Introduction to Digital Communication

This chapter introduces some basic terminology, describes a model for communication systems and networks and describes some characteristics of data.
After this chapter you should be able to: define the terms introduced this chapter; compute information, entropy, bit, symbol, bit error and frame error rates; compute throughput; convert between characters, Unicode code points and their UTF-8 encodings; convert numbers between different number bases and bit and byte orders.

Model of a Communication System

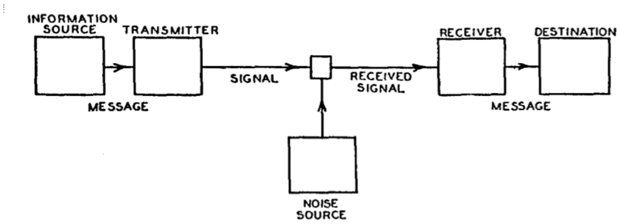


Fig. 1—Schematic diagram of a general communication system.

The diagram above shows a model for a communication system that includes the following[1]:

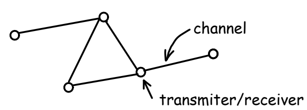
- information source - generates a sequence of "messages," taken from a limited set of possible values. A message might be a sequence of voltage levels that convey a sound or image. A message might also convey more abstract information called "data" which could represent, for example, the characters in a document or perhaps numbers whose meaning is unknown ("opaque") to the communication system
- transmitter - a device that converts the messages into a time-varying voltage or current (a "signal") that can be carried over the channel
- channel - carries the signal from the transmitter to the receiver. The channel often distorts the signal and adds a random signal to it called "noise"
- receiver - a device that attempts to recover the messages that were transmitted from the received signal
- data destination - (sometimes called a "sink") such as a person or computer that makes use of the information

[1]The diagram is from Claude Shannon's fundamental paper, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

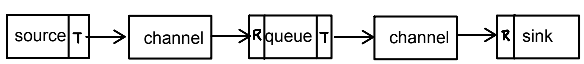
Exercise 1: Give an example of a communication system. If you can, identify the source, transmitter, channel, receiver and destination.

Networks

In many cases information travels over a network consisting of multiple channels and their associated transmitters and receivers:



Digital communication networks typically operate in a "store-and-forward" manner where messages are transmitted in pieces called "frames" or "packets" and each received frame is queued (stored) before being transmitted (forwarded) over another channel:



In this course we will study the basic principles underlying digital communication systems and networks.

Exercise 2: Give an example of a communication network. If you can, identify different the type(s) of channels used.

Overview

In this course we will try to answer the following fundamental questions about communication systems:

- How do we encode information into messages?
- How do we convert bits to signals? (line codes)

- What are some useful metric for a channel? (e.g. bandwidth, linearity, noise)
- How fast can we transmit data through a channel? (information rates and capacity)
- How often do errors happen? (error rates) How can we detect and recover from these errors? (error detection and correction)
- How do we group bits for transmission over a channel? (framing)
- How can several users share the same channel? (multiple access and multiplexing)

Digital Communications

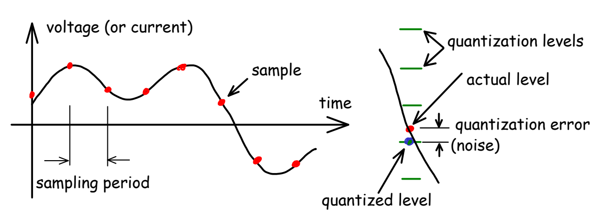
Digital communication systems communicate messages by encoding them as sequences of **b**inary dig**its** which are called bits.

Digital communication systems can be used to transmit digitized speech or video as well as more abstract information such as text or computer software, usually called "data".

The terms "data communications" and "digital communications" are often used interchangeably because today the same communication systems carry both data and digitized waveforms.

Sampled Waveforms

To represent speech or video waveforms as binary data, the analog signal – a voltage that is continuous in time and level – is sampled (measured) at a regular rate called the "sampling rate." Each sample is then quantized into a binary number with a fixed number of bits. Thus the signal becomes discrete in both time and level. In other courses you will learn why the sampling rate for a signal must be twice the highest bandwidth and that quantizing a signal to B bits introduces "quantization noise" with a power about $6B$ dB less than the signal power.



Exercise 3: Speech is intelligible if you restrict the sounds to frequencies below about 4 kHz. What is the minimum sampling rate that should be used to sample speech so that it will be intelligible?

A signal-to-noise power ratio of about 48 dB is considered "toll quality" (the SNR conventional telephone networks provide). How many bits of quantization are required to obtain a quantization SNR equivalent to "toll quality" speech?

What if the signal was a video signal with a 5 MHz bandwidth and required a quantization SNR of 40 dB?

What are the resulting bit rates in the two examples above ?

Words, Byte and Bit Orders

Bits are often organized into "words". Words of 8 bits are called bytes (or "octets" in many standards documents). Words composed of multiple bytes (2, 4 and 8 bytes) are also common.

It is important that communication systems preserve the order of the bytes and the order of the bits within the bytes.

Values stored in computer memory can only be addressed by byte. Thus byte order (but not bit order) is also a concern when multi-byte words are stored in memory or in files.

The bytes in a word can be ordered as the bytes are written, from "most significant byte" (MSB) to "least significant byte" (LSB). This is sometimes called "big endian" order. The reverse order is called "little endian". Most Internet protocols use "big endian" byte order which is sometimes called "network order."

Data is often transmitted over a communication system as individual bits. In this case the bits in a byte can be transmitted least-significant bit first (lsb-first) or most-significant-bit first (msb-first).

Many communication systems (e.g. Ethernet) transmit data with bytes ordered MSB-first but with the bits ordered lsb-first.

Exercise 4: Write the sequence of bits that would be transmitted if the 16-bit value 525 was transmitted with the bytes in little-endian order and the bits lsb-first. Write the sequence of bits that would be transmitted in "network order" and the bits msb-first.

Binary Numbers

When a sequence of bits represents a number, the number may be "unsigned" (non-negative) or signed, typically in two's-complement format. You should be familiar with these encodings from previous courses.

Binary numbers are often written in hexadecimal notation because it allows 8-bit values (bytes) to be written using two hexadecimal digits. Hexadecimal digits are `0` to `9` and `a` through `f` (representing values from 10 through 15). The prefix, "0x" as used in the C programming language, will be used to indicate hexadecimal notation.

Files

Data is often stored in files. A file is an ordered sequence of bytes. Files often have associated metadata such as a file name, length, date, type, etc. The bytes stored in a file can represent text or they may have other interpretation such as a computer program, sound or video.

Character Encodings

Data often represents printable characters or "glyphs".

A standard called Unicode assigns a unique number called a "code point" to over 100,000 of the characters used by more than 100 languages and scripts. Unicode is used by most operating systems and Internet applications.

Exercise 5: How many bits would be required to uniquely identify 100,000 different characters? (Hint: $2^{16} = 65536$).

We could representing each character using 32 bits (4 bytes). However, the UTF-8 (Unicode Transformation Format - 8 bits) format is more widely used since it can represent many documents using fewer bytes. It also has some practical advantages[2]

ASCII (American Standard Code for Information Interchange) was an earlier character encoding that used 7 bits to encode letters from the English alphabet, numbers and the most common punctuation symbols. The UTF-8 encoding is the same as the ASCII encoding for these first 127 characters. This means that ASCII-encoded documents are already encoded as UTF-8.

The table below shows the ASCII table which is the first code chart from the Unicode standard. The columns are labelled with the most significant (first) hex digit and the rows with the least-significant (second) hex digit of the numerical value of each character.

[2]The UTF-8 encoding has no zero bytes and all values <0x80 represent ASCII characters.

ASCII also includes some non-printable control codes (values 0 to 31) that were used to control printers. For example the line feed (LF) character would move the paper in the printer up one line.

Other Unicode characters require between 2 and 4 bytes according to the rules summarized in Table 3-6 of the Unicode standard shown above. Code points between 128 and 2047 include most characters from European languages and can be encoded in two bytes. Code points from 2048 to 65535 include most CJK (Chinese, Japanese and Korean) characters and require three bytes. Some rarely-used symbols (e.g. emoticons or Mahjong tiles) have four-byte encodings.

To encode a code point into UTF-8 format, first convert it to a binary number and find the first row in the table that can represent that number. Then copy the bits indicated by x, y, z and u from the binary number into the corresponding locations in the 1, 2 or 3 bytes of the UTF-8 encoding.

Exercise 6: The Chinese character for "Rice" (the grain) is 米 with Unicode value (code point) U+7C73. What is the UTF-8 encoding for this character?

Text versus Binary Number Representations

It's important to understand the difference between binary data and text that represents a number. For example, the number 49 is the ASCII encoding (and the Unicode code point) for the character '**1**'. This number can also be written as an 8-bit binary number 00110001 or in hexadecimal notation as 0x31.

Numbers can be stored in files or transmitted over communication systems in either binary format (e.g. one 8-bit value per byte) or in text format (as a sequence of numeric characters). Numbers in text format can be more easily interpreted by humans since they are sequences of printable characters but they may require more bits to transmit.

Exercise 7: Convert the decimal number 525 to a 16-bit (two-byte) binary number. How would you write this in hexadecimal notation?

Find the ASCII codes for the *characters* '525'. Write out the bits of the sequence that would be transmitted assuming each character is encoded in UTF-8. *Hint: the UTF-8 character code for a digit is 0x30 plus the value of the digit.*

Which of these two sequences of bits is the text format and which is the binary format? How many more bits would need to be stored or transmitted for the text format?

Table 3-6. UTF-8 Bit Distribution

Scalar Value	First Byte	Second Byte	Third Byte	Fourth Byte
00000000 0xxxxxxx	0xxxxxxx			
00000yyy yyxxxxxx	110yyyyy	10xxxxxx		
zzzzyyyy yyxxxxxx	1110zzzz	10yyyyyy	10xxxxxx	
000uuuuu zzzzyyyy yyxxxxxx	11110uuu	10uuzzzz	10yyyyyy	10xxxxxx

Characteristics of Data Sources, Channels and Sinks

Information Rate

We can model sources as generating one of a limited number of messages. For example, the messages might be letters, words, pixel values, or measurements. Different messages will often have different probabilities. The probability of a particular message is the fraction of messages of that type.

Exercise 8: We observe a source that outputs letters. Out of 10,000 letters 1200 were 'E'. What would be a reasonable estimate of the probability of the letter 'E'?

We define the information that is transmitted by a message that occurs with a probability P as:

$$I = -\log_2(P) \quad \text{bits}$$

For example, a message with a probability of $\frac{1}{2}$ conveys 1 bit of information. While one with a probability of $\frac{1}{4}$ carries 2 bits of information. Thus, less likely messages carry more information.

The information rate (also known as the "entropy") of a source in units of bits per message can be computed as the average information generated by the source:

$$H = \sum_i (-\log_2(P_i) \times P_i) \quad \text{bits/message}$$

where P_i is the probability of the i'th message.

Exercise 9: A source generates four different messages. The first three have probabilities 0.125, 0.125, 0.25. What is the probability of the fourth message? How much information is transmitted by each message? What is the entropy of the source? What is the average information rate if 100 messages are generated every second? What if there were four equally-likely messages?

We will see later in the course that there is a limit, called the "capacity," (C) for the information rate that can be transmitted over a given channel.

Data and Bit Rates

The rate at which bits, symbols or messages are generated by a particular system, independent of their probabilities, is the "data rate".

Data rates can be specified in different units. When the data rate is measured in bits per second (bps or b/s) it is also called the bit rate. Unfortunately, the same acronym is often used for "bytes per second". To avoid confusion it's best to spell out the units if the meaning is not clear from the context.

Some people use the convention that a capital 'B' indicates bytes and a lower-case 'b' indicates bits, but you should not assume this convention is universally understood.

In this course, and in almost all technical documents (equipment specifications, interoperability standards, data sheets, etc.), "bps" or "b/s" means "bits per second".

Computer storage is often measured in units with prefixes that are powers of 2 (e.g. kilo means 2^{10} or 1024 rather than 1000 or 10^3). However, communication system data rates always use units that are powers of 10 (e.g. 1 kb/s is 1000 b/s).

Exercise 10: How long will it take to transfer 1 MByte at a rate of 10 kb/s?

Symbol and Baud Rates

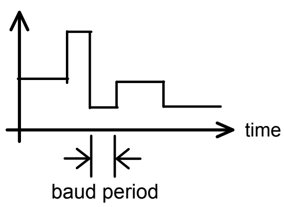
A signal is a voltage or current that varies with time; typically to convey information. The purpose of a transmitter is to convert data, in the form of bits, to a signal that carries information for transmission over a channel.

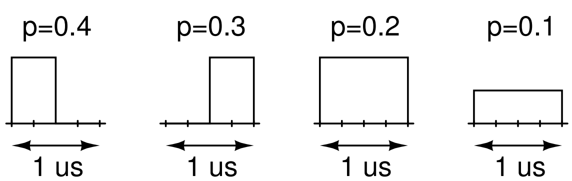
The IEEE's standards dictionary[3] defines a symbol as *"The smallest unit of data transmission on the medium."*

Often the transmitter uses a group of n bits to select from one of 2^n different "symbols", to transmit. The rate at which these signals are transmitted is called the "symbol rate". The symbol rate is equal to or lower than the bit rate(*why?*).

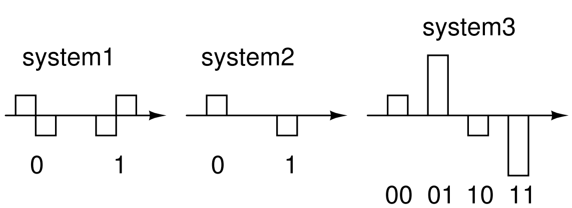
One possible set of signals is voltages of different levels. The maximum rate at which these levels can change is called the "baud rate[4]". The baud rate can be lower or higher than the bit rate depending on the design of the transmitter.

The IEEE dictionary defines a baud as *"A unit of signaling speed equal to the number of discrete conditions or signal events per second, or the reciprocal of the time of the shortest signal element in a character."*





Exercise 11: A communication system transmits one of the symbols above each microsecond. The probability of each symbol being transmitted is given above each symbol. What are the bit rate, the symbol rate, the information rate and the baud rate?



Exercise 12: Another system, as shown above, encodes each bit using two pulses of opposite polarity (H-L for 0 and L-H for 1). A second system encodes bits using one pulse per bit (H for 0 and L for 1). A third system encodes two bits per pulse by using four different pulse levels (-3V for 00, -1V for 01, +1V for 10 and +3V for 11). Assuming each system transmits at 1000 bits per second, what are the baud rates in each case? How many different symbols are used by each system? What are the symbol rates? Assuming each symbol is equally likely, what are the information rates?

[3] *The Authoritative Dictionary of IEEE Standards Terms*, Seventh Edition, IEEE, 2000.

[4] This is the definition found in the IEEE LAN standards. However, some people consider baud rate to be a synonym for symbol rate.

Error Rates

The bit error rate (BER, P_e) is the average fraction of bits that are received incorrectly.

When these bits are grouped into "frames" we are often interested in the average fraction of the frames that contain one or more errors. This is known as the FER (Frame Error Rate). Sometimes frames include additional bits that allow us to detect most, but not all, errors. We usually want the UEP (Undetected Error Probability) to be very small (e.g. one undetected error per many years).

Exercise 13: You receive 1 million frames, each of which contains 100 bits. By comparing the received frames to the transmitted ones you find that 56 frames had errors. Of these, 40 frames had one bit in error, 15 had two bit errors and one had three errors. What was the FER? The BER?

Throughput

The average rate at which data arrives at the destination can be different than the instantaneous rate at which the transmitter sends data to the channel because:

- the channel may have to be shared between different users
- the transmitter may add (and the receiver remove) "overhead" bits for addressing, error detection, etc
- incorrectly received data may have to be retransmitted
- there may be gaps between frames

The long-term average data rate arriving at the destination is called the "throughput". It can be computed by dividing the total amount of data delivered to the sink by the elapsed time. Each type of communication system involves different overheads and thus requires a different throughput computation.

Exercise 14: A system transmits data at an (instantaneous) rate of 1 Mb/s in frames of 256 bytes. 200 of these bytes are data and the rest are overhead. The time available for transmission over the channel is shared equally between four users. A 200 μs gap must be left between each packet. What throughput does each user see? Now assume 10% of the frames are lost due to errors. What is the new throughput per user?

Compression

When data is not random and we can make use of the redundancy to reduce the amount of data that needs to be transmitted. Both lossless and lossy compression are examples of "source coding."

Lossless. Some types of data contains redundancy such as sequences of bits or bytes that occur more often than others. This type of data can be compressed before transmission and then decompressed at the receiver without loss of information. An example of this "lossless" compression is the 'zip' compression used for computer files.

Another definition of information rate is "the minimum data rate, assuming the best possible lossless compression". Lossless compression does not reduce the information rate but it may reduce the bit rate.

Lossy. Data representing speech and video can often be compressed with little degradation because humans cannot perceive certain details of sounds and images. These details can be removed resulting in lower data rates. Examples of these "lossy" compression techniques include "MP3" for compressing audio and MPEG-4 for video.

Data Rate Variability

The data rate of a source can be:

- constant: "isochronous" sources generate data at a constant bit rate (CBR) and are typical of regularly sampled waveforms such as (uncompressed) audio or video sources.

- variable: variable bit rate (VBR) sources are typical of compressed speech and video because different parts of the speech or video signal have different amounts of redundancy and can be compressed to different bit rates.

- bursty: bursty data sources generate large amounts of information at instants of time in-between pauses where no information is generated. This is typical of systems involving human-machine interaction such as web surfing.

Exercise 15: Plot some sample data rate versus time curves for these three types of sources. Can you think of some characteristics of a video source that might result in a variable bit rate when it is compressed? (*Hint: what types of redundancy are there in video?*) .

Tolerance To Impairments

Data sinks vary in their tolerance to channel impairments such as errors, delay and variability of delay (delay "jitter").

For example, computer data transmission systems usually require very low undetected error rates (e.g. one undetected error in decades) but can often tolerate high delay and delay variability (seconds). On the other hand telephone systems can tolerate loss of a small percentage of the speech waveform but become difficult to use if delays exceed a significant fraction of a second.

Exercise 16: For each of the following communication systems identify the tolerance it is likely to have to errors and delay: a phone call between two people, "texting", downloading a computer program, streaming a video over a computer network. What do you think might be the maximum tolerable delay for each?

Exercise 17: Highlight or underline each term where it is defined in these lecture notes.

	000	001	002	003	004	005	006	007
0	NUL 0000	DLE 0010	SP 0020	0 0030	@ 0040	P 0050	` 0060	p 0070
1	SOH 0001	DC1 0011	! 0021	1 0031	A 0041	Q 0051	a 0061	q 0071
2	STX 0002	DC2 0012	" 0022	2 0032	B 0042	R 0052	b 0062	r 0072
3	ETX 0003	DC3 0013	# 0023	3 0033	C 0043	S 0053	c 0063	s 0073
4	EOT 0004	DC4 0014	$ 0024	4 0034	D 0044	T 0054	d 0064	t 0074
5	ENQ 0005	NAK 0015	% 0025	5 0035	E 0045	U 0055	e 0065	u 0075
6	ACK 0006	SYN 0016	& 0026	6 0036	F 0046	V 0056	f 0066	v 0076
7	BEL 0007	ETB 0017	' 0027	7 0037	G 0047	W 0057	g 0067	w 0077
8	BS 0008	CAN 0018	(0028	8 0038	H 0048	X 0058	h 0068	x 0078
9	HT 0009	EM 0019	) 0029	9 0039	I 0049	Y 0059	i 0069	y 0079
A	LF 000A	SUB 001A	* 002A	: 003A	J 004A	Z 005A	j 006A	z 007A
B	VT 000B	ESC 001B	+ 002B	; 003B	K 004B	[005B	k 006B	{ 007B
C	FF 000C	FS 001C	, 002C	< 003C	L 004C	\ 005C	l 006C	| 007C
D	CR 000D	GS 001D	- 002D	= 003D	M 004D	] 005D	m 006D	} 007D
E	SO 000E	RS 001E	. 002E	> 003E	N 004E	^ 005E	n 006E	~ 007E
F	SI 000F	US 001F	/ 002F	? 003F	O 004F	_ 005F	o 006F	DEL 007F