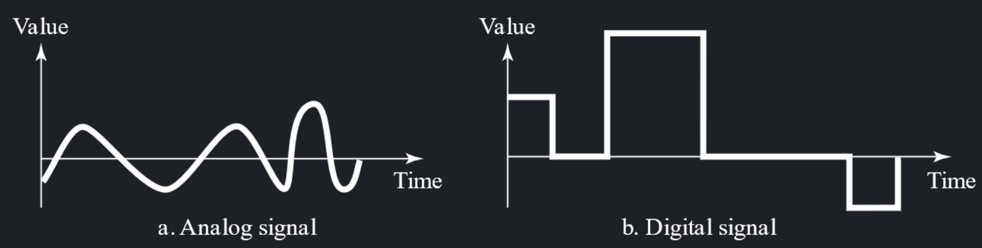
Chapter 3: Data and Signals

Table of Contents

3.1 Analogue and Digital

The difference between analogue and digital data is that analogue data had infinitely many levels while digital data has discrete values. Consider an analogue clock. It has a continuous range of values, even between the time that we actually read. A digital clock on the other hand has a discrete set of values and it switches from one value to another without going over anything in between. Another example of analogue data is the human voice, which we can convert to an analogue signal.

For analogue and digital signals, the same basic concept applies. Since an analogue signal passes through infinitely many values within its lower and upper bounds, we can get infinitely many values for its amplitude depending on where we test it. The digital signal on the other hand, can only give us some specific values.

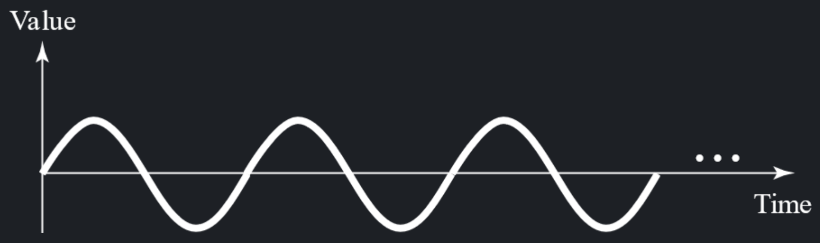


Periodic and Aperiodic Signals

Periodic signals follow a repetitive pattern. This means that over some period of time, the signal will follow the same pattern repeatedly. Aperiodic signals on the other hand do not follow a repetitive pattern.

Both analogue and digital signals can be periodic or aperiodic. For periodic signals, the completion of one full pattern is called a cycle. The concepts of frequency and time period also come into play here. In digital data communications, we use two types of signals, periodic analogue and aperiodic digital signals.

3.2 Periodic Analogue Signals

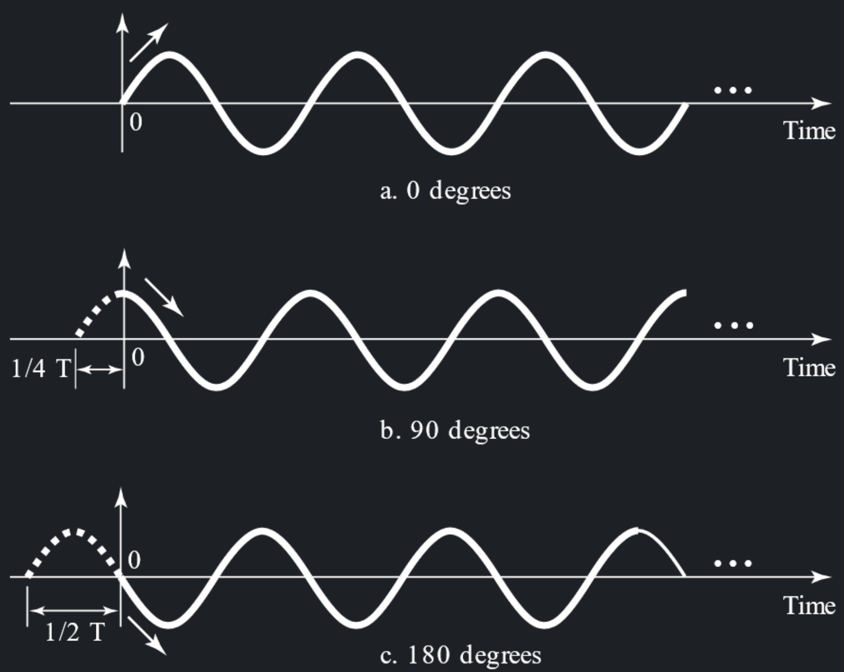


This is a normal sine wave. One notable feature is that we cannot make any comments on the frequency of the wave, since the time period is not given. If we assume the end of the wave to be at 1s, then the frequency would be $3\ \text{Hz}$.

The equation for a sine wave is $y = A\sin(\omega t + \theta)$. For now, we are assuming the phase change is $0°$. ω is the angular velocity, given by $\omega = 2\pi f$. Thus, to represent a sine wave, we need the frequency, the peak amplitude and the phase change.

When we see something like a battery with a voltage of 1.5V, it essentially means that the peak amplitude is 1.5V.

The effect of phase changes can be understood from the diagrams below.

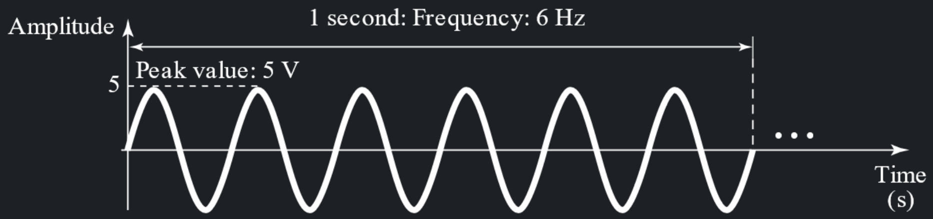


Phase changes will be important when we study modulation. In this case, it is important to note that it is not meaningful to think of phase changes as a shift of a signal to the left, since that would mean the start of the signal is being shifted to negative time. Instead, think of it as our observation time starting later.

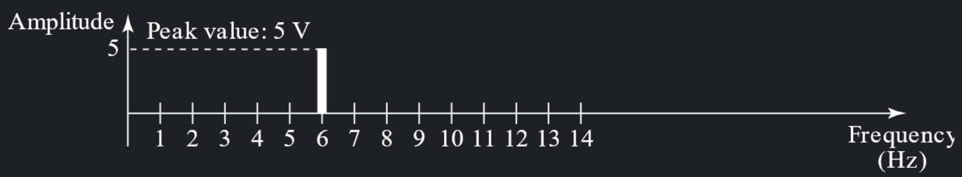
Another property of periodic analogue signals is the wavelength, which is the only property dependant on the transmission medium. $\lambda = \frac{c}{f}$ where c is the propagation speed of the wave.

Time and Frequency Domains

The time and frequency domains are simply different ways to represent a signal. In the time domain, the signal is shown as the value of the instantaneous amplitude at each time. In the frequency domain, the signal is shown as the maximum amplitude achieved at a particular frequency.



a. A sine wave in the time domain (peak value: 5 V, frequency: 6 Hz)



b. The same sine wave in the frequency domain (peak value: 5 V, frequency: 6 Hz

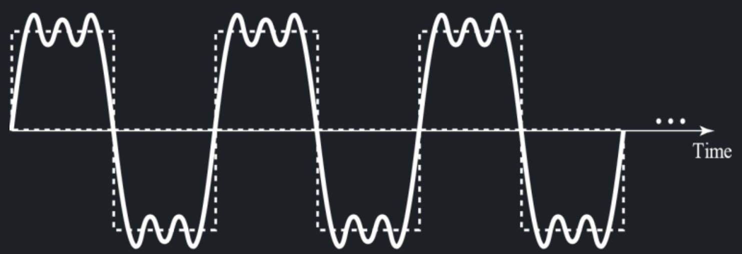
Composite Signals

In data communications, what we actually use are composite signals. A single frequency signal is not very useful to us, so we combine multiple single frequency signals to create a composite signal. On the other hand, given a composite signal, there are many tools that allow us to retrieve the original signals again, one of which is called Fourier analysis. We do not need to know the details of Fourier analysis.

Composite signals can be periodic or aperiodic. For periodic composite signals, when we decompose it, we get a series of signals with discrete frequencies. For aperiodic composite signals, the decomposition gives us a series of sine waves with continuous frequencies.

a. Time-domain decomposition of a composite signal

b. Frequency-domain decomposition of the composite signal



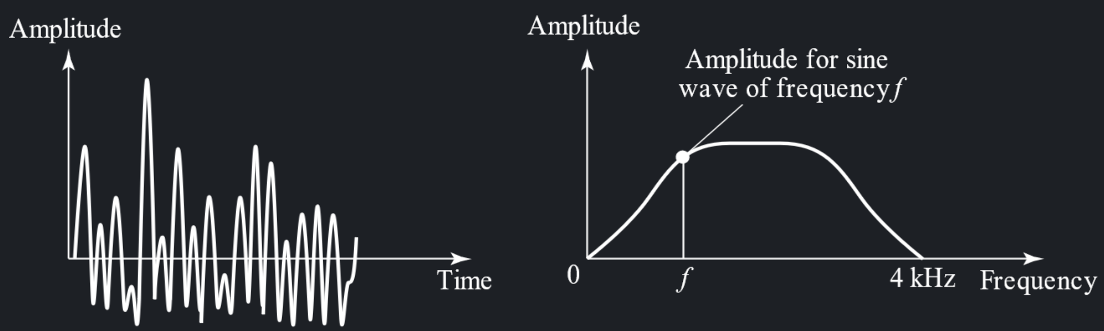
Another point to note is that it is possible to create a digital signal with a squarish shape by combining infinitely many sine waves. In the diagram above, three different sine waves have been combined (identifiable by the three smaller peaks at the peaks of the wave), and a tentative digital signal has been drawn with dotted lines. Getting extremely close to an actual digital signal is extremely difficult, but not impossible. Adding more signals would make the generated digital signal more accurate. However, this can only be maintained over short distances, for example in the bus of a CPU. This is called baseband communication. For longer distances, where broadband communication is used, it is not possible to maintain the shape.

In reality, when we take input of some analogue signal like the human voice, we first convert it to a digital signal. During transmission, this digital signal is again converted to an analogue signal. The reason we need to go through this weird process is because our voices actual work with frequencies that cannot be transmitted, since they are not supported by the transmission medium. This technique is called broadband communication.

The frequency of the general shape we see above is the same as that of one of the sine waves that create it (frequency f in the top diagram). That frequency is called the fundamental frequency or the first harmonic of the composite signal.

The small changes in amplitude at the peaks of the composite signal are simple enough to understand, since they are the result of adding up the values of all three curves from the top diagram at that point of time.

For aperiodic composite signals, the graph for the signal in the frequency domain looks like this:



a. Time domain   b. Frequency domain

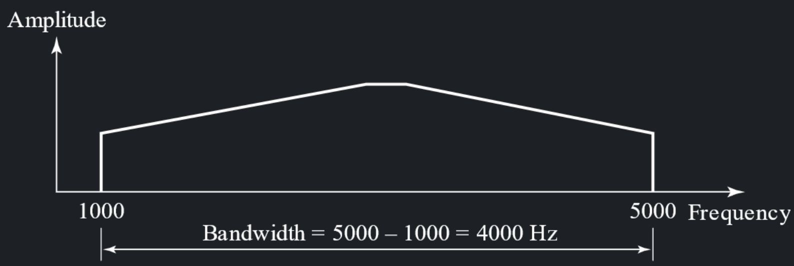
Frequency Spectrum and Bandwidth

The frequency spectrum of a signal is the combination of all the sine wave signals that make that signal.

Bandwidth is the difference between the highest and the lowest frequencies in the signal, or the width of the frequency spectrum.

The two diagrams below show the frequency spectrums for a periodic and an aperiodic signal in the frequency spectrum. The bandwidths are also marked.

a. Bandwidth of a periodic signal

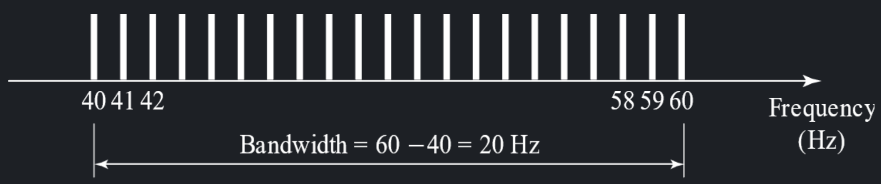


b. Bandwidth of a nonperiodic signal

Example

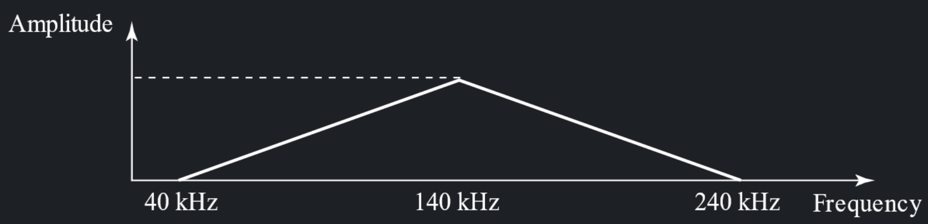
A periodic signal has a bandwidth of $20\ \mathrm{Hz}$. The highest frequency is $60\ \mathrm{Hz}$. What is the lowest frequency? Draw the spectrum if the signal contains all frequencies of the same amplitude.

Lowest Frequency $= 60 - 20 = 40\ \mathrm{Hz}$



Example

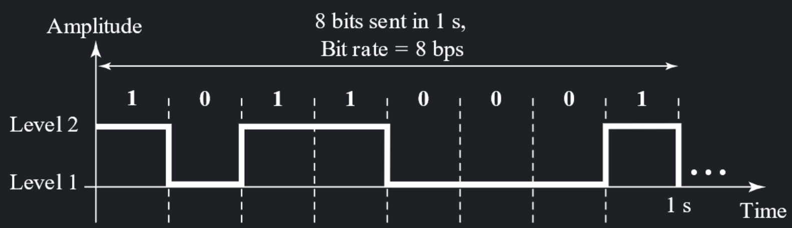
A nonperiodic composite signal has a bandwidth of $200\ \mathrm{kHz}$, with a middle frequency of $140\ \mathrm{kHz}$ and peak amplitude of $20\ \mathrm{V}$. The two extreme frequencies have an amplitude of 0. Draw the frequency domain of the signal.



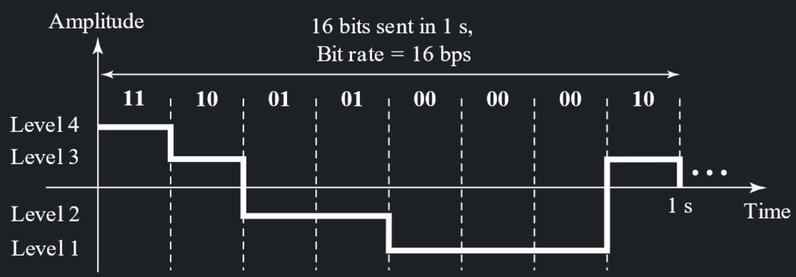
3.3 Digital Signals

A digital signal is a composite analogue signal with an infinite bandwidth. Recalling the discussion in the previous section about how a digital signal can be mimicked with a large number of analogue signals, if we have an infinite number of analogue signals with an infinite range of frequencies, we can accurately depict a digital signal.

Digital signals are composed of just a few discrete levels.



a. A digital signal with two levels



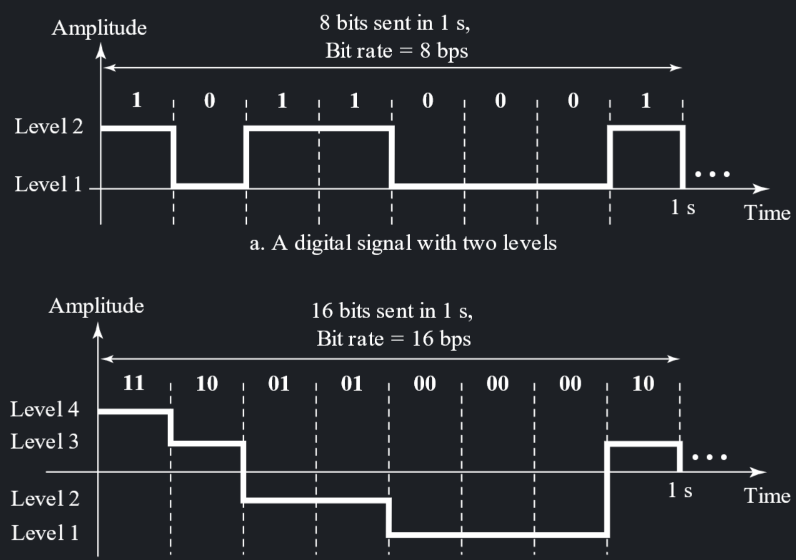
b. A digital signal with four levels

A fundamental difference between the two signals above is that they use a different number of levels. With 2 levels, a single bit could be represented, but with 4 levels, 2-bit data could be represented. Thus, with more levels, we can send more data. However, this will also make it more difficult for the receiver to accurately decode the data, so we would need more complicated devices.

The number of bits that can be represented with n levels is given by $\log_2 n$.

With digital signals, we use bit lengths instead of wave lengths, but the two are similar. The bit length is also the product of the propagation speed and bit duration.

We have previously discussed how, in digital communications, aperiodic digital signals are used. Consider the graph of a digital signal.



Here, notice that we maintain the same level for a certain period of time. For that period of time, while the level does not change, the frequency is simply 0. At the point where levels change, the frequency suddenly changes to another value, meaning the frequency is infinity. Essentially, we have a bandwidth between 0 and infinity.

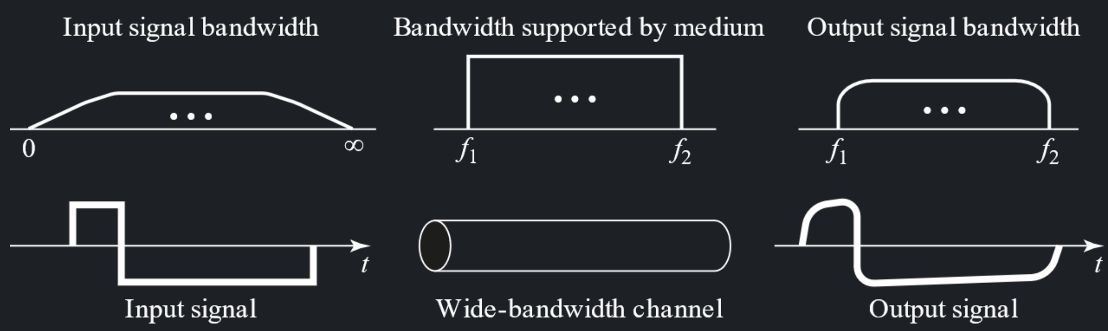
Transmission of Digital Signals

Digital signals have an infinite bandwidth, which means to transmit a digital signal, we need a transmission medium that has an infinite bandwidth. This is of course, not possible in real life. However, we can get very close. Some of the unsupported frequencies are lost, which means the exact square shape of a digital signal will not be maintained, but the signal can be more or less be recovered.

There are two techniques to transmitting digital signals – broadband and baseband. Baseband does not convert the digital signal to an analogue signal, but broadband converts it to a composite analogue signal with infinite bandwidth, a process called modulation, using a modem. Baseband communication is only possible over short distances using a dedicated channel, such as a computer bus.

Baseband Transmission

Baseband communication makes use of a low-pass channel, which allows frequencies of $0\mathrm{Hz}$ or very close to $0\mathrm{Hz}$. This requirement is set so as to allow very low frequency signals through. Transmission mediums commonly available in the market do not support this feature. Instead, they are bandpass channels. The baseband channel can have a wide bandwidth, which would allow a better shape, or a narrow bandwidth, which would require us to approximate the shape at the receiver side.

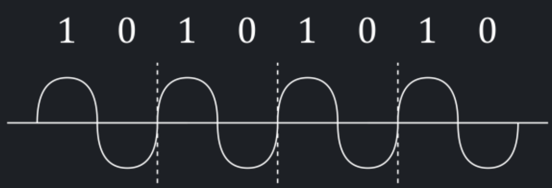


Essentially, frequencies from f_1 to f_2 are allowed by the channel in the figure above. This results in some data at the ends of the input signal with very low or very high frequency to be lost, giving the output signal the curved shaped. This distortion is entirely ignorable.

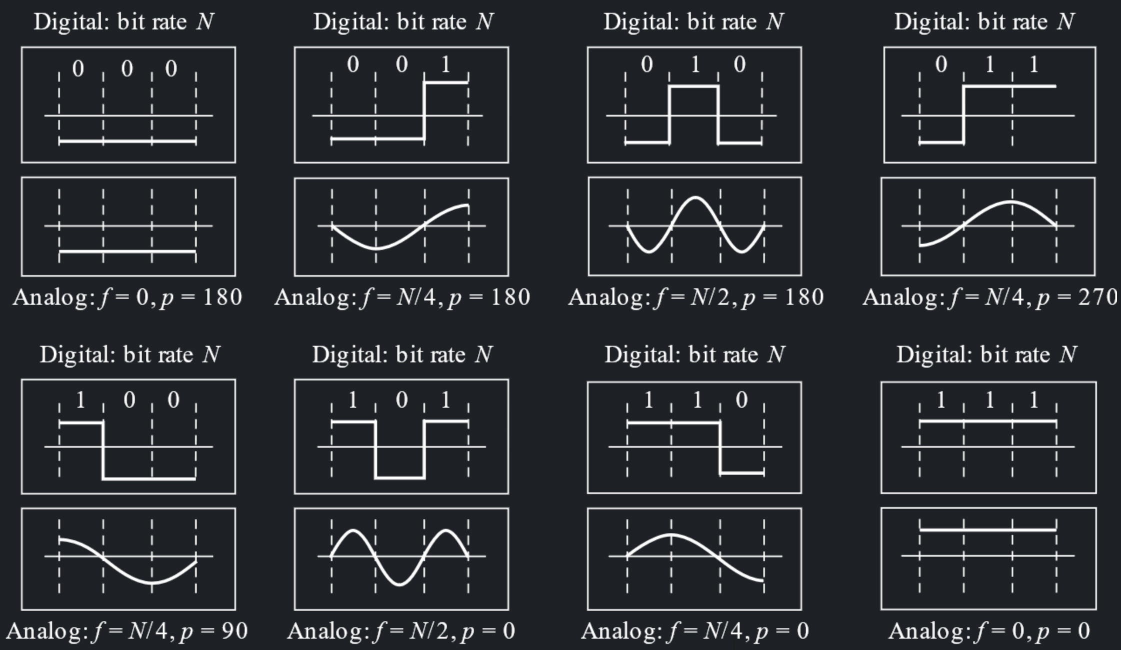
Now consider a scenario where we have a low-pass channel with a limited bandwidth. We need to know what the minimum bandwidth we need for the signal to be able to pass and maintain a recognizable shape.

Consider this figure:

Clearly, with a single cycle, we can represent two bits of data, a 1 and a 0. In the worst case, our digital data will be repeatedly fluctuating, like 10101010. In this case, if we consider that the digital data rate is N, meaning N bits are sent every time unit, the analogue signal needs to have a frequency of $\frac{\mathrm{N}}{2}$. For a data rate of $10\mathrm{kbps}$, we need a minimum bandwidth of $5\mathrm{kHz}$ to obtain a very rough approximation of the original signal. This is the first harmonic.



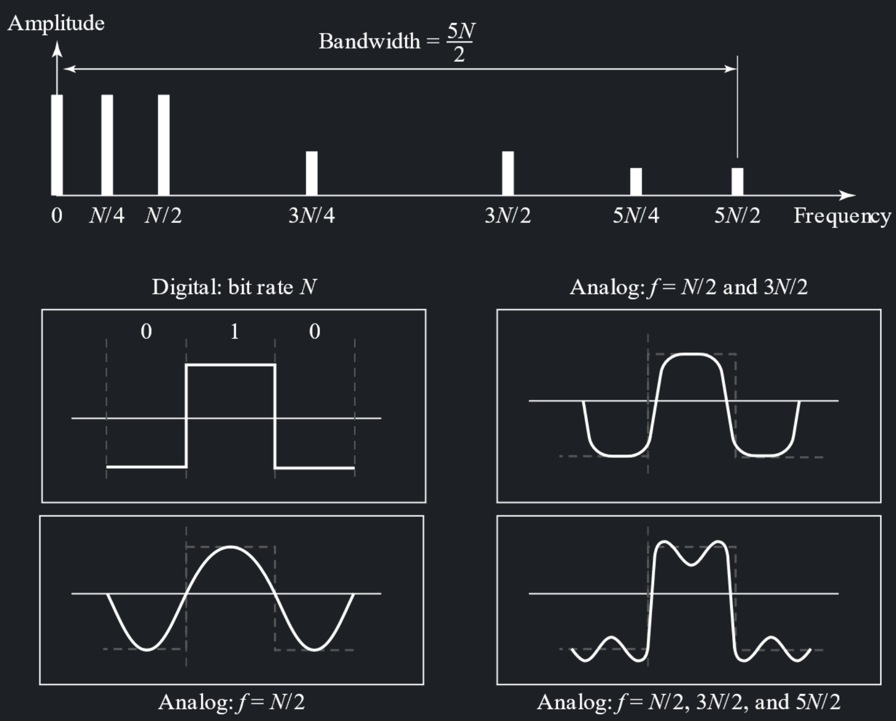
The frequency of this signal is the maximum signal that we need to support. In any other case, the signal will have a lower frequency.



In the above diagram, 3 data bits are sent per time period. Other than the two worst cases, there are four cases where the analogue signal can have a frequency of $\frac{N}{4}$, since the first two or last two bits are the same and can thus be represented using a single half of the analogue signal.

In the best case, the frequency is 0, since the three data bits are the same.

The reason larger bandwidths help is because we can use multiples of the first harmonic to get the signal closer to the original digital signal.



Broadband Transmission

If the channel we have available is a bandpass channel, then we cannot use it to directly send digital signals. This is what we have to deal with in our day-to-day lives. Bandpass means only a specific frequency band is supported, which means some digital to analogue conversion is needed. This conversion process is called modulation.

3.4 Transmission Impairment

Transmission media are not perfect because of impairments in the signal sent through the medium, i.e. the signals at the beginning and end of the medium are not the same. This impairment is mainly caused by three factors – attenuation, distortion and, most importantly, noise.

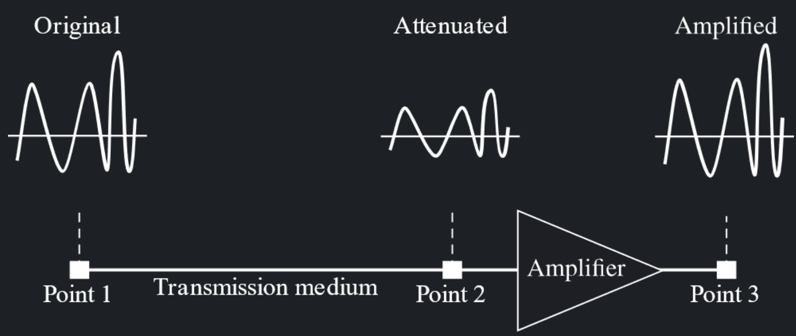
Attenuation

Attenuation refers to loss of energy. When signals are travelling for a long time, it loses energy to resistance, thus becoming weaker. To resolve this, we can use amplifiers to amplify the signal. The amplification process may be done with an amplifier or a repeater. A repeater only amplifies the signal, while an amplifiers amplifies the signal as well as any noise in the signal.

Attenuation and amplification are measured in decibels as $1\text{dB} = 10\log_{10}\left(\frac{p_2}{p_1}\right)$, where p_2 is the output power and p_1 is the input power. A positive value in decibels indicates amplification, while a negative value indicates attenuation.

$$p_2 = \frac{1}{2}p_1 \qquad \rightarrow \qquad 10\log_{10}\frac{0.5p_1}{p_1} = -3\text{dB}$$

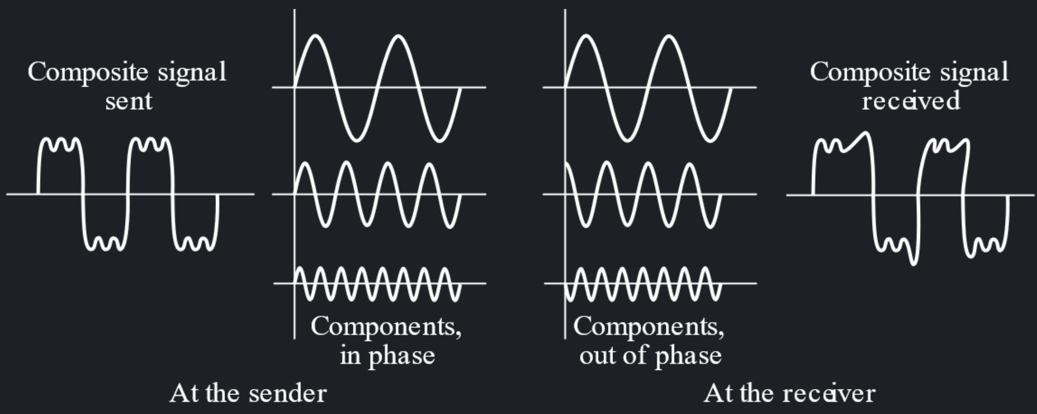
Thus, a 3dB attenuation means a loss of half the power.



Distortion

Distortion refers to a signal changing its form or shape. Distortion is also sometimes called delay distortion. This is because the distortion occurs due to a change in phase.

Consider the figure below:



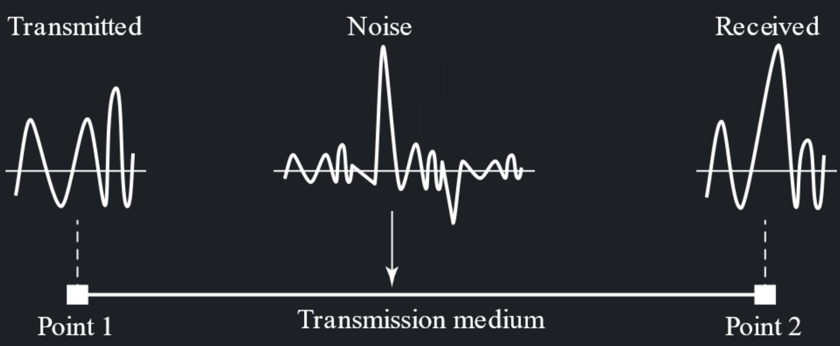
Due to the different propagation speeds of different frequencies, it is possible that the signals that make up a composite signal become out of phase. This makes the output composite signal distorted.

Noise

Noise is the addition of some unwanted elements to the signal being transmitted. Ideally, a noiseless transmission is preferred, but this is never practically possible.

There are a few types of noise. These include:

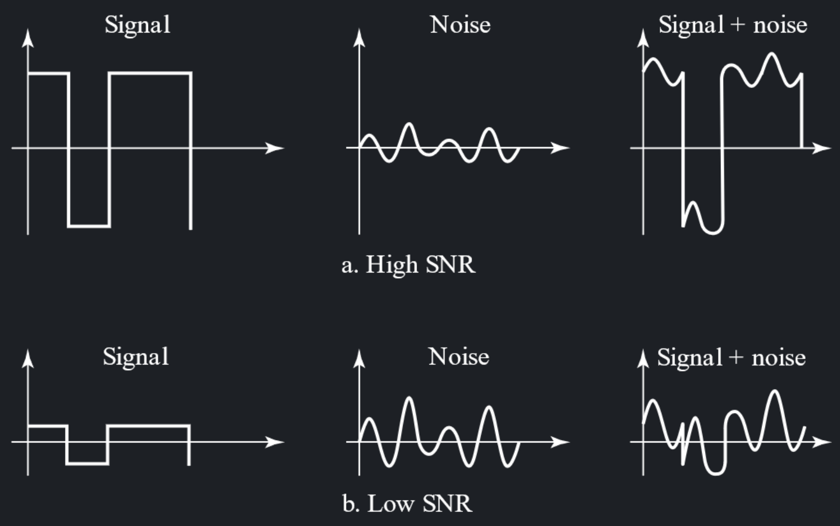
- Thermal Noise – Some noise is caused simply due to the movement of electrons.
- Induced Noise – These come from motors or appliances working in the system.
- Crosstalk – This noise is caused by information from one connection getting mixed with that in another.
- Impulse Noise – These are spikes caused by power lines, lighting, etc.



Since noise is unavoidable, the best we can do is limit its effects. It is not acceptable for a signal to be heavily affected by noise, to an extent that the original signal can no longer be retrieved.

The signal to noise ratio is called SNR. It allows us to measure how badly our signal is affected by noise. SNR is calculated as $\text{average signal power} / \text{average noise power}$. It can be converted to a decibel value as $\text{SNR}_{\text{dB}} = 10\log_{10}\text{SNR}$. For a theoretical noiseless signal, the value of SNR and SNR_{dB} is ∞.

A lower SNR indicates that noise is becoming stronger than the signal. This can result in the output signal being too distorted to properly read. A higher SNR indicates that the signal is stronger than the noise, which means it might still be readable.



There is also something called a signal to interference and noise ratio, or SINR. This deals with both noise and interference. Interference refers to any unwanted signals produced by the system that makes the original signal weaker. It is different from noise in that noise is a fixed, constant distortion of the signal while interference is a variable distortion. For example, the signal of a phone can face interference when it is working near sources of signals that have similar frequencies to that which it uses. These interfering sources could even be other phones.

3.5 Data Rate Limits

With the information we have now, we are in a position to talk about the maximum data rate of a channel. The data rate depends on:

- The available bandwidth
- The levels of signal we can use
- The quality of the channel (i.e. the amount of noise)

There are two formulas that help us calculate the maximum data rate of a channel, the Nyquist Bit Rate theorem, which is used with theoretical noiseless channels, and the Shannon Capacity formula, which works with noisy channels.

The Nyquist Bit Rate theorem states that

$\text{Bit Rate} = 2 \times \text{Bandwidth} \times \log_2 \text{L}$ where L is the number of signal levels.

The Shannon Capacity formula states that

$\text{Capacity} = \text{Bandwidth} \times \log_2(1 + \text{SNR})$

Both of these can be used together. The Shannon Capacity formula will give us the maximum capacity, and how many levels are needed to achieve that capacity can be found from the Nyquist Bit Rate theorem.

Example

Consider a channel with a $1\mathrm{MHz}$ bandwidth and an SNR of 63. We need to find the appropriate bit rate and signal level.

$$\mathrm{C} = \mathrm{B}\log_2(1 + \mathrm{SNR}) = 10^6 \log_2 64 = 6\mathrm{Mbps}$$

$$6 = 2 \times 1 \times \log_2 \mathrm{L}$$

$$\therefore \mathrm{L} = 8$$

3.6 Performance

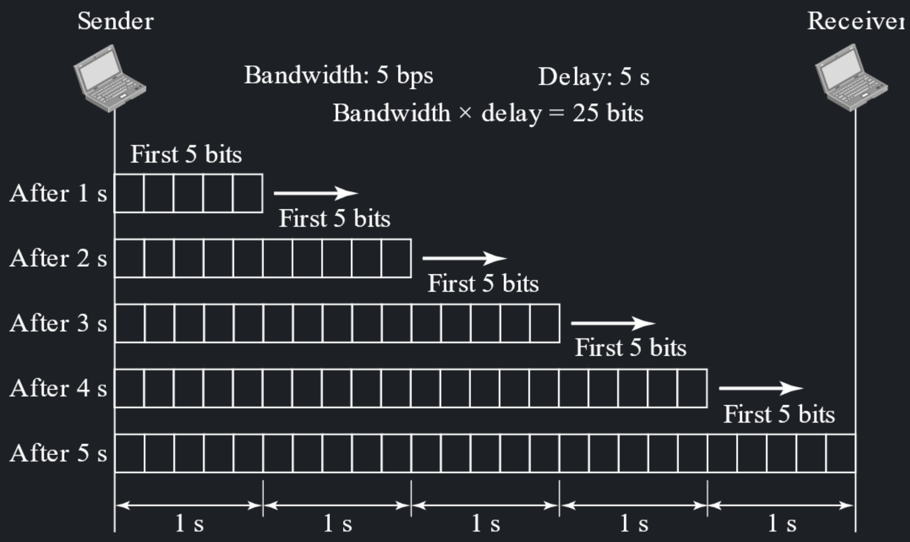
Four things are considered when measuring the performance of a network:

- **Bandwidth** – There are two types of bandwidths. The first refers to the range of frequencies that a channel can pass, measured in hertz. The second is the theoretical speed of bit transmission in a channel or link, measured in bits per second.

- **Throughput** – This is the highest practically achievable bit rate. It is the measurement of how fast data can pass through a point.

- **Latency (Delay)** – This is a measure of how long it takes for an entire transmission unit to arrive at the destination from the moment the first bit is sent out from the source. This is called nodal delay or end-to-end nodal delay. Latency actually consists of four parts:
    - **Propagation Delay** – This is the time needed for a bit to travel from the source to the destination.
    - **Processing Delay** – This is related to intermediate devices such as routers. It is nominal.
    - **Transmission Delay** – This is the time taken for all the data to be pushed from the buffer to the transmission medium at the source and vice versa at the receiver end.
    - **Queuing Delay** – Any wait times fall into this category.

- **Bandwidth-Delay Product** – This defines the maximum number of bits that can fill the link. A larger bandwidth means more bits can be sent per second, and a larger delay means more bits are in the path between the source and destination.


Sender
Receiver
Bandwidth: 5 bps
Delay: 5 s
Bandwidth × delay = 25 bits
First 5 bits
After 1 s
First 5 bits
After 2 s
First 5 bits
After 3 s
First 5 bits
After 4 s
First 5 bits
After 5 s
1 s
1 s
1 s
1 s
1 s