Inter-symbol Interference (ISI)

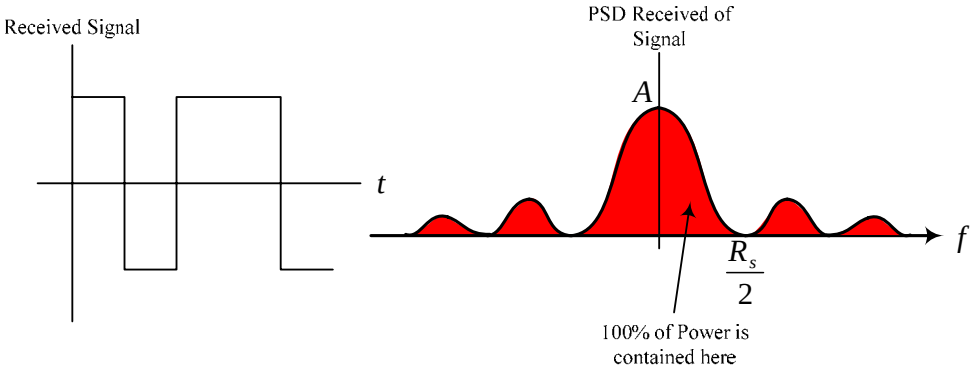
Unlike analog signals, which are usually smooth in nature, digital signals are composed of pulses with often vertical transitions. The fact that digital signals sometimes have vertical transitions increases their bandwidth significantly since it requires infinite bandwidth to represent a signal with vertical transitions. Compare for example the bandwidth of two baseband signals given by a sine wave with frequency f_0 and a square wave with frequency f_0. The sine wave has a single frequency component at f_0 Hz. However, the square wave has infinite frequency components at f_0 and integer multiples of it. If we consider the bandwidth of a signal to be the minimum frequency that encloses all frequency components of the signal (the signal has no frequency components at all above that frequency), then the sine wave will have a bandwidth of f_0 Hz because it has no frequency components above that frequency, while the square wave has an infinite bandwidth because it theoretically has frequency components that extend to infinity.

The fact that any communication system has limited bandwidth to transmit digital data indicates that certainly a transmitted square pulse will be received differently at the receiver as the channel will filter some components of it. The difference depends on how narrow the bandwidth of the channel compared to the symbol rate in the signal. The effect of filtering part of the transmitted signal by the channel on the quality of the received signal may be significant that a phenomenon called "Inter-symbol Interference (ISI)" occurs. ISI causes the transmitted pulses to get mixed together, meaning that a pulse that is transmitted between time instants will smear into adjacent pulses affecting the process of data detection and possibly causing errors not as a result of noise but as a result of symbols mixing together.

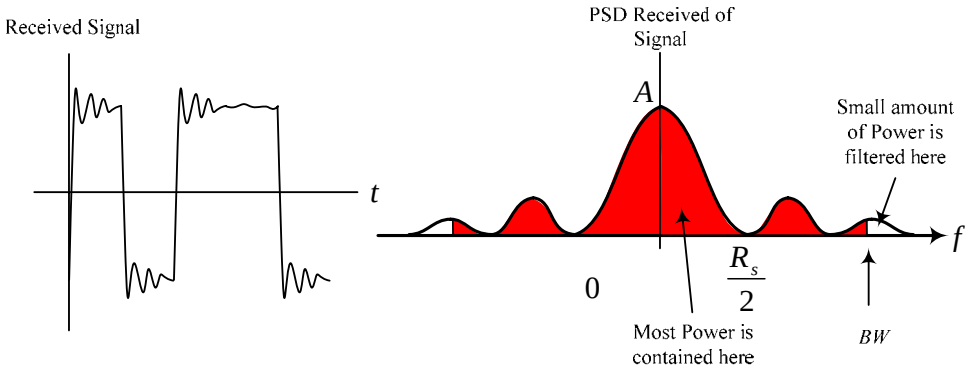
Effect of Channel Bandwidth Limitation on ISI

Consider a baseband digital signal with symbol rate R_s symbols/second that is composed of a sum of square pulses and that is transmitted through different baseband channels with different bandwidths:

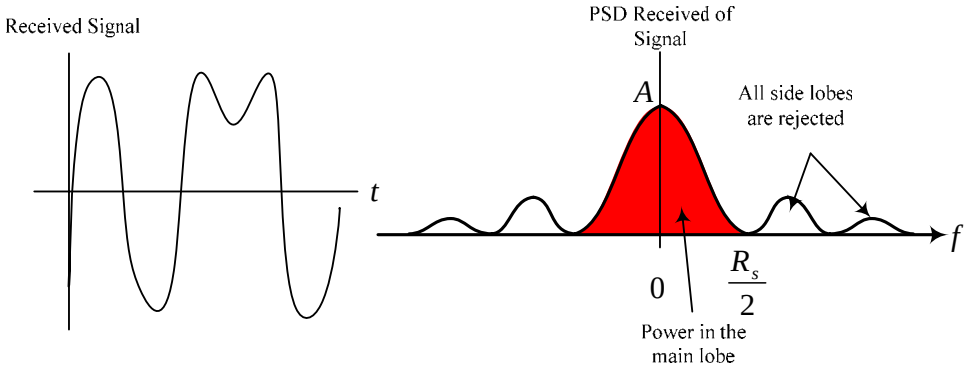
1. **Channel with Infinite Bandwidth:** Such a channel passes all signal components. In this case, the received signal will be exactly the same as the transmitted square wave since the complete signal is passed. So, the transmitted data will not experience any ISI at all.



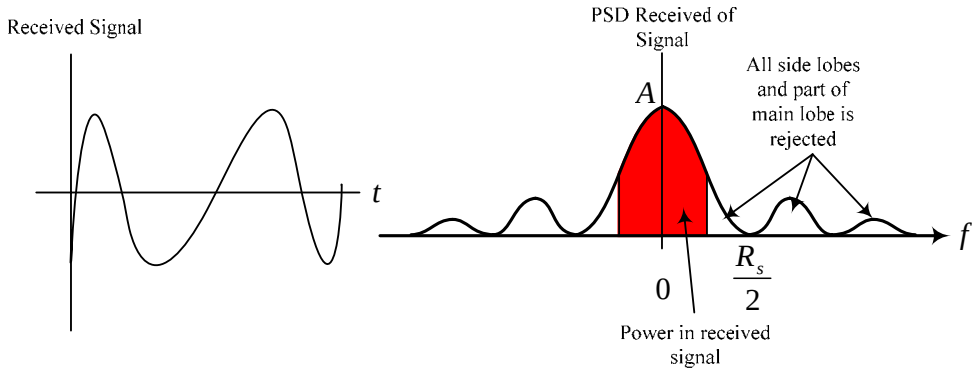
2. **Wideband Channel with channel larger than $R_s/2$:** the bandwidth of the channel in this case is wide but not infinite, so a relatively large amount of the signal power will pass and a small amount at high frequencies will be rejected. The data in this case experiences some ISI but data can easily be recovered since the ISI is limited.



3. **Channel Bandwidth is Equal to One Half Symbol Rate $R_s/2$:** The first null (zero) in the power spectrum density of transmitted data occurs at one half the sample rate $R_s/2$. The received signal in this situation experiences significant amount of ISI. However, the data is still recoverable using some signal processing algorithms. This represents the minimum channel bandwidth that would allow us to recover the data completely. Any channel bandwidth below this would cause a problem.



4. **<u>Channel Bandwidth is Lower than One Half Symbol Rate $R_s/2$:</u>** in this case, the ISI is huge and loss of data will occur. It is not possible to recover back the data completely no matter what signal processing algorithms are used.



We see from the previous 4 cases that when transmitting square pulses and the bandwidth of the channel is not infinite, then ISI will occur. However, as long as the bandwidth of the channel is greater than one half the symbol rate, data can be recovered but possibly using some signal processing algorithms to remove the effect of ISI. If the channel bandwidth is less than that, then loss of data will certainly occur.

Pulse Shaping to Control ISI

The use of rectangular-shape pulses to transmit digital information makes sense because they have flat tops which fit the shapes of digital signals perfectly. In addition, a rectangular pulse that extends over a bit (or symbol) period avoids interference between consecutive pulses as long as the exact shape of the pulses is preseved. However, the power spectral density of rectangular-pulse shapes is very wide (remember that the spectrum of a rectangular pulse is a “sinc” function). The wide spectrum of rectangular pulses means that such pulses must be transmitted over very wideband channels even for relatively low bit (or symbol) rates or else part of the transmitted signal will be filtered out by the channel and the received signal will be a distorted version of the transmitted signal. Filtering out part of the transmitted signal results in the rectangular pulses getting mixed up with preceding and succeeding pulses in what we called above Inter-Symbol Interference (ISI).

To combat ISI, the pulses that we use to transmit data must have limited bandwidth so that when transmitted over limited bandwidth channels, the complete spectrum of these signals is retained and no part of it is filtered out. This will guarantee that the signal does not change as it is transmitted through the channel. However, limiting the bandwidth of the pulses we use to transmit data causes their duration in time to be infinite (remember that time limited signals are frequency unlimited and frequency limited signals are time unlimited). A pulse with an infinite time duration (or at least very

long time duration) means that each pulse extends over a very large number of bit periods. This is not necessarily bad if the pulse is designed properly. What we mean by designed properly is that each pulse needs to be equal to a constant (1 V) at the time instant of the start of the bit that this pulse represents and at which this bit will be sampled and be zero (0 V) at all time instants of future and past bits so not to interfere with these bits at the moments that they are sampled for detection. A class of pulses called “Nyquest Pulses” satisfies all these requirements. A famous class of Nyquest pulses is called “Raised Cosine” pulses

Raised Cosine Pulses

The class of Raised Cosine pulses include the famous “sinc” function. Although the “sinc” The “sinc” function has the narrowest bandwidth of all Nyquest pulses, it decays at a very slow rate that is proportional to 1/t. This means that the generation of the “sinc” pulse corresponding to a specific symbol must start many symbol periods before the time of the symbol represented by this pulse and must continue for many symbol periods after the time of the symbol represented by this pulse. This exerts a relatively large computational requirements on the system in additional to a delay before and after the transmission of data. Other Raised Cosine pulses provide a compromise between the bandwidth (they require more bandwidth than the “sinc” pulse) with the length of tails of the pulse (they have much shorter tails than the “sinc” pulse that extend only few symbol periods before and after the time of their symbol).

The general format for a raised cosine pulse is

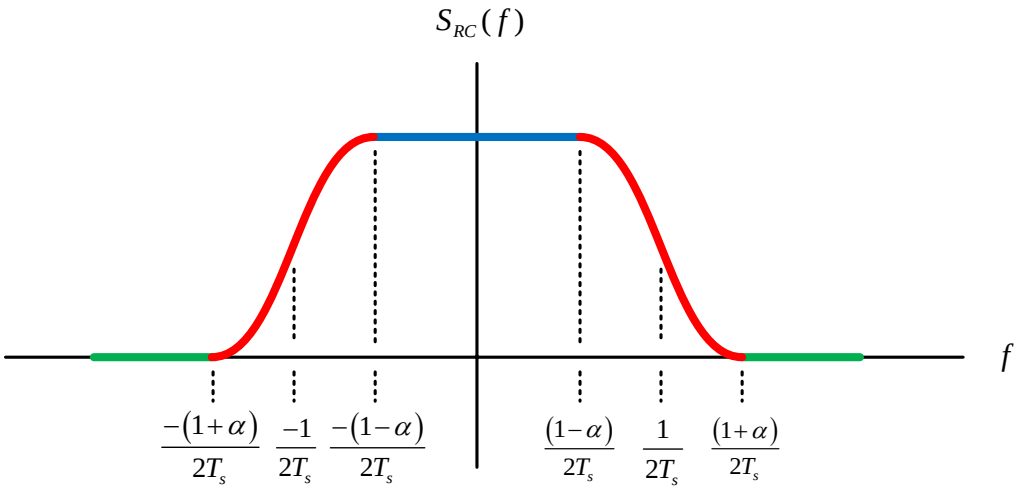
$$s_{RC}(t) = \frac{\sin\left(\frac{\pi t}{T_s}\right)}{\pi t} \cdot \frac{\cos\left(\frac{\pi \alpha t}{T_s}\right)}{1 - \left(\frac{4\alpha t}{2T_s}\right)^2}$$

where α is a parameter that provides the tradeoff between the bandwidth and tail length of the raised cosine function, and T_s is the symbol period. The first component in the raised cosine pulse shown above is a “sinc” pulse. The tails of the “sinc” pulse are attenuated further by the second component at the rate of t^2. So the raised cosine tails drop at the rate of t^3 which means that for a properly designed raised cosine, the tails die out after few (3 to 5) bit or symbol periods only. The raised cosine pulse becomes the “sinc” when the parameter $\alpha = 0$.

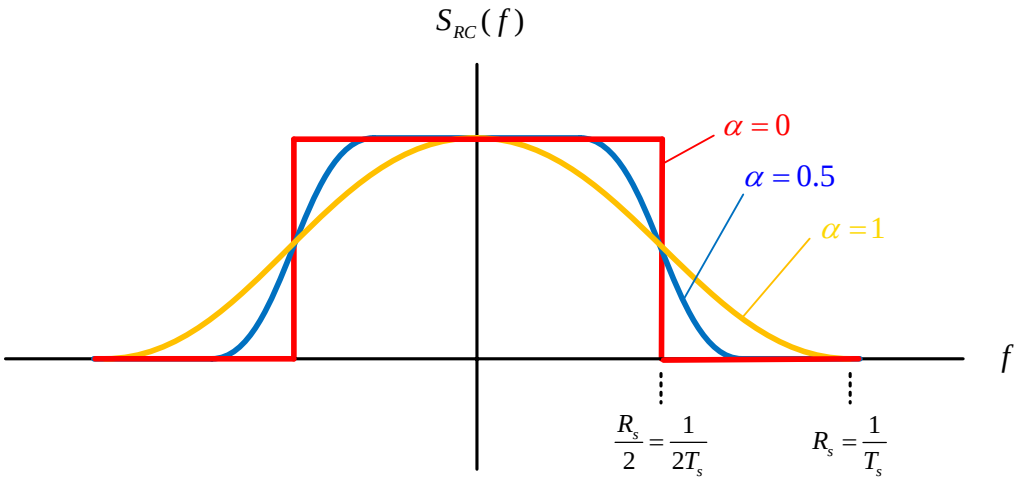
The spectrum of raised cosine pulses is

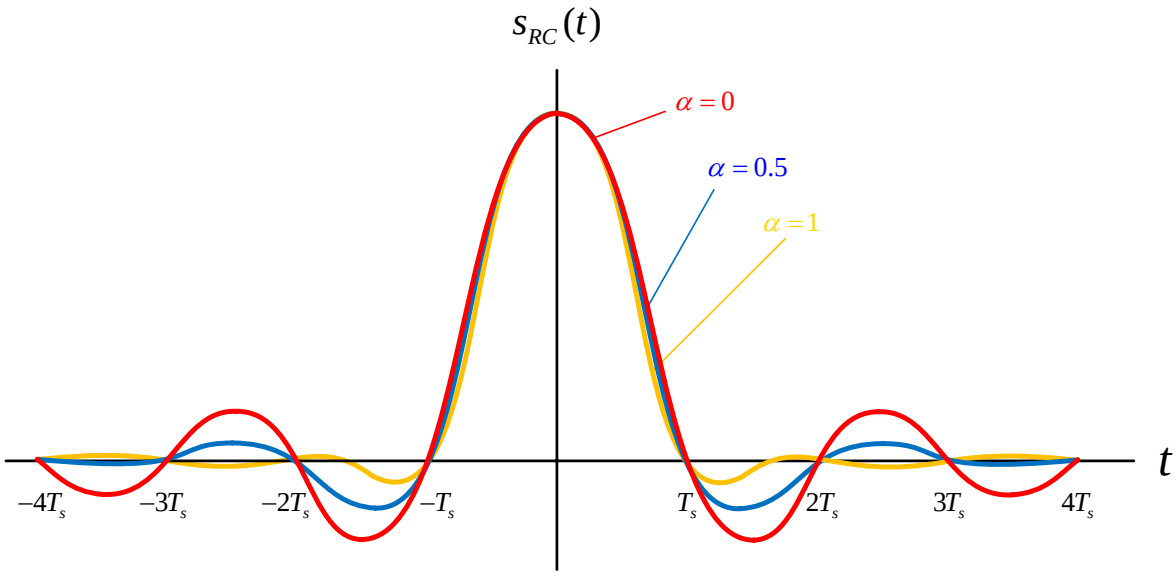
$$S_{RC}(f)=\begin{cases}1 & 0\le|f|\le\dfrac{(1-\alpha)}{2T_s}\\ \dfrac{1}{2}\left[1+\cos\left(\dfrac{\pi\left(2T_s|f|-1+\alpha\right)}{2\alpha}\right)\right] & \dfrac{(1-\alpha)}{2T_s}\le|f|\le\dfrac{(1+\alpha)}{2T_s}\\ 0 & \dfrac{(1+\alpha)}{2T_s}<|f|\end{cases}$$

The spectrum is divided into three regions that are shown in the figure below



The spectrums and time-domain pulse shapes of several raised cosine pulses are shown below for different values of α.





For baseband transmission , the symbol rate of the transmitted data that can be transmitted using a Raised Cosine pulse is related to α and the bandwidth of the signal B by the relation

$$R_s = \frac{1}{T_s} = \frac{2B}{1+\alpha} \qquad \text{(baseband transmission)}$$

and for passband transmission, the rate is half of the above value, or

$$R_s = \frac{1}{T_s} = \frac{B}{1+\alpha} \qquad \text{(passband transmission)}$$

Important Notes:

1. The required bandwidth for transmitting a digital data signal is a function of the symbol rate R_s not the bit rate R_b.
2. For a signal with

$$M = 2^{N \text{ Bits/Symbol}}$$

$$N \text{ Bits/Symbol} = \log_2 M$$

$$T_b = \frac{1}{N} T_s$$

$$R_b = N \cdot R_s$$

Relation between Probability of Error and C/N Ratio

Unlike analog signals in which quality of the signal is measured in terms of the received signal power relative to the noise power, the quality of digital signals is measured in terms of number of errors that occur in the received data as a result of added noise. The probability of error (also called bit error rate) is related to the C/N ratio of the received signal. To find this relation, let us consider a baseband binary transmission (for the case of passband transmission the same concept stands) where we are transmitting one of two pulses that have amplitudes +1 V and –1 V with equal probability. That is

$T_1 = +1 \quad \text{V}$ Prob(Transmitting T_1) = 0.5

$T_2 = -1 \quad \text{V}$ Prob(Transmitting T_2) = 0.5

Clearly, in the absence of any thermal noise or other sources of noise, the received pulses corresponding to the above transmissions are

$R_1 = +1 \quad \text{V}$ (No Noise)

$R_2 = -1 \quad \text{V}$ (No Noise)

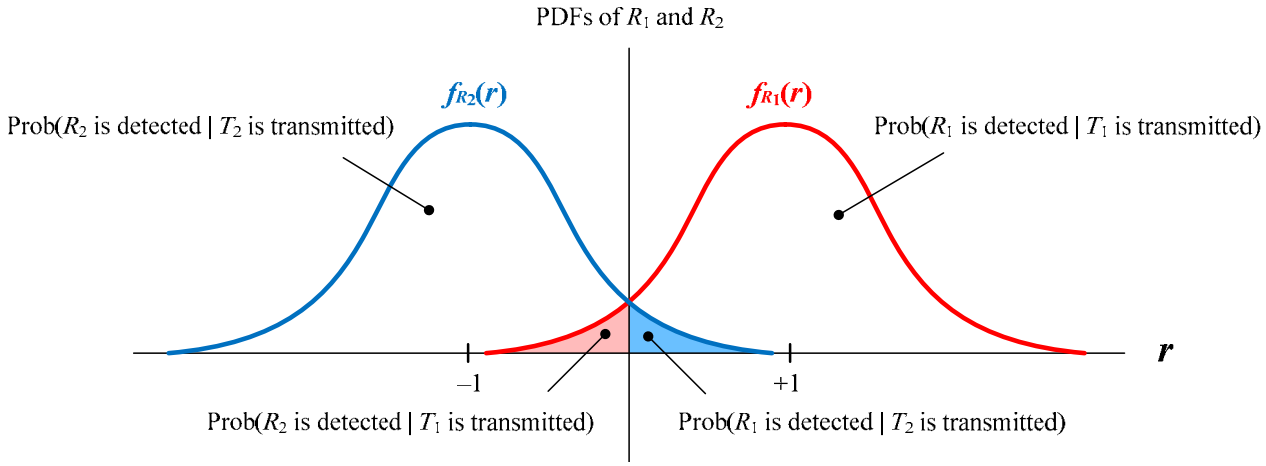
So, clearly the transmitted data can be recovered with zero probability of error.

In the presence of thermal noise, the received pulses become accompanied by normally distributed noise (noise that has probability density function that follows the Gaussian distribution). So, the received signals become:

$R_1 = +1 + N \quad \text{V}$ (with Noise)

$R_2 = -1 + N \quad \text{V}$ (with Noise)

where N is a Gaussian random variable. The probability density functions of the two random variables R_1 and R_2 will be similar to the probability density function of N (a zero-mean Gaussian random variable) except that the means of two will be +1 and –1, respectively. This is shown below:



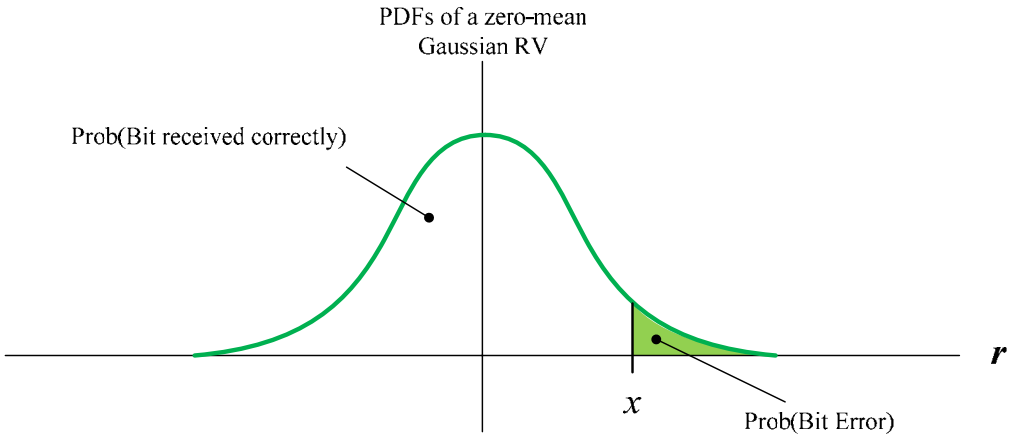
The bit error probability P_b can be computed using conditional probability as

$$P_b = P\left(\text{Detecting } R_1 | T_2 \text{ was transmitted}\right) \cdot P\left(T_2 \text{ was transmitted}\right)$$
$$+P\left(\text{Detecting } R_2 | T_1 \text{ was transmitted}\right) \cdot P\left(T_1 \text{ was transmitted}\right)$$

Since Prob(Transmitting T_1) = Prob(Transmitting T_2) = 0.5 and since the above two areas are equal to each other, then the bit error probability P_b becomes

$$P_b = P\left(\text{Detecting } R_1 | T_2 \text{ was transmitted}\right) = P\left(\text{Detecting } R_2 | T_1 \text{ was transmitted}\right)$$

which is equivalent to the area in probability density function of the zero mean Gaussian random variable shown below:



Therefore, the probability of bit error can be written as

$$P_b = \frac{1}{\sqrt{\pi}} \int_x^{\infty} e^{-u^2} du$$

This integration does not have a closed form. Instead, it is often expressed in terms of a function called the “error function complement (erfc). The above expression becomes

$$P_b = \frac{1}{\sqrt{\pi}} \int_x^{\infty} e^{-u^2} du$$
$$= \frac{1}{2} \operatorname{erfc}(x)$$

What is x in the expression above? The value of x is what determines the probability of bit error. This quantity is related to the signal power and noise power. In fact, this quantity is expressed as

$$x = \sqrt{\frac{E_b}{N_0}}$$

where E_b is the bit energy (energy contained in a bit) while N_0 is the thermal noise power per unit Hz of bandwidth.

Relation between $\frac{E_b}{N_0}$ and $\frac{C}{N}$

You may ask, why is the probability of bit error expressed in terms of E_b/N_0 ratio and not in terms of C/N ?

The answer is simple. Assuming equal amounts of noise power are added to the digital signals transmitted by two different systems, the probability of error in the received data of the two systems may be different. The reason is that although both systems are transmitting equal amounts of power, what counts is how much energy is allocated per bit in each system. To see this, consider that one of the systems transmits much more data than the other, yet the transmitted power by each system is the same. Clearly, the system that transmits more data allocates smaller amounts of energy per bit, and it is expected that the probability of bit error for that system would be worse (higher proability of bit error). However, this does not mean that the system has a worse performance than the other. So, to have fare comparison, it is important to compare two systems with equal bit energy rather than equal transmitted power. Now, consider two systems that transmit equal amounts of data. However, one of them uses much more bandwidth than the other. Clearly, the system that uses more bandwidth may have a lower probability of bit error because of the fact that it is using wider bandwidth. Also, to have fair comparison, the thermal noise should be evaluated in terms of noise per Hz. For both of these, it is seen that what determines the probability of bit error is E_b/N_0 ratio and not C/N ratio.

The ratios of E_b/N_0 and C/N can be related to each other by observing that Energy = Power * Time and that Noise per Hz = Total Noise / Bandwidth, or

C = Carrier Power (W)

E_b = Bit Energy (J) $= C \cdot T_b = \frac{C}{R_b}$

N = Noise Power (W)

N_0 = Noise Power per Hz (W/Hz = J) $= \frac{N}{BW}$

Therefore,

$$\frac{E_b}{N_0} = \frac{C / R_b}{N / BW} = \frac{\left(\frac{C}{N}\right)}{\left(\frac{R_b}{BW}\right)}$$

The quantity $\frac{R_b}{BW}$ represents the total bit rate of the system divided by the amount of bandwidth the system uses to transmit this data. This is called the "Throughput" of the communication system, which is the number of bits/s that the system transmits in each Hz of bandwidth that is allocated for it, or

$$\text{Throughput} = \frac{R_b}{BW}$$

Comparison of Different Systems

a) BPSK

$T_s = T_b$

$R_s = R_b$

$\text{Throughput}_{\text{BPSK, Zero-ISI}} = \frac{1}{2}$

a) 4-PSK (QPSK)

$$T_s = 2T_b$$

$$R_s = \frac{1}{2}R_b$$

$$\text{Throughput}_{\text{QPSK, Zero-ISI}} = 1$$

$$P_{\text{s,QPSK}} = 2P_{\text{s,BPSK}} = 2P_{\text{b,BPSK}}$$

$$P_{\text{b,QPSK}} = \frac{1}{2}P_{\text{b,QPSK}}$$

$$\boxed{P_{\text{b,QPSK}} = P_{\text{b,BPSK}}}$$

The benefit of using QPSK over BPSK is that a higher bit rate can be achieved without any deterioration in bit error probability performance.

b) 8-PSK

$$T_s = 3T_b$$

$$R_s = \frac{1}{3}R_b$$

$$\text{Throughput}_{\text{QPSK, Zero-ISI}} = 1.5$$

$$P_{\text{s, 8-PSK}} = 4P_{\text{s, BPSK}}$$

$$\boxed{P_{\text{b, 8-PSK}} = \frac{3}{2}P_{\text{b, BPSK}}}$$