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.

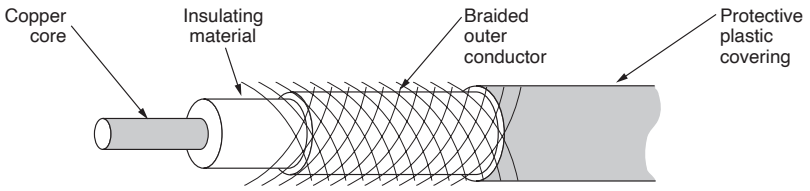


Figure 2-2. A coaxial cable.

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.

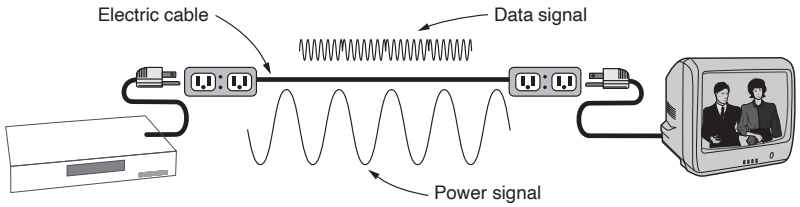


Figure 2-3. A network that uses household electrical wiring.

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.

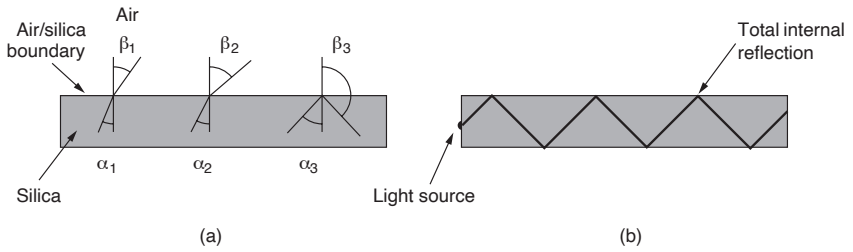


Figure 2-4. (a) Three examples of a light ray from inside a silica fiber impinging on the air/silica boundary at different angles. (b) Light trapped by total internal reflection.

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

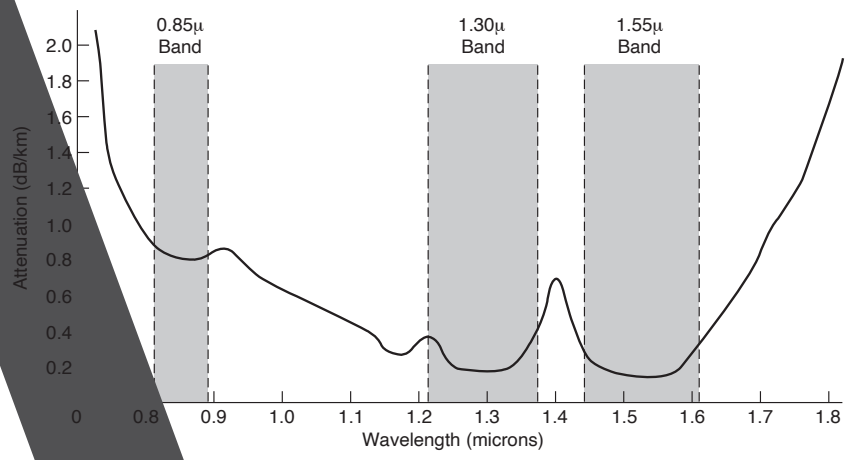


Figure 2-5. Attenuation of light through fiber in the infrared region.

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.

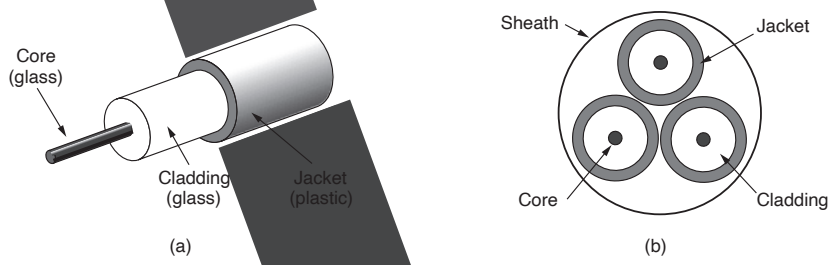


Figure 2-6. (a) Side view of a single fiber. (b) End view of a sheath with three fibers.

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,

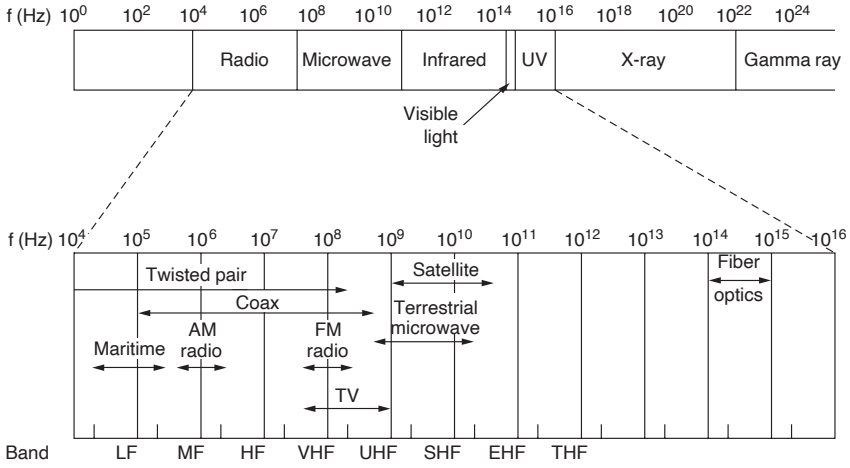


Figure 2-8. The electromagnetic spectrum and its uses for communication.

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).

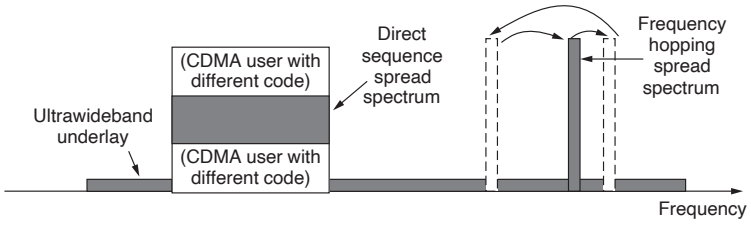


Figure 2-9. Spread spectrum and ultra-wideband (UWB) communication.

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

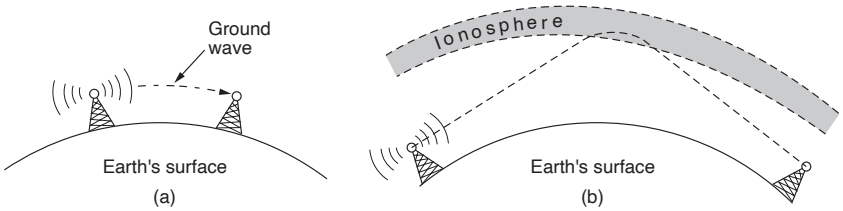


Figure 2-10. (a) In the VLF, LF, and MF bands, radio waves follow the curvature of the earth. (b) In the HF band, they bounce off the ionosphere.

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

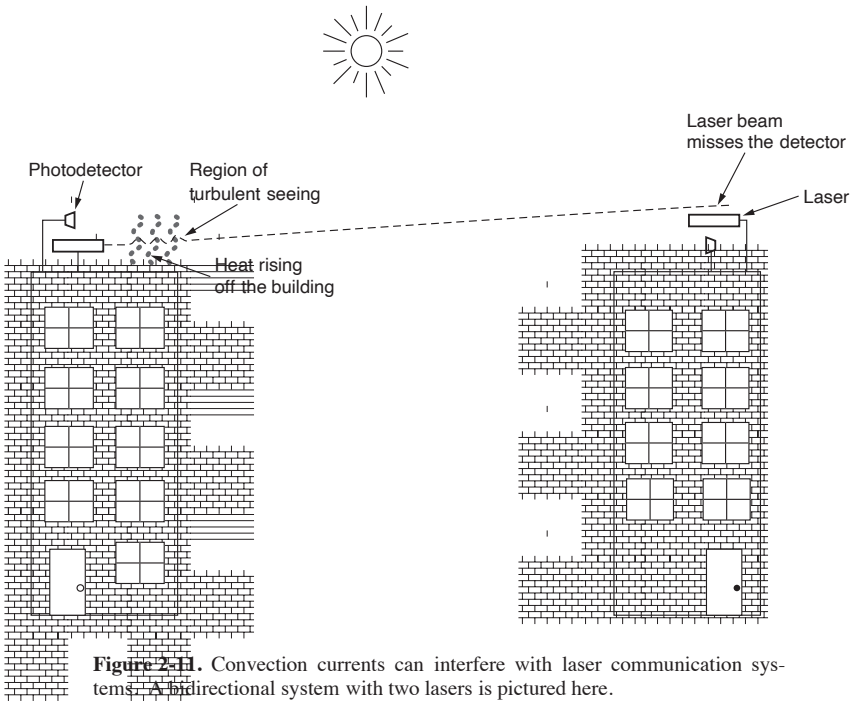


Figure 2-11. Convection currents can interfere with laser communication systems. A bidirectional system with two lasers is pictured here.

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

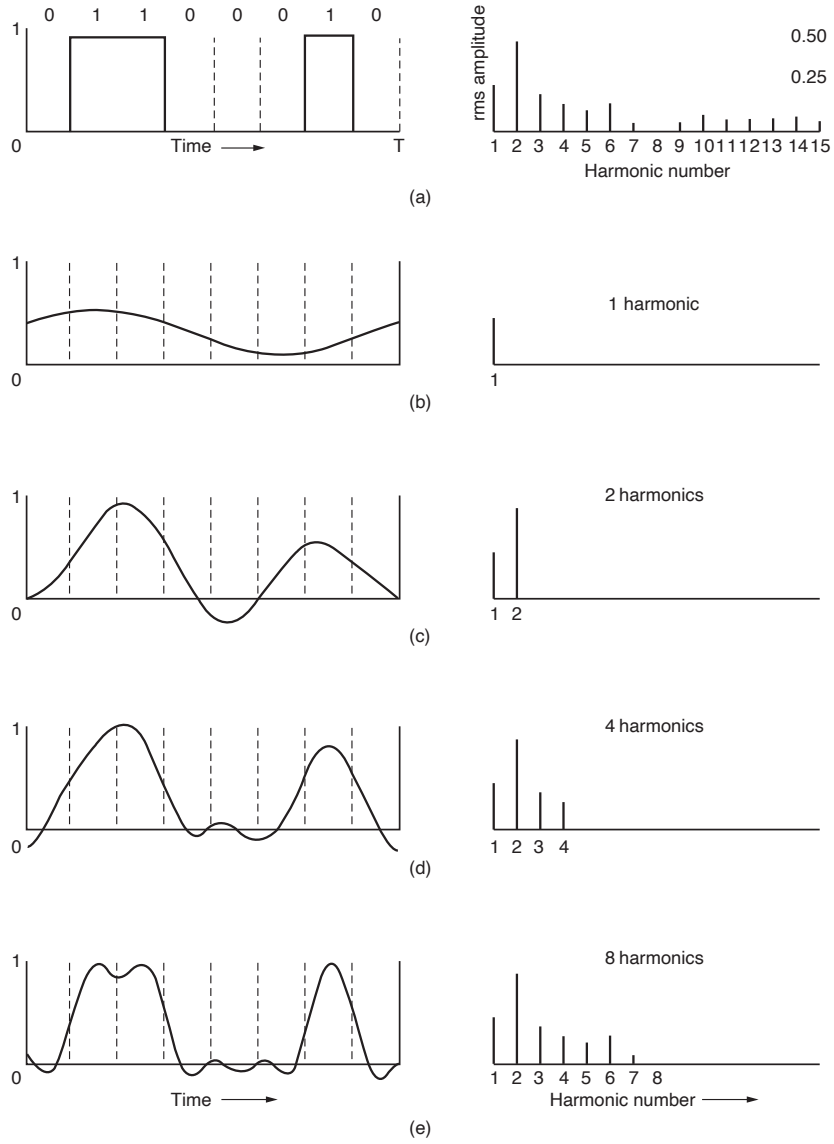


Figure 2-12. (a) A binary signal and its root-mean-square Fourier amplitudes. (b)–(e) Successive approximations to the original signal.

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).

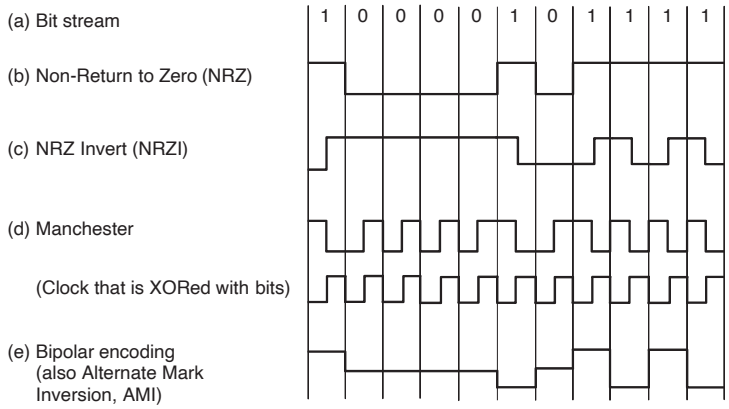


Figure 2-14. Line codes: (a) Bits, (b) NRZ, (c) NRZI, (d) Manchester, (e) Bipolar or AMI.

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.

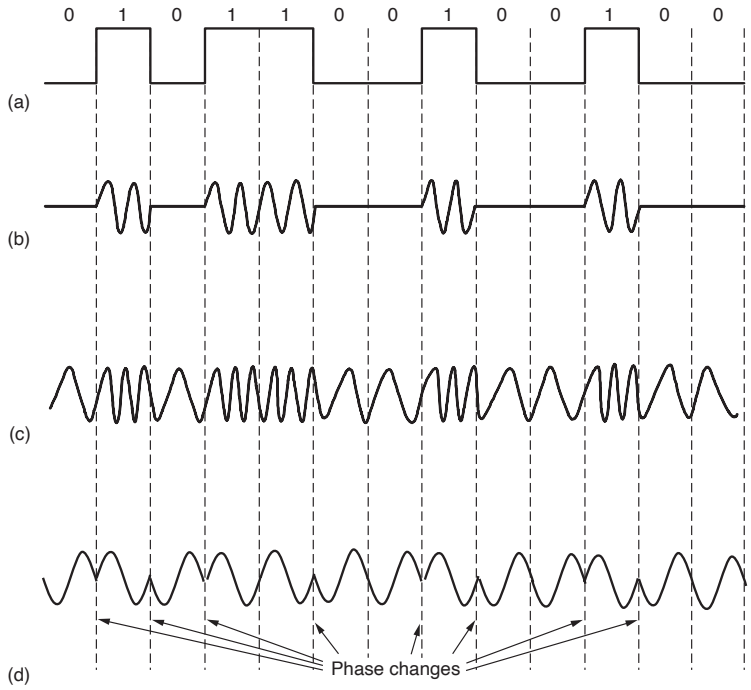


Figure 2-16. (a) A binary signal. (b) Amplitude shift keying. (c) Frequency shift keying. (d) Phase shift keying.

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.

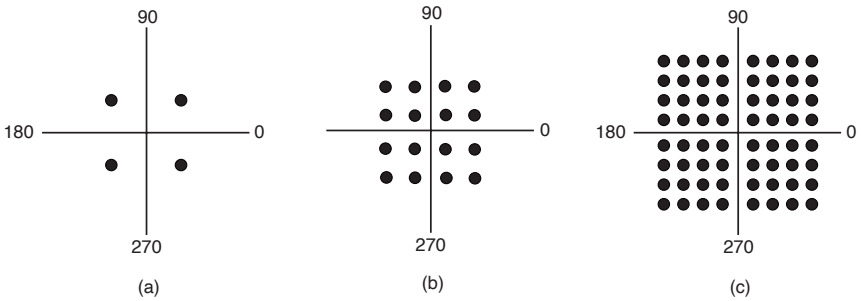
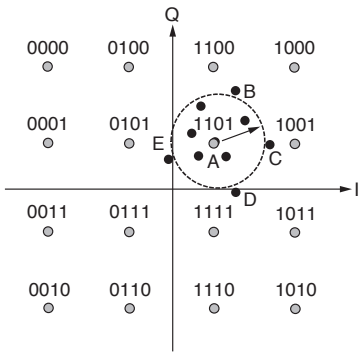


Figure 2-17. (a) QPSK. (b) QAM-16. (c) QAM-64.

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**.