# Introduction to Data Communication

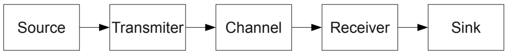
*This lecture introduces some basic terminology, describes a simple model of a communication link and a network and describes some characteristics of data.*

*After this lecture you should be able to: define the terms introduced this lecture, be able to convert numbers between different number bases and bit/byte orders and convert between ASCII characters and their character codes.*

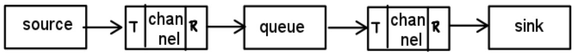
## Communication Systems Model

A common model for a communication system includes the following:

- data source - generates the information
- transmitter - converts the information into an electrical signal that can be transmitted over the channel
- channel - distorts the signal and adds noise
- receiver - attempts to recover the information that was transmitted
- data sink - accepts the information



In many cases information travels over a network consisting of multiple channels and their associated transmitters and receivers. In some cases the information is stored ("queued") before being forwarded over the next channel.



In this course we will study the implementation of digital communication systems and networks in detail.

**Exercise 1**: For each of the following digital communication services identify the source, sink and the channel(s) involved: the Ethernet connection between a computer and a router; a cell phone call ; watching a YouTube video at home.

## Digital and Data Communications

Digital communication systems communicate information that is in the form of digits, almost always binary digits or "bits".

Digital communication systems include those used to transmit digitized speech or video waveforms as well more abstract information such as text, images, or computer software (often called "data").

However, today the terms "data communications" and "digital communications" are often used interchangeably because the same networks are used to carry all types of digital information.

## Characteristics of Data Sources and Sinks

Data sources are often compressed to reduce the amount of data that needs to be transmitted.

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.

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

Both lossless and lossy compression are examples of "source coding".

One characteristic of a data source is whether the rate at which the data is generated is:

- constant - constant bit rate (CBR) or "isochronous" sources are typical of regularly sampled waveforms such as uncompressed audio

- variable - variable bit rate (VBR) sources are typical of compressed speech and video
- bursty - bursty data sources are typical of systems involving human interaction such as web surfing

**Exercise 2**: What features of speech and video waveforms might result in variable bit rates when these sources are compressed?

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

For example, computer data transmission systems usually provide very low undetected error rates (e.g. once per several years) 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 3**: What units would be used to specify error rate, delay, and delay variability? For each of the following data sources/sinks identify the relative data rate variability and the tolerance it is likely to have to errors, to the absolute delay and to the delay variability: a phone call between two people, downloading a computer program, streaming a video over a computer network. Try to guess typical values.

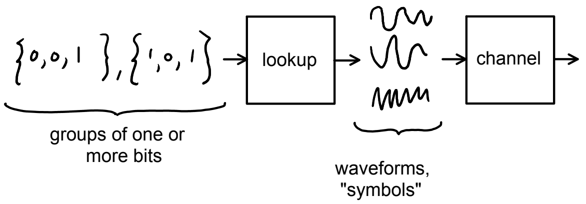
## Data Rates

The rate at which the source generates data, the "data rate", is specified in bits per second (bps or b/s).

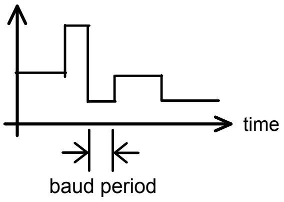
Unfortunately, particularly in consumer-facing data communications applications, the same acronym is used for "bytes per second". To avoid confusion it's best to spell out the units if the meaning is not clear from the context. In this course "bps" will always mean "bits per second". This is the convention used in almost all technical documents (equipment specifications, interoperability standards, data sheets, etc).

Sometimes the data is not completely random and can be compressed. The theoretical minimum data rate, assuming the best possible compression has been applied, is called the "information rate". We will see later on in the course that there are certain limits to the information rate that can be transmitted over a given channel.

A waveform is a voltage or current that varies with time. The transmitter converts the data to a signal, a waveform that carries information, for transmission over the channel. Often the transmitter uses a number of bits to select from one of several possible waveforms, or "symbols", to transmit. The rate at which these waveforms are transmitted is called the "symbol rate". The symbol rate is equal to or lower than the bit rate(*why?*).



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



## Throughput

In addition to the data rate at the source, we can measure the average data rate at the receiver. This is called the "throughput". This can be different than the rate at the transmitter because the transmitter may need to:

- share the channel with other transmitters
- add "overhead" bits for addressing, error detection, etc
- retransmit data that was not received correctly

The term "goodput" is sometimes used to refer to the throughput as measured by a computer application.

**Exercise 4**: A system transmits data at a rate of 1 Mb/s in "packets" of 128 bytes. 100 of these are bytes data and the rest are overhead. The channel is shared between four users. There is a 10 $\mu$s gap between each packet. 10% of the frames are lost due to errors. What throughput does each user see?

## Bit and Byte Order

The bits generated by a data source are usually organized into "words". Words of 8 bits are called bytes (or "octets" in some standards documents). Words of 4 bits are often called nibbles (or nybbles). Words composed of other even multiples of 8 bits (16, 32, 48 and 64 bits) are also common.

It is important that the communication system preserve the order of the bits and bytes between the source and sink.

If the bits in a word represent a binary number they can be ordered from "most significant bit" (MSB) to "least significant bit" (LSB). This is sometimes called "big endian" order. The reverse order is called "little endian".

Often the bits (in either bit order) are part of bytes which themselves are concatenated to form words. The bytes in each word can also be ordered MSB(yte) first (big-endian) or LSB(yte) first (little endian).

Most Internet protocols use big-endian bit and byte order which is sometimes called "network order". In network order the bits and bytes are transmitted in the order they are written.

**Exercise 5**: Convert the decimal number 3525 to a 16-bit (two-byte) binary number. Write the sequence of bits that would be transmitted if both the bytes and bits were transmitted in little-endian order. Write the sequence of bits that would be transmitted in "network order".

## Notation

When collections of bits are interpreted as numbers they can be written using digits from various bases.

The most common notation is hexadecimal 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). Typically a special prefix of non-numeric character(s) is used to indicate that the number is written in binary, octal or hexadecimal notation. Typical prefixes for hexadecimal notation include "$", "#", "0x", and "0H".

Octal numbers (representing 3 bits per digits) are occasionally used as well.

**Exercise 6**: Write the 16-bit number above in hexadecimal notation.

## Character Codes

Data can also represent printable characters ("glyphs").

A standard, called Unicode, has been developed to assign a unique number (called a "code point") to over 100,000 of the characters used by over 100 different languages and scripts. Unicode is used by most modern operating systems and most Internet applications.

**Exercise 7**: How many bits would be required to uniquely identify 100,000 different characters? (Hint: $2^{16} = 65536$).

The first 127 characters of Unicode correspond to an earlier character encoding called ASCII (American Standard Code for Information Interchange).

16 bits is insufficient for representing all of the Unicode characters and even if it were, it would double the size of typical text files and web pages that primarily use ASCII characaters. UTF-8 (Unicode Transformation Format - 8 bits) was developed as a way to represent typical Unicode documents in an efficient way. The UTF-8 encoding is the same as the ASCII encoding for the first 127 characters. This makes existing ASCII files compatible with software that expects UTF-8 encoded Unicode. Other Unicode characters require between 2 and 4 bytes according to the rules summarized in Table 3-6 of the Unicode standard shown above.

ASCII also includes some non-printable control codes that will be discussed later in the course.

The next page shows the ASCII table as given in version 6.2 of 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.

**Exercise 8**: Find the ASCII codes for the *characters* '3525'. Write out the first 16 bits of the sequence that would be transmitted assuming each character is encoded using 8 bits per character

**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 |

and little-endian bit order. *Hint: the character code for a digit is 0x30 plus the value of the digit.*

**Exercise 9**: 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?

**Exercise 10**: Highlight or underline each term where it is defined in these lecture notes.

## The Telecom Ecosystem

The following are the main types of companies and organizations involved in the data communications industry:

**Component Manufacturers** make components such as ICs, capacitors, cables, etc

**Distributors** buy components from manufacturers

**Equipment Manufacturers** buy components and build equipment

**Retailers** buy equipment and sell to customers

**Service Providers** buy equipment and provide services to customers

**Standards Organizations** write and publish standards so that equipment from different manufacturers will interoperate

**Regulators** publish and enforce regulations to ensure efficient use of public resources and protect the safety of the public

**Exercise 11**: Draw a diagram showing the flow of of goods and services between these various entities. Classify the following: Intel, Xilinx, Analog Devices, Belden, Avnet, DigiKey, Samsung, Apple, Cisco, Ericsson, Telus, Netflix, Walmart, Amazon, IEEE-SA, IETF, Industry Canada, CRTC, FCC. Look these up if you’re not familiar with them.

# C0 Controls and Basic Latin

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