle? Assume that the user interface is
(a) 10-Mbps Ethernet
(b) 100-Mbps Ethernet
(c) 54-Mbps Wireless.

54. The 66 low-orbit satellites in the Iridium project are divided into six necklaces around the earth. At the altitude they are using, the period is 90 minutes. What is the average interval for handoffs for a stationary transmitter?

55. Consider a satellite at the altitude of geostationary satellites but whose orbital plane is inclined to the equatorial plane by an angle ϕ. To a stationary user on the earth's surface at north latitude ϕ, does this satellite appear motionless in the sky? If not, describe its motion.

56. Calculate the end-to-end transit time for a packet for both GEO (altitude: 35,800 km), MEO (altitude: 18,000 km), and LEO (altitude: 750 km) satellites.

57. What is the latency of a call originating at the North Pole to reach the South Pole if the call is routed via Iridium satellites? Assume that the switching time at the satellites is 10 microseconds and earth's radius is 6371 km.

58. How long will it take to transmit a 1-GB file from one VSAT to another using a hub as shown in Fig. 2-50? Assume that the uplink is 1 Mbps, the downlink is 7 Mbps, and circuit switching is used with 1.2 sec circuit setup time.

59. Calculate the transmit time in the previous problem if packet switching is used instead.

Assume that the packet size is 64 KB, the switching delay in the satellite and hub is 10 microseconds, and the packet header size is 32 bytes.

60. Multiplexing STS-1 multiple data streams, called tributaries, plays an important role in SONET. A 3:1 multiplexer multiplexes three input STS-1 tributaries onto one output STS-3 stream. This multiplexing is done byte for byte. That is, the first three output bytes are the first bytes of tributaries 1, 2, and 3, respectively. The next three output bytes are the second bytes of tributaries 1, 2, and 3, respectively, and so on. Write a program that simulates this 3:1 multiplexer. Your program should consist of five processes. The main process creates four processes, one each for the three STS-1 tributaries and one for the multiplexer. Each tributary process reads in an STS-1 frame from an input file as a sequence of 810 bytes. They send their frames (byte by byte) to the multiplexer process. The multiplexer process receives these bytes and outputs an STS-3 frame (byte by byte) by writing it to standard output. Use pipes for communication among processes.

61. Write a program to implement CDMA. Assume that the length of a chip sequence is eight and the number of stations transmitting is four. Your program consists of three sets of processes: four transmitter processes (t0, t1, t2, and t3), one joiner process, and four receiver processes (r0, r1, r2, and r3). The main program, which also acts as the joiner process first reads four chip sequences (bipolar notation) from the standard input and a sequence of 4 bits (1 bit per transmitter process to be transmitted), and forks off four pairs of transmitter and receiver processes. Each pair of transmitter/receiver processes (t0,r0; t1,r1; t2,r2; t3,r3) is assigned one chip sequence and each transmitter process is assigned 1 bit (first bit to t0, second bit to t1, and so on). Next, each transmitter process computes the signal to be transmitted (a sequence of 8 bits) and sends it to the joiner process. After receiving signals from all four transmitter processes, the joiner process combines the signals and sends the combined signal to the four receiver processes. Each receiver process then computes the bit it has received and prints it to standard output. Use pipes for communication between processes.