1 Introduction

This textbook provides an introduction to the conceptual underpinnings of communication technologies. Most of us directly experience such technologies daily: browsing (and audio/video streaming from) the Internet, sending/receiving emails, watching television, or carrying out a phone conversation. Many of these experiences occur on mobile devices that we carry around with us, so that we are always connected to the cyberworld of modern communication systems. In addition, there is a huge amount of machine-to-machine communication that we do not directly experience, but which is indispensable for the operation of modern society. This includes, for example, signaling between routers on the Internet, or between processors and memories on any computing device.

We define *communication* as the process of *information transfer across space or time.* Communication across space is something we have an intuitive understanding of: for example, radio waves carry our phone conversation between our cell phone and the nearest base station, and coaxial cables (or optical fiber, or radio waves from a satellite) deliver television from a remote location to our home. However, a moment's thought shows that that communication across time, or storage of information, is also an everyday experience, given our use of storage media such as compact discs (CDs), digital video discs (DVDs), hard drives, and memory sticks. In all of these instances, the key steps in the operation of a communication link are as follows:

(a) insertion of information into a signal, termed the *transmitted signal,* compatible with the physical medium of interest;
(b) propagation of the signal through the physical medium (termed the *channel*) in space or time; and
(c) extraction of information from the signal (termed the *received signal*) obtained after propagation through the medium.

In this book, we study the fundamentals of modeling and design for these steps.

Chapter plan

In Section 1.1, we provide a high-level description of analog and digital communication systems, and discuss why digital communication is the inevitable design choice in modern systems. In Section 1.2, we briefly provide a technological perspective on recent

developments in communication. We do not attempt to provide a comprehensive discussion of the fascinating history of communication: thanks to the advances in communication that brought us the Internet, it is easy to look it up online! A discussion of the scope of this textbook is provided in Section 1.3.

1.1 Analog or digital?

Even without defining information formally, we intuitively understand that speech, audio, and video signals contain information. We use the term *message signals* for such signals, since these are the messages we wish to convey over a communication system. In their original form – both during generation and consumption – these message signals are *analog*: they are continuous-time signals, with the signal values also lying in a continuum. When someone plays the violin, an analog acoustic signal is generated (often translated to an analog electrical signal using a microphone). Even when this music is recorded onto a digital storage medium such as a CD (using the digital communication framework outlined in Section 1.1.2), when we ultimately listen to the CD being played on an audio system, we hear an analog acoustic signal. The transmitted signals corresponding to physical communication media are also analog. For example, in both wireless and optical communication, we employ electromagnetic waves, which correspond to continuous-time electric and magnetic fields taking values in a continuum.

1.1.1 Analog communication

Given the analog nature of both the message signal and the communication medium, a natural design choice is to map the analog message signal (e.g., an audio signal, translated from the acoustic to the electrical domain using a microphone) to an analog transmitted signal (e.g., a radio wave carrying the audio signal) that is compatible with the physical medium over which we wish to communicate (e.g., broadcasting audio over the air from an FM radio station). This approach to communication system design, depicted in Figure 1.1, is termed *analog communication.* Early communication systems were all analog: examples include AM (amplitude modulation) and FM (frequency modulation) radio, analog television, first-generation cellular-phone technology (based on FM), vinyl records, audio cassettes, and VHS or Betamax videocassettes

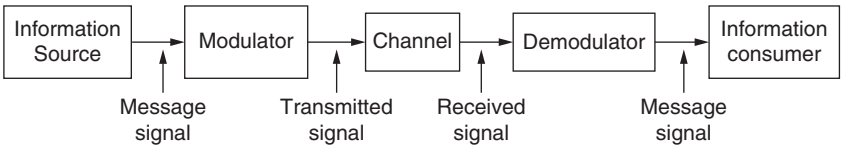

Figure 1.1 A block diagram for an analog communication system. The modulator transforms the message signal into the transmitted signal. The channel distorts and adds noise to the transmitted signal. The demodulator extracts an estimate of the message signal from the received signal arriving from the channel.

While analog communication might seem like the most natural option, it is in fact obsolete. Cellular-phone technologies from the second generation onwards are digital; vinyl records and audio cassettes have been supplanted by CDs, and videocassettes by DVDs. Broadcast technologies such as radio and television are often slower to upgrade because of economic and political factors, but digital broadcast radio and television technologies are either replacing or sidestepping (e.g., via satellite) analog FM/AM radio and television broadcast. Let us now define what we mean by digital communication, before discussing the reasons for the inexorable trend away from analog and towards digital communication.

1.1.2 Digital communication

The conceptual basis for digital communication was established in 1948 by Claude Shannon, when he founded the field of information theory. There are two main threads to this theory.

- **Source coding and compression.** Any information-bearing signal can be represented efficiently, to within a desired accuracy of reproduction, by a digital signal (i.e., a discrete-time signal taking values from a discrete set), which in its simplest form is just a sequence of binary digits (zeros or ones), or *bits*. This is true irrespective of whether the information source is text, speech, audio, or video. Techniques for performing the mapping from the original source signal to a bit sequence are generically termed *source coding*. They often involve *compression,* or removal of redundancy, in a manner that exploits the properties of the source signal (e.g., the heavy spatial correlation among adjacent pixels in an image can be exploited to represent it more efficiently than a pixel-by-pixel representation).
- **Digital information transfer.** Once the source encoding has been done, our communication task reduces to reliably transferring the bit sequence at the output of the source encoder across space or time, without worrying about the original source and the sophisticated tricks that have been used to encode it. The performance of any communication system depends on the relative strengths of the signal and noise or interference, and the distortions imposed by the channel. Shannon showed that, once we have fixed these operational parameters for any communication channel, there exists a maximum possible rate of reliable communication, termed the *channel capacity*. Thus, given the information bits at the output of the source encoder, in principle, we can transmit them reliably over a given link as long as the information rate is smaller than the channel capacity, and we cannot transmit them reliably if the information rate is larger than the channel capacity. This sharp transition between reliable and unreliable communication differs fundamentally from analog communication, where the quality of the reproduced source signal typically degrades gradually as the channel conditions get worse.

A block diagram for a typical digital communication system based on these two threads is shown in Figure 1.2. We now briefly describe the role of each component, together with simplified examples of its function.

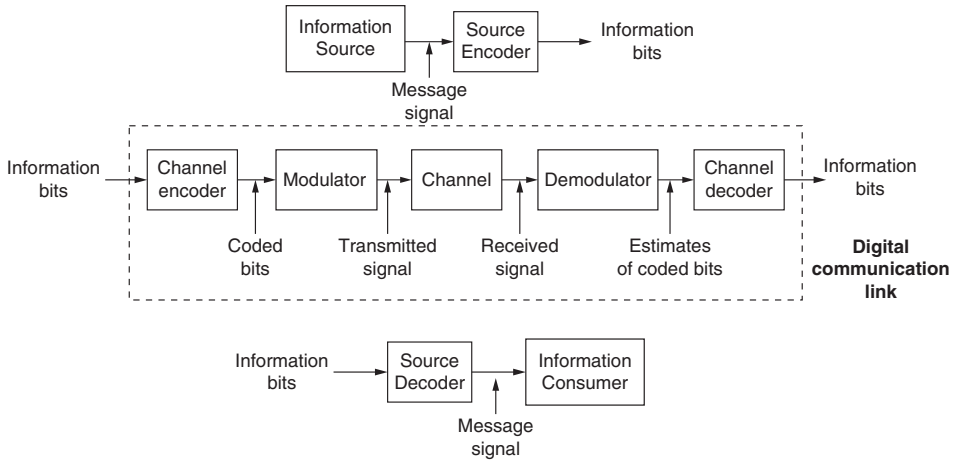

Figure 1.2 Components of a digital communication system.

Source encoder As already discussed, the source encoder converts the message signal into a sequence of information bits. The information bit rate depends on the nature of the message signal (e.g., speech, audio, video) and the application requirements. Even when we fix the class of message signals, the choice of source encoder is heavily dependent on the setting. For example, video signals are heavily compressed when they are sent over a cellular link to a mobile device, but are lightly compressed when sent to a high-definition television (HDTV) set. A cellular link can support a much smaller bit rate than, say, the cable connecting a DVD player to an HDTV set, and a smaller mobile display device requires lower resolution than a large HDTV screen. In general, the source encoder must be chosen such that the bit rate it generates can be supported by the digital communication link we wish to transfer information over. Other than this, source coding can be decoupled entirely from link design (we comment further on this a bit later).

Example. A laptop display may have resolution 1024×768 pixels. For a grayscale digital image, the intensity for each pixel might be represented by 8 bits. Multiplying by the number of pixels gives us about 6.3 million bits, or about 0.8 Mbyte (a byte equals 8 bits). However, for a typical image, the intensities for neighboring pixels are heavily correlated, which can be exploited for significantly reducing the number of bits required to represent the image, without noticeably distorting it. For example, one could take a two-dimensional Fourier transform, which concentrates most of the information in the image at lower frequencies, and then discard many of the high-frequency coefficients. There are other possible transforms one could use, and also several more processing stages, but the bottom line is that, for natural images, state-of-the-art image-compression algorithms can provide $10\times$ compression (i.e., reduction in the number of bits relative to the original uncompressed digital image) with hardly any perceptual degradation. Far more aggressive compression ratios are possible if we are willing to tolerate more distortion. For video, in addition to the spatial correlation exploited for image compression, we can also exploit temporal correlation across successive frames.

Channel encoder The channel encoder adds redundancy to the information bits obtained from the source encoder, in order to facilitate error recovery after transmission over the channel. It might appear that we are putting in too much work, adding redundancy just after the source encoder has removed it. However, the redundancy added by the channel encoder is tailored to the channel over which information transfer is to occur, whereas the redundancy in the original message signal is beyond our control, so that it would be inefficient to keep it when we transmit the signal over the channel.

Example. The noise and distortion introduced by the channel can cause errors in the bits we send over it. Consider the following abstraction for a channel: we can send a string of bits (zeros or ones) over it, and the channel randomly flips each bit with probability 0.01 (i.e., the channel has a 1% error rate). If we cannot tolerate this error rate, we could repeat each bit that we wish to send three times, and use a majority rule to decide on its value. Now, we only make an error if two or more of the three bits are flipped by the channel. It is left as an exercise to calculate that an error now happens with probability approximately 0.0003 (i.e., the error rate has gone down to 0.03%). That is, we have improved performance by introducing redundancy. Of course, there are far more sophisticated and efficient techniques for introducing redundancy than the simple repetition strategy just described; see Chapter 7.

Modulator The modulator maps the coded bits at the output of the channel encoder to a transmitted signal to be sent over the channel. For example, we may insist that the transmitted signal fit within a given frequency band and adhere to stringent power constraints in a wireless system, where interference between users and between co-existing systems is a major concern. Unlicensed WiFi transmissions typically occupy 20–40 MHz of bandwidth in the 2.4- or 5-GHz bands. Transmissions in fourth-generation cellular systems may often occupy bandwidths ranging from 1 to 20 MHz at frequencies ranging from 700 MHz to 3 GHz. While these signal bandwidths are being increased in an effort to increase data rates (e.g., up to 160 GHz for emerging WiFi standards, and up to 100 MHz for emerging cellular standards), and new frequency bands are being actively explored (see the epilogue for more discussion), the transmitted signal still needs to be shaped to fit within certain spectral constraints.

Example. Suppose that we send bit value 0 by transmitting the signal $s(t)$, and bit value 1 by transmitting $-s(t)$. Even for this simple example, we must design the signal $s(t)$ so it fits within spectral constraints (e.g., two different users may use two different segments of spectrum to avoid interfering with each other), and we must figure out how to prevent successive bits of the same user from interfering with each other. For wireless communication, these signals are voltages generated by circuits coupled to antennas, and are ultimately emitted as electromagnetic waves from the antennas.

The channel encoder and modulator are typically jointly designed, keeping in mind the anticipated channel conditions, and the result is termed a *coded modulator.*

Channel The channel distorts and adds noise, and possibly interference, to the transmitted signal. Much of our success in developing communication technologies has resulted from being able to optimize communication strategies based on accurate mathematical models

for the channel. Such models are typically statistical, and are developed with significant effort using a combination of measurement and computation. The physical characteristics of the communication medium vary widely, and hence so do the channel models. Wireline channels are typically well modeled as linear and time-invariant, while optical-fiber channels exhibit nonlinearities. Wireless mobile channels are particularly challenging because of the time variations caused by mobility, and due to the potential for interference due to the broadcast nature of the medium. The link design also depends on system-level characteristics, such as whether or not the transmitter has feedback regarding the channel, and what strategy is used to manage interference.

Example. Consider communication between a cellular base station and a mobile device. The electromagnetic waves emitted by the base station can reach the mobile's antennas through multiple paths, including bounces off streets and building surfaces. The received signal at the mobile can be modeled as multiple copies of the transmitted signal with different gains and delays. These gains and delays change due to mobility, but the rate of change is often slow compared with the data rate, hence, over short intervals, we can get away with modeling the channel as a linear time-invariant system that the transmitted signal goes through before arriving at the receiver.

Demodulator The demodulator processes the signal received from the channel to produce bit estimates to be fed to the channel decoder. It typically performs a number of signal-processing tasks, such as synchronization of phase, frequency, and timing, and compensating for distortions induced by the channel.

Example. Consider the simplest possible channel model, where the channel just adds noise to the transmitted signal. In our earlier example of sending $\pm s(t)$ to send 0 or 1, the demodulator must guess, based on the noisy received signal, which of these two options is true. It might make a hard decision (e.g., guess that 0 was sent), or hedge its bets, and make a soft decision, saying, for example, that it is 80% sure that the transmitted bit is a zero. There are many other aspects of demodulation that we have swept under the rug: for example, before making any decisions, the demodulator has to perform functions such as synchronization (making sure that the receiver's notion of time and frequency is consistent with the transmitter's) and equalization (compensating for the distortions due to the channel).

Channel decoder The channel decoder processes the imperfect bit estimates provided by the demodulator, and exploits the controlled redundancy introduced by the channel encoder to estimate the information bits.

Example. The channel decoder takes the guesses from the demodulator and uses the redundancies in the channel code to clean up the decisions. In our simple example of repeating every bit three times, it might use a majority rule to make its final decision if the demodulator is putting out hard decisions. For soft decisions, it might use more sophisticated combining rules with improved performance.

While we have described the demodulator and decoder as operating separately and in sequence for simplicity, there can be significant benefits from iterative information exchange between the two. In addition, for certain coded modulation strategies in which

channel coding and modulation are tightly coupled, the demodulator and channel decoder may be integrated into a single entity.

Source decoder The source decoder processes the estimated information bits at the output of the channel decoder to obtain an estimate of the message. The message format may, but need not, be the same as that of the original message input to the source encoder: for example, the source encoder may translate speech to text before encoding into bits, and the source decoder may output a text message to the end user.

Example. For the example of a digital image considered earlier, the compressed image can be translated back to a pixel-by-pixel representation by taking the inverse spatial Fourier transform of the coefficients that survived the compression.

We are now ready to compare analog and digital communication, and discuss why the trend towards digital is inevitable.

1.1.3 Why digital?

On comparing the block diagrams for analog and digital communication in Figures 1.1 and 1.2, respectively, we see that the digital communication system involves far more processing. However, this is not an obstacle for modern transceiver design, due to the exponential increase in the computational power of low-cost silicon integrated circuits. Digital communication has the following key advantages.

Optimality For a point-to-point link, it is optimal to separately optimize source coding and channel coding, as long as we do not mind the delay and processing incurred in doing so. Owing to this *source–channel separation principle,* we can leverage the best available source codes and the best available channel codes in designing a digital communication system, independently of each other. Efficient source encoders must be highly specialized. For example, state-of-the-art speech encoders, video-compression algorithms, and text-compression algorithms are very different from each other, and are each the result of significant effort over many years by a large community of researchers. However, once source encoding has been performed, the coded modulation scheme used over the communication link can be engineered to transmit the information bits reliably, regardless of what kind of source they correspond to, with the bit rate limited only by the channel and transceiver characteristics. Thus, the design of a digital communication link is *source-independent* and *channel-optimized.* In contrast, the waveform transmitted in an analog communication system depends on the message signal, which is beyond the control of the link designer, hence we do not have the freedom to optimize link performance over all possible communication schemes. This is not just a theoretical observation: in practice, huge performance gains are obtained from switching from analog to digital communication.

Scalability While Figure 1.2 shows a single digital communication link between source encoder and decoder, under the source–channel-separation principle, there is nothing preventing us from inserting additional links, putting the source encoder and decoder at the end points. This is because digital communication allows *ideal regeneration* of the information bits, hence every time we add a link, we can focus on communicating reliably

over that particular link. (Of course, information bits do not always get through reliably, hence we typically add error-recovery mechanisms such as retransmission, at the level of an individual link or "end-to-end" over a sequence of links between the information source and sink.) Another consequence of the source–channel-separation principle is that, since information bits are transported without interpretation, the same link can be used to carry multiple kinds of messages. A particularly useful approach is to chop the information bits up into discrete chunks, or *packets*, which can then be processed independently on each link. These properties of digital communication are critical for enabling massively scalable, general-purpose, *communication networks* such as the Internet. Such networks can have large numbers of digital communication links, possibly with different characteristics, independently engineered to provide "bit pipes" that can support data rates. Messages of various kinds, after source encoding, are reduced to packets, and these packets are switched along different paths along the network, depending on the identities of the source and destination nodes, and the loads on different links in the network. None of this would be possible with analog communication: link performance in an analog communication system depends on message properties, and successive links incur noise accumulation, which limits the number of links which can be cascaded.

The preceding makes it clear that source–channel separation, and the associated bit-pipe abstraction, is crucial in the formation and growth of modern communication networks. However, there are some important caveats that are worth noting. Joint source–channel design can provide better performance in some settings, especially when there are constraints on delay or complexity, or if multiple users are being supported simultaneously on a given communication medium. In practice, this means that "local" violations of the separation principle (e.g., over a wireless last hop in a communication network) may be a useful design trick. Similarly, the bit-pipe abstraction used by network designers is too simplistic for the design of wireless networks at the edge of the Internet: physical properties of the wireless channel such as interference, multipath propagation, and mobility must be taken into account in network engineering.

1.1.4 Why analog design remains important

While we are interested in transporting bits in digital communication, the physical link over which these bits are sent is analog. Thus, analog and mixed-signal (digital/analog) design continue to play a crucial role in modern digital communication systems. Analog design of digital-to-analog converters, mixers, amplifiers, and antennas is required in order to translate bits to physical waveforms to be emitted by the transmitter. At the receiver, analog design of antennas, amplifiers, mixers, and analog-to-digital converters is required in order to translate the physical received waveforms to digital (discrete-valued, discrete-time) signals that are amenable to the digital signal processing that is at the core of modern transceivers. Analog circuit design for communications is therefore a thriving field in its own right, which this textbook makes no attempt to cover. However, the material in Chapter 3 on analog communication techniques is intended to introduce digital communication system designers to some of the high-level issues addressed by analog circuit designers.

The goal is to establish enough of a common language to facilitate interaction between system and circuit designers. While much of digital communication system design can be carried out by abstracting out the intervening analog design (as done in Chapters 4 through 8), closer interaction between system and circuit designers becomes increasingly important as we push the limits of communication systems, as briefly indicated in the epilogue.

1.2 A technology perspective

We now discuss some technology trends and concepts that have driven the astonishing growth in communication systems in the past two decades, and that are expected to impact future developments in this area. Our discussion is structured in terms of big technology "stories."

Technology story 1: the Internet Some of the key ingredients that contributed to its growth and the essential role it plays in our lives are as follows.

- Any kind of message can be chopped up into packets and routed across the network, using an Internet Protocol (IP) that is simple to implement in software.
- Advances in optical-fiber communication and high-speed digital hardware enable a super-fast "core" of routers connected by very high-speed, long-range links, that enable worldwide coverage;
- The World Wide Web, or web, makes it easy to organize information into interlinked hypertext documents, which can be browsed from anywhere in the world.
- The digitization of content (audio, video, books) means that ultimately "all" information is expected to be available on the web.
- Search engines enable us to efficiently search for this information.
- Connectivity applications such as email, teleconferencing, videoconferencing and online social networks have become indispensable in our daily lives.

Technology story 2: wireless Cellular mobile networks are everywhere, and are based on the breakthrough concept that ubiquitous tetherless connectivity can be provided by breaking the world into cells, with "spatial reuse" of precious spectrum resources in cells that are "far enough" apart. Base stations serve mobiles in their cells, and hand them off to adjacent base stations when the mobile moves to another cell. While cellular networks were invented to support voice calls for mobile users, today's mobile devices (e.g., "smart phones" and tablet computers) are actually powerful computers with displays large enough for users to consume video on the go. Thus, cellular networks must now support seamless access to the Internet. The billions of mobile devices in use easily outnumber desktop and laptop computers, so that the most important parts of the Internet today are arguably the cellular networks at its edge. Mobile service providers are having great difficulty keeping up with the increase in demand resulting from this convergence of cellular and Internet; by some estimates, the capacity of cellular networks must be scaled up by several orders of magnitude, at least in densely populated urban areas! As discussed in the epilogue, a major

challenge for the communication researcher and technologist, therefore, is to come up with the breakthroughs required to deliver such capacity gains.

Another major success in wireless is WiFi, a catchy term for a class of standardized wireless local-area network (WLAN) technologies based on the IEEE 802.11 family of standards. Currently, WiFi networks use unlicensed spectrum in the 2.4- and 5-GHz bands, and have come into widespread use in both residential and commercial environments. WiFi transceivers are now incorporated into almost every computer and mobile device. One way of alleviating the cellular capacity crunch that was just mentioned is to offload Internet access to the nearest WiFi network. Of course, since different WiFi networks are often controlled by different entities, seamless switching between cellular and WiFi is not always possible.

It is instructive to devote some thought to the contrast between cellular and WiFi technologies. Cellular transceivers and networks are far more tightly engineered. They employ spectrum that mobile operators pay a great deal of money to license, hence it is critical to use this spectrum efficiently. Furthermore, cellular networks must provide robust wide-area coverage in the face of rapid mobility (e.g., automobiles at highway speeds). In contrast, WiFi uses unlicensed (i.e., free!) spectrum, must provide only local coverage, and typically handles much slower mobility (e.g., pedestrian motion through a home or building). As a result, WiFi can be more loosely engineered than cellular. It is interesting to note that, despite the deployment of many uncoordinated WiFi networks in an unlicensed setting, WiFi typically provides acceptable performance, partly because the relatively large amount of unlicensed spectrum (especially in the 5-GHz band) allows channel switching on encountering excessive interference, and partly because of naturally occurring spatial reuse (WiFi networks that are "far enough" from each other do not interfere with each other). Of course, in densely populated urban environments with many independently deployed WiFi networks, the performance can deteriorate significantly, a phenomenon sometimes referred to as a tragedy of the commons (individually selfish behavior leading to poor utilization of a shared resource). As we briefly discuss in the epilogue, both the cellular and the WiFi design paradigms need to evolve to meet our future needs.

Technology story 3: Moore's law Moore's "law" is actually an empirical observation attributed to Gordon Moore, one of the founders of Intel Corporation. It can be paraphrased as saying that the density of transistors in an integrated circuit, and hence the amount of computation per unit cost, can be expected to increase exponentially over time. This observation has become a self-fulfilling prophecy, because it has been taken up by the semiconductor industry as a growth benchmark driving their technology roadmap. While the growth in density implied by Moore's law may be slowing down somewhat, it has already had a spectacular impact on the communications industry by drastically lowering the cost and increasing the speed of digital computation. By converting analog signals to the digital domain as soon as possible, advanced transceiver algorithms can be implemented in digital signal processing (DSP) using low-cost integrated circuits, so that research breakthroughs in coding and modulation can be quickly transitioned into products. This leads to economies of scale that have been critical to the growth of mass-market products in both wireless (e.g., cellular and WiFi) and wireline (e.g., cable modems and DSL) communication.