1

Radio Frequency (RF) Filter Networks for Wireless Communications—The System Perspective

This chapter is dedicated to an overview of communication systems, especially the relationship between the communication channel and other elements of the system. The intent here is to provide the reader with sufficient background to be able to appreciate the critical role and requirements of radio frequency (RF) filters in communication systems. A number of standard texts [1–8] have been referred to in developing a significant portion of this chapter.

This chapter is divided into three parts: Part I presents a simple model of a communication system: the radio spectrum and its utilization, the concept of information, and system link budgets. Part II describes the noise and interference environment in a communication channel, the nonideal amplitude and phase characteristics of the channel, the choice of modulation–demodulation schemes, and how these parameters affect the efficient use of the allocated bandwidths. Part III discusses the impact of system design on the requirements and the specifications of microwave filter networks in satellite and cellular communication systems.

PART I Introduction to a Communication System, Radio Spectrum, and Information

1.1 Model of a Communication System

Communication refers to the process of conveying information-bearing signals from one point to another that are physically separate. In ancient times, people communicated over long distances by various means, such as smoke signals, drum beating, homing pigeons, and horseback riders. All such means were slow in the transmission of information over any appreciable distance. It was the invention of electricity that changed the means of communication. Communication became almost instantaneous by transmission of electrons through wires or electromagnetic (EM) waves through empty space or fibers, which is limited only by the speed of light—a fundamental constraint of our universe.

At the highest (simplest) level, communication involves an information source, a transmitter, a communication medium (or channel), a receiver, and an information destination (sink), as depicted in Figure 1.1. Until 1980s, most information was communicated in an

Microwave Filters for Communication Systems: Fundamentals, Design, and Applications, Second Edition.
Richard J. Cameron, Chandra M. Kudsia, and Raafat R. Mansour.

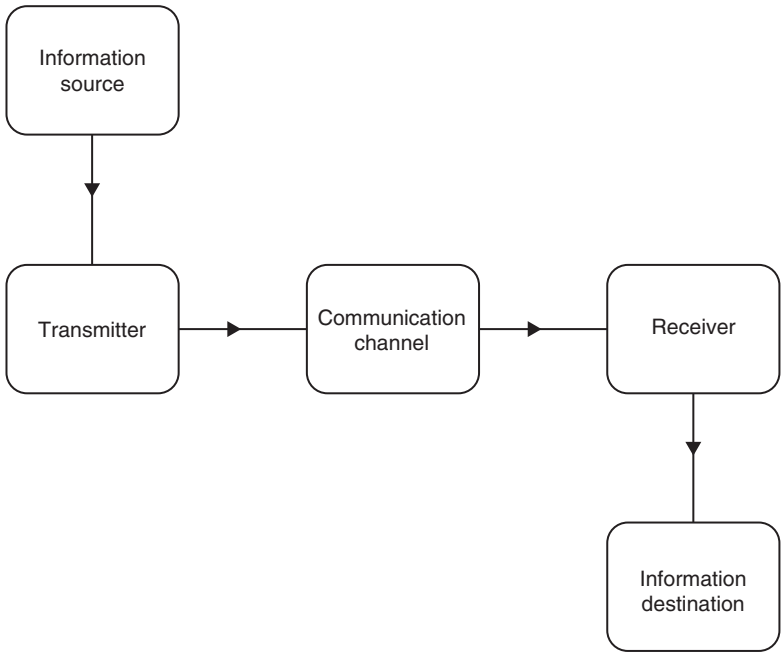


Figure 1.1 Simple model of a communication system.

analog format called *analog communication*. Today, most information is communicated in a digital format called *digital communication*. Even information in analog format is routinely converted into digital format for transmission and then converted back to analog format at the destination.

All communication systems are required to be linear. For such systems, the law of superposition holds. It allows the use of common media for the transmission and reception of an arbitrary number of independent signals, subject only to the constraints of the available bandwidths and adequate power levels. However, all the components of a communication system need not be linear, as long as the overall system is linear over the specified range of bandwidth within an acceptable degree of nonlinearity. In fact, all active devices are inherently nonlinear, which is essential for the purposes of frequency generation, modulation, demodulation, and amplification of signals. However, such intentional nonlinearities can be controlled for a specific application. For broadband wireless communication systems such as line-of-sight (LOS) and satellite systems used for long-distance networks, the frequency spectrum is divided into a number of RF channels, often referred to as *transponders*. Within each RF channel, there can be more than one RF carrier, depending on the system requirements. The channelization of the frequency spectrum provides the flexibility for the communication traffic flow in a multiuser environment. High-power amplifiers (HPAs) can operate with relatively high efficiency, since they are required to amplify a single carrier, or a limited number of signals, on a channel-by-channel basis, incurring a minimum distortion.

Irrespective of the communication format, it should be recognized that the passage of the transmitted signal through the communication channel is a strictly analog operation. The communication channel is a nonideal, lossy medium, and the reception of signals at the receiver involves recovery of the transmitted signal in the presence of impairments, in particular, thermal noise at the receiver, signal distortions (within the channel and from

nonideal transmitter and receiver), and interference from other signals or echoes (multipath) seen by the receiver.

1.1.1 Building Blocks of a Communication System

In this section, the building blocks shown in Figure 1.2 are described for both analog and digital communication systems.

1.1.1.1 Information Source

The information source consists of a large number of individual signals that are combined in a suitable format for transmission over the communication medium. Such a signal is referred to as the *baseband signal*. The transducers shown in Figure 1.2 are required to convert the energy of the individual information sources, either acoustic (voice) or electrical, into an appropriate electrical signal suitable for transmission. For an analog system, all the individual signals, as well as the combined baseband signal, are in an analog format as illustrated in Figure 1.2a.

For digital systems, the baseband signal is a digital datastream, whereas the individual signals constituting the baseband can be digital or analog. Consequently, the individual analog signals need to be converted into their equivalent digital format via analog-to-digital (A/D) converters. Another feature of the information source in a digital system is the use of data compression to conserve bandwidth. A compressor takes the digital data and exploits its redundancy and other features to reduce the amount of data that need to be transmitted but still permits the information to be recovered. The information source for a digital communication system is illustrated in Figure 1.2b.

1.1.1.2 Transmitter

The block diagram of a transmitter is presented in Figure 1.3, and the functionality of each element is described as follows:

Encoder. In digital systems, an encoder introduces error correction data into the baseband information stream that permits recovery of the digital information even after significant impairments are created in the communication channel.

Modulator. It transcribes the baseband signal onto a higher intermediate carrier frequency (IF) as an intermediate step in the transmission and reception of the information bearing signals. Use of IF simplifies the filtering and signal processing circuitry in the modulator. The modulator can shift the signal frequencies, change the bandwidth occupancy, or materially alter the form of the signal, making it more suitable and efficient for transmission over the communication medium.

Upconverter. Also referred to as a mixer, shifts the modulated IF carrier frequency to the microwave range of radio frequencies (RF) within the allocated frequency band for RF transmission.

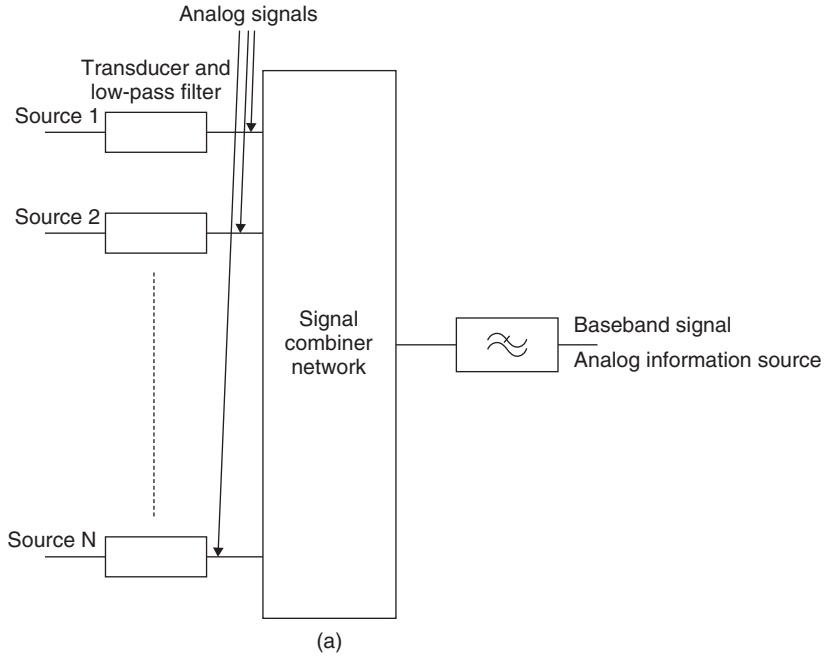
RF amplifier. Is used to amplify the RF signal. RF power has a direct bearing on the communication capacity of a RF channel.

RF multiplexer. It is used to combine the power of a number of RF channels into a composite broadband signal for transmission via a common antenna.

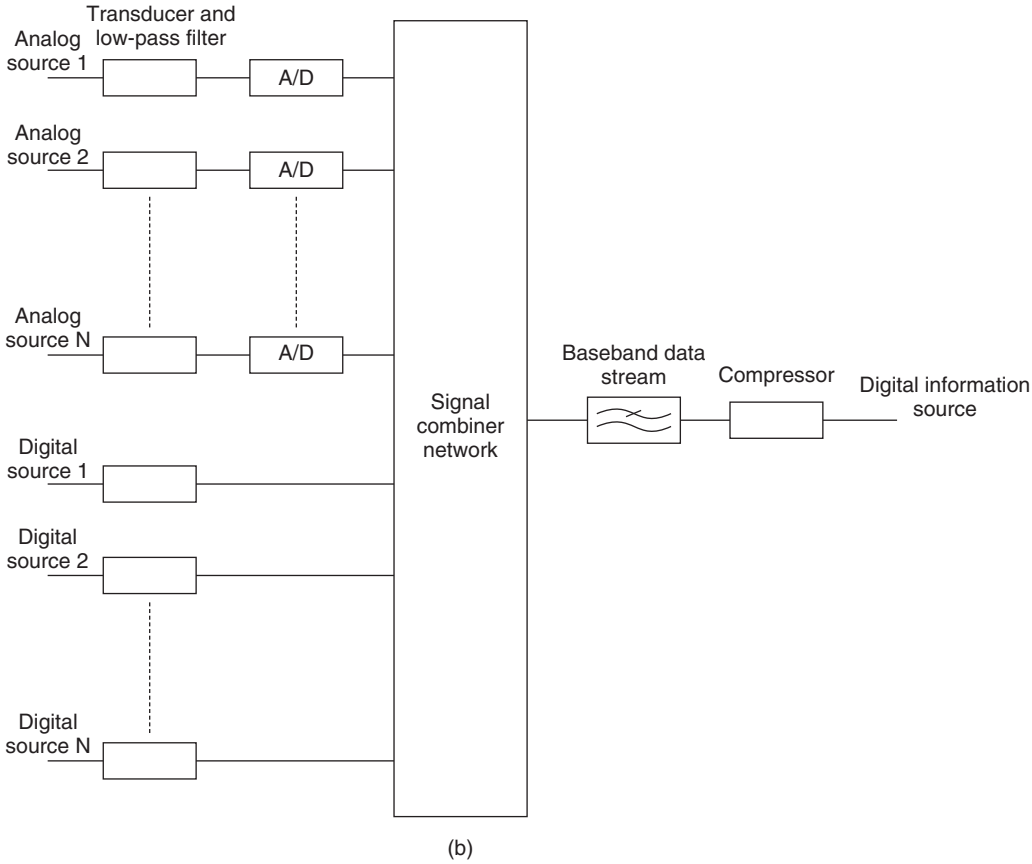
Transmit antenna. Launches the RF power into space and focuses it toward the receiving station.

1.1.1.3 Communication Channel

For wireless systems, the communication channel is free space. As a consequence, the properties of space, including the atmosphere, play a critical role in system design.



(a)



(b)

Figure 1.2 Information source: (a) analog system; (b) digital system.

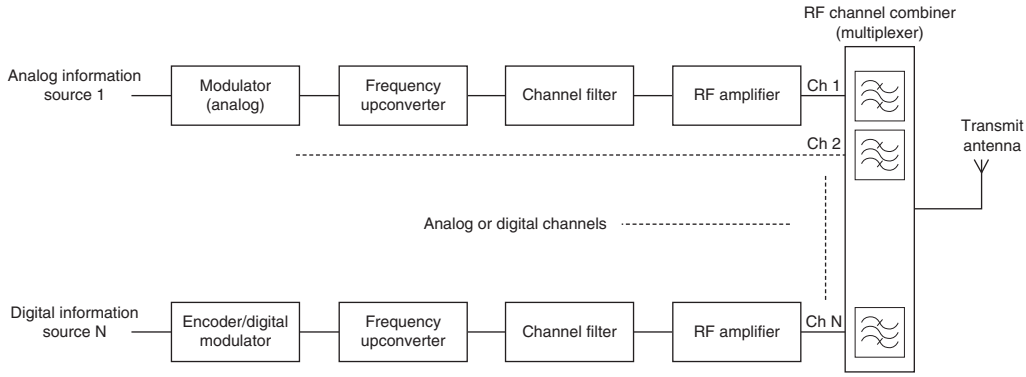


Figure 1.3 Transmitter block diagram.

1.1.1.4 Receiver

The block diagram of a receiver is shown in Figure 1.4, and the functionality of each element is described as follows:

Receive Antenna. It intercepts the RF power and focuses it to a transmission line connected to the low-noise amplifier (LNA).

LNA. It amplifies the very weak received signal with a minimum addition of noise to it.

Downconverter. It serves the function of frequency conversion, which is similar to that required in the transmitter chain. The downconverter converts the uplink frequency band into the downlink frequency band.

Demodulator. It extracts the baseband signal from the RF carrier by a process opposite to that of a modulator.

Decoder. It exploits the error correction data previously inserted into the information stream and uses it to correct the errors made during digital demodulator's recovery of the data.

1.1.1.5 Information Destination

The functionality of the information destination block is opposite to that of the information source. In the case of digital systems, an expander is used to reverse the operation of the compressor as shown in Figure 1.5.

Since a communication system is costly to install, its commercial viability is critically dependent on the number of users who must share the transmission medium. As a result, the

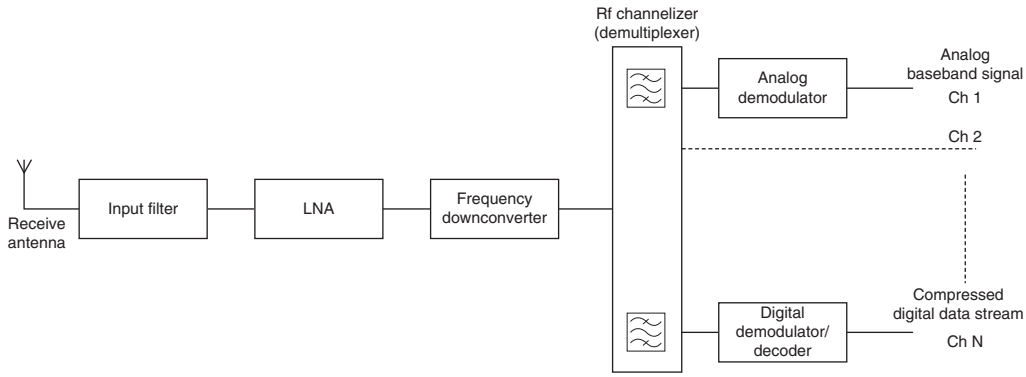


Figure 1.4 Receiver block diagram.

Figure 1.5 Information sink for a digital system.

information source usually consists of a large number of signals, occupying a finite range of frequencies. The width of the range of frequencies is called the *bandwidth* of the system.

Some key questions then arise: Is there a limitation on the available bandwidth for a communication system? What are the limitations of transmitting information over the chosen transmission media in the available bandwidth? What are the cost-sensitive parameters in a communications system?

1.2 Radio Spectrum and its Utilization

To understand the limitations on the available bandwidths for a communication system, it is necessary to understand the radio spectrum and its utilization [4].

EM waves cover an extremely broad spectrum of frequencies, from a few cycles per second to γ rays with frequencies of up to 10^{23} cycles per second. The radio spectrum is that portion of the EM spectrum which can be electronically and effectively radiated from one point in space and received at another. This includes frequencies anywhere from 9 kHz to 400 GHz. Although most of the commercial use takes place between 100 kHz and 44 GHz, some experimental systems reach as high as 100 GHz. The signals employing these same frequencies can also be transmitted over long distances by wire, coaxial cables, and glass fibers. However, since such signals are not intended to be radiated, they are not considered as part of the radio spectrum. A communication system is allowed to access only a portion of the radio spectrum, due to not only technology limitations but also regulatory reasons. The radio spectrum is subdivided into smaller frequency "bands" by national and international agencies, where each band is restricted to a limited set of types of operation. In addition, each band is a controlled commodity that often has a license fee associated with its use. Obviously, this represents a huge incentive to make the most efficient use of the allocated frequency spectrum.

1.2.1 Radio Propagation at Microwave Frequencies

There are many sources of energy loss in the free-space communication medium. The most serious ones include rainfall and presence of oxygen in the atmosphere. The atmospheric losses as a function of frequency are presented in Figure 1.6. The radio energy is absorbed and scattered by the raindrops, and this effect becomes more intense as the wavelength approaches the size of the raindrops. Consequently, rainfall and water vapor produce intense attenuation effects at higher microwave frequencies. The first absorption band, due to water vapor, peaks at approximately 22 GHz, and the first absorption band, due to the presence of oxygen in the atmosphere, peaks at about 60 GHz.

For fixed line of sight (LOS) terrestrial microwave radio links, multipath fading is another major cause of signal loss. Fading results from the variations in the refractive index of air for the first few tenths of a kilometer above the earth's surface. Such gradients in refraction bend the rays, which, on reflection from the ground or other layers, combine with the direct rays, causing coherent interference.

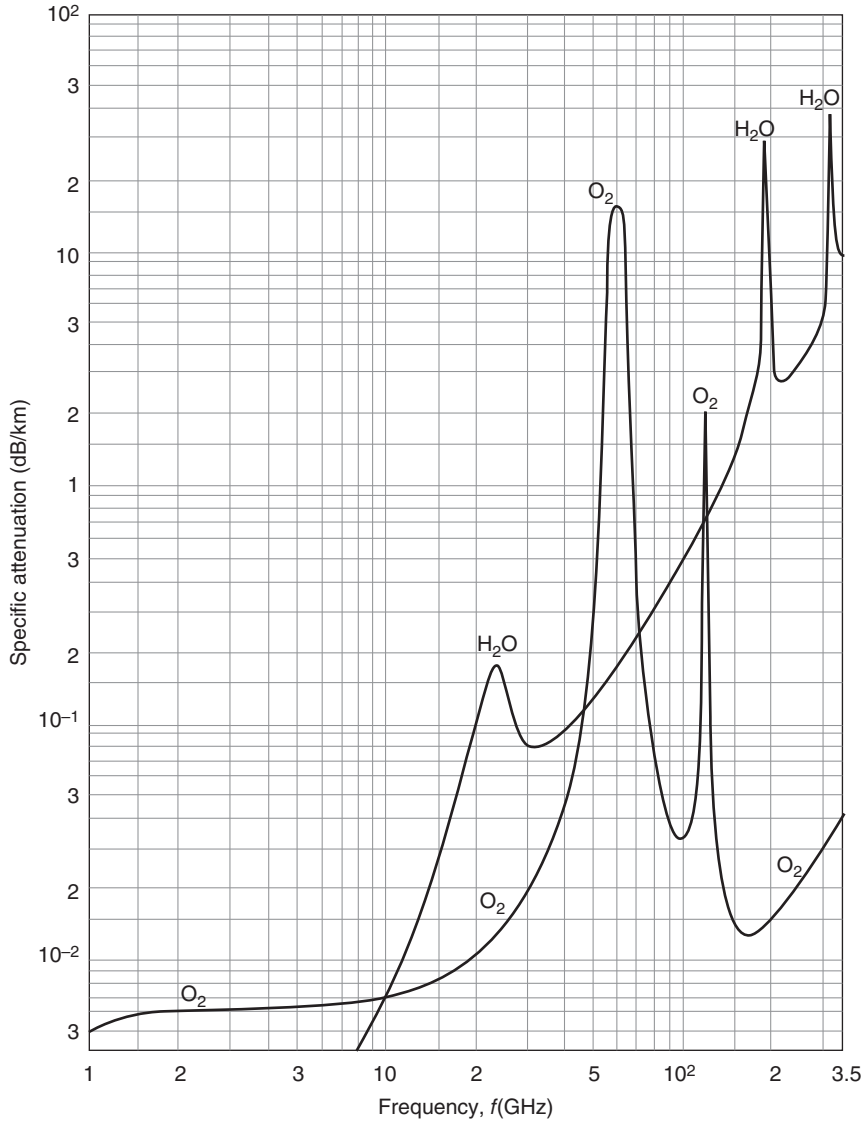


Figure 1.6 Atmospheric losses as a function of frequency. (*Source*: Freeman 2007 [7]. Reproduced with the permission of John Wiley and Sons. CCIR Report., 1990. Reproduced with permission International Radio Consultative Committee.)

Mobile communication adds a new dimension to the propagation problem. Besides the requirements of omnidirectional coverage and the mobility of end users, the communication system must deal with non-LOS and multipath problems caused by the signal reflections from tall buildings, trees, valleys, and other large objects in an urban environment. In addition, the coverage area for mobile services includes the area inside buildings. Also, mobility incurs the *Doppler shift*, a change in the frequency of the received signal, further complicating the issue.

From the foregoing analysis, it is evident that the attenuation due to rain and other atmospheric effects, coupled with multipath fading in an urban environment, severely constrains the available frequency spectrum suitable for commercial communications.

1.2.2 Radio Spectrum as a Natural Resource

The radio spectrum is a natural resource unlike any other communication system. It is a completely renewable resource that can never be permanently depleted. It is also universally available. The limitation comes with its usage. The radio spectrum has a finite capacity that, if exceeded, results in interference, incapacitating the system. For this reason, national governments grant users the privilege of using radio spectra in exchange for agreements to abide by usage rules. Because RF signals often cross national borders where they can interfere with radio frequencies allocated in another country, nations must cooperate to find ways to coordinate their individual allocations. Since communication is vital to all nations, there are national and international agencies that regulate the allocation and usage of the RF spectrum.

The International Telecommunications Union (ITU), part of the United Nations, is the international body that determines the worldwide radio spectrum allocations. The ITU accomplishes this through World Administrative Radio Conferences (WARC), where ITU member nations attempt to reach a consensus on proposals by different countries. These meetings require a consensual approach to decision-making, rendering the process very tedious, and often resulting in delays. Once a consensus is reached, the ITU publishes tables of the frequency allocations, deemed as the *radio regulations* of the ITU. Each country then makes its own detailed frequency allocation plan consistent with the radio regulation tables. In addition, there are other consultative committees such as the International Radio Consultative Committee (CCIR), a part of the ITU assigned to study and recommend the standards for interoperability and guidelines and the control of interference from various services. On the domestic front, most countries have government agencies such as the Federal Communications Commission (FCC) of the United States that regulate all the nonfederal government use of the frequency spectrum. There are other agencies that control frequency allocations for government use, including the military.

In most countries, the priority followed by the regulators for allocating radio spectrum is as follows:

- Military
- Public safety functions such as aeronautical and marine emergency communications, police, fire, and other emergency services
- National telecommunication companies for telephone
- Broadcast radio and television
- Private users such as mobile systems and other services.

Because of the complex web of procedures, priorities, government policies, and so on, once a service has been established and uses a particular portion of the spectrum, it is seldom changed. Incumbency tends to be a big advantage, and often an obstacle in the efficient use of the radio spectrum. The emergence of widespread wireless communications and services has added enormous pressure in both domestic and international regulatory bodies, as well as in original equipment manufacturers (OEMs) to ensure that the frequency allocations and spectrum usage represent the most efficient use of this natural resource.

1.3 Concept of Information

What are the limitations for transmitting information over the chosen transmission media? At the fundamental level, the answer lies in the seminal work of Shannon [9], who developed the concepts of information theory.

In 1948, Claude Shannon from the Bell Labs pointed out that "for a signal to carry information, the signal must be changing; and to convey information, the signal must resolve uncertainty."

Thus, the measure of the amount of information is a probabilistic one. Shannon defined the uncertainty of the outcome between two equally likely message probabilities as a unit of information; thereby, he used, as his measure, the binary digit or bit. Also, he showed that the information capacity of a system is fundamentally limited by a very few parameters. Specifically, he demonstrated that the maximum information capacity C of a channel is limited by the channel bandwidth B and the signal-to-noise ratio (S/N) in the channel as described by the relationship

$$C = B \log_M \left(1 + \frac{S}{N}\right) \quad \text{information messages per second} \tag{1.1}$$

where M is the number of possible source information message states. Here, S/N defines the conditions at the demodulator input, where S is the signal power in watts and N is white Gaussian noise with a mean power of N watts distributed uniformly over the channel bandwidth at the same location. In analog communication, M is difficult to define; however, in digital communication, the binary data ($M = 2$) is considered and the capacity limit is given by

$$C = B \log_2 \left(1 + \frac{S}{N}\right) \quad \text{information bits per second} \tag{1.2}$$

For digital communication, this capacity is defined in the context of information bits, measured at the output of the compressor function before the encoding function, and not at the data source as might be expected; B and S/N are defined within the channel at the input of the receiver section.

Note that the information is not defined at the data source, since there is a significant distinction between "data" and "information":

Data are the raw output from the source, for example, a digitized voice, digitized video, or a sequence of text characters, defining a text document.

Information is the salient content in the raw data. This content is often much less than the data used to represent the raw source output.

For example, consider a voice digitizer. It must continually sample the voice analog source and produce output, even when a person pauses between the words or sentences. Similarly, in a video signal, much of the picture does not change from frame to frame. In a text document, some characters or words are repeated more often than others, but the same number of bits is used to represent each character or word at the data source.

Digital compression exploits the knowledge about the characteristics of the data source type to map the data from one format of representation into another format that reduces the number of digital bits required to provide the salient information. The details of this are not addressed in this book. However, it is important to realize that there can be substantial reductions of raw data prior to their transmission with minimal or no loss of the useful information. For example, the current generation digital TV uses MPEG2 compression that provides at least one order of magnitude reduction in the average data rate required to represent the source data compared with the digitization of the raw video. Digital compression has certainly attracted the interest of the research community over the last few decades, resulting in continual improvement.

Shannon's information theorem proves that as long as the information transmission rate R is less than C, it is possible to limit the error in transmission to an arbitrarily small value. The technique for approaching this limit in digital communication is called *coding*, another favored area of research over the last few decades. The current technology comes within a few

tenths of a decibel of the Shannon limit. The theorem states that for $R < C$, transmission can be accomplished without error in the presence of noise. This is a surprising result in the presence of Gaussian noise, since its probability density extends to infinity.

This theorem, although restricted to the Gaussian channel, is fundamentally important for two reasons: (1) the channels encountered in physical systems are generally Gaussian and (2) the results obtained for the Gaussian channel often provide a *lower bound* on the performance of a system, indicating that it has the highest probability of error. Thus, if a particular encoder–decoder is used with a Gaussian channel and an error probability P_e results, then, with a non-Gaussian channel, another encoder–decoder can be designed so that P_e is smaller. Similar channel capacity equations have been derived for a number of non-Gaussian channels.

Shannon's theorem indicates that a noiseless Gaussian channel ($S/N = \infty$) has an infinite capacity, irrespective of the bandwidth, but the channel capacity does not become infinite as the bandwidth becomes infinite. This is due to the increase in the noise power as the bandwidth is increased. As a result, for a fixed signal power and in the presence of white Gaussian noise, the channel capacity approaches the upper limit with the increasing bandwidth. By using $N = \eta B$ in equation (1.2), where η is the noise density (W/Hz), we obtain

$$C = B\log_2\left(1 + \frac{S}{\eta B}\right) = \frac{S}{\eta}\log_2\left[1 + \frac{S}{\eta B}\right]^{\eta B/S} \tag{1.3}$$

and

$$\lim_{B\to\infty} C \approx \frac{S}{\eta}\log_2 e = 1.44\frac{S}{\eta} \tag{1.4}$$

It is evident that there is also an absolute lower bound on the received signal power in a system for a given capacity, irrespective of the bandwidth utilized. For a given capacity

$$S \geq \frac{C\eta}{1.44} \tag{1.5}$$

In accordance with equation (1.2), once minimum received signal power is exceeded, the bandwidth is traded off with the S/N ratio and vice versa. For example, if $S/N = 7$ and $B = 4$ kHz, $C = 12 \times 10^3$ bps. If the signal-to-noise ratio (SNR) is increased to $S/N = 15$ and B is decreased to 3 kHz, the channel capacity remains the same. With a 3 kHz bandwidth, the noise power is three-fourth as large as with a 4 kHz bandwidth. As a result, the signal power must be increased by the factor $\frac{3}{4} \times \frac{15}{7} = 1.6$. Therefore, this 25% reduction in bandwidth requires a 60% increase in signal power.

Once the basic parameters of radio spectrum and channel bandwidths have been established, the next question is: *What are the cost-sensitive parameters in the communication system?* This leads to an evaluation of the communication channel and system link parameters.

1.4 Communication Channel and Link Budgets

Communication capacity of a channel is bounded by signal power S, bandwidth B, and noise N in the channel. The parameters and the resultant trade-offs are investigated in the subsequent sections.

1.4.1 Signal Power in a Communication Link

Wireless communication systems require transmission of signals as radio frequencies in space. RF signals travel through earth's atmosphere and are collected by the receiver. The direction of

the propagation can vary from a horizontal direction for terrestrial fixed and mobile systems to a vertical direction for satellite systems. The received signal power is primarily a function of four elements of the communication system:

- The transmitter output electrical power at the RF
- The fraction of the transmitter electrical power directed at the receiver as an RF free-space wave, defined by the transmit antenna gain
- The loss of energy in the communication medium, including the loss due to spherical spreading of the energy
- The fraction of the received free-space RF power at the receiver converted into electrical energy, defined by the receive antenna gain.

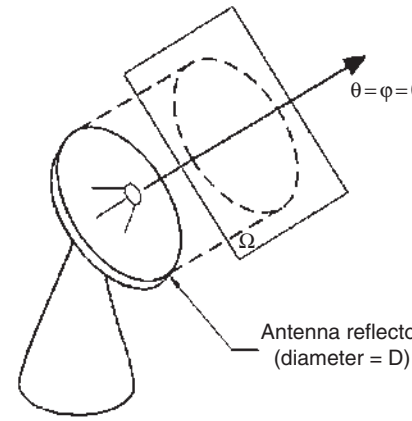
1.4.1.1 RF Power

For a given allocation of bandwidth and a certain level of noise, the communication capacity is governed by the available power for radio transmission. At first glance, it appears that the capacity can be arbitrarily increased by increasing the level of the RF power. There are two problems with this approach: (1) the generation of RF power is expensive and represents a cost-sensitive parameter and (2) a substantial amount of the RF power is wasted in the process of radio transmission, rendering the power the most costly portion of a wireless system. It behooves us to examine the basic limitations of power transmission in a wireless system.

1.4.2 Transmit and Receive Antennas

Antennas are used to launch EM energy from a transmission line into space and vice versa. Antennas are linear, reciprocal elements, and, as a consequence, the properties of an antenna are the same in transmission and in reception. The reciprocity theorem holds for all the antenna characteristics. A physical antenna (reflector type) consists of a reflecting surface and a radiating or absorbing feed network. The reflector is used to focus the energy in a given direction. The feed element converts electric currents into EM waves in a transmitting system or converts EM waves into electric currents in a receiving system. A typical parabolic microwave antenna is depicted in Figure 1.7. A transmit antenna focuses energy in the direction of a receive station or over a specific geographic area. For a receive antenna, the aperture collects the power and focuses it on the input feed element of the receive system. The general properties of the antennas described here can be applied to a broad range of frequency spectra.

Figure 1.7 Parabolic antenna.



1.4.2.1 Antenna Gain

The gain of an antenna is defined with respect to an isotropic antenna. An isotropic transmitting antenna is equivalent to a point source, radiating uniform spherical waves equally in all directions as denoted in Figure 1.8a. The power p_0 in watts, radiated from such a source has a uniform power flux density (pfd) of $(p_0/4\pi r^2)$ watts per square meter at distance r from the source. Assume that the power source is located at the input terminal of an antenna. The power radiated from this source is proportional to $p_0/4\pi$ in any direction (θ, ϕ) of the surrounding space. A directional antenna radiates power $p(\theta, \phi)$ in direction θ, ϕ as shown in Figure 1.8b. The gain of the directional antenna, with respect to an isotropic source, is then given by

$$g(\theta, \phi) = \frac{p(\theta, \phi)}{(p_0/4\pi)} \tag{1.6a}$$

The gain varies, depending on the specific values θ and ϕ. The maximum value of the gain, as determined by the maximum value of power within the (θ, ϕ) envelope, is simply defined as the antenna gain g. It is usually expressed in decibels:

$$G = 10 \log g \quad \text{dBi or simply dB} \tag{1.6b}$$

Note that this definition is independent of the physical attributes of the antenna and depends solely on the geometry of the radiation patterns of the antennas. For most applications employing uniform parabolic antennas, this maximum gain occurs along the boresight of the antenna, where $\theta = \phi = 0$.

1.4.2.2 Effective Aperture and Antenna Gain

Physical antennas are designed to radiate and capture energy in the desired directions with a minimum of loss and spillage of energy outside the region of interest. Effective aperture A_e of an antenna is defined as the equivalent physical area of the antenna that captures or radiates energy within the desired direction of θ and ϕ. It is defined by

$$A_e = \eta A \tag{1.7}$$

where $\eta(< 1)$ is the efficiency of the antenna and A is the physical aperture used to radiate or capture energy. The effective aperture represents the projected area in the direction of the beam

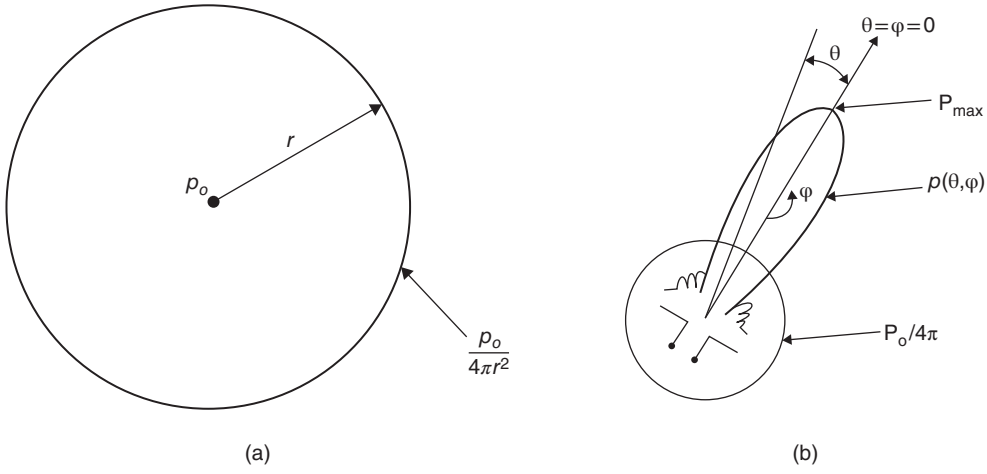


Figure 1.8 Radiated power of a transmit antenna: (a) uniform spherical waves from an isotropic source; (b) a directional antenna beam.

to achieve the maximum gain and includes the degradation due to the losses and nonuniformities of the structure and the nonuniformity in the illumination of the aperture. If an antenna is perfect and the energy propagation is uniform, A_e equals the actual projected area A, and η is unity. For practical applications, η varies between 0.5 and 0.8.

The next question is how to determine the amount of power that an effective aperture is able to focus in a given frequency band. This is a complex problem, but the result is simple and elegant. From the theory of antennas [10], the gain of an antenna is related to the effective aperture by

$$g = \frac{4\pi A_e}{\lambda^2} \tag{1.8}$$

where λ is the wavelength of the RF. This demonstrates that the gain of an antenna depends on its effective aperture and operating frequency. The higher the frequency, the higher the gain of the antenna for a given aperture size. This relationship indicates that there are inherent limitations in terms of the achievable gain for a given size of the aperture. Moreover, the surface of the aperture needs to be accurate to within small fractions of λ, making higher-frequency structures more expensive. Another useful relationship can be derived by considering the solid angle Ω within which an antenna focuses the power. This can be readily derived from equation (1.8) as

$$\Omega = \frac{\lambda^2}{A_e} \text{ rad}^2 \tag{1.9}$$

and

$$g = \frac{4\pi}{\Omega}$$

For a satellite-based system, once the antenna footprint is specified (i.e., the coverage area is delineated), the solid angle to cover that area from the satellite gets fixed. This implies that the antenna gain (and therefore the aperture size) is determined by the coverage area. In other words, the limit on the achievable gain from the satellite is set by the coverage area and not by the design or physical structure of the antenna. This relationship also shows that in order to achieve a high directivity, that is, a smaller solid angle, the area of the aperture must be much larger than the operating wavelength. In the microwave region, the physical apertures required to achieve relatively high gains are reasonable and readily implemented. For such antennas, the radiated fields do not interact strongly with nearby objects that are not in the direction of the transmission. Thus, the ground effects, which play a major role in the design of megahertz region (AM, FM, and TV broadcasts) antennas, are generally not important above 1 GHz.

1.4.2.3 Power Flux Density

The radiation of EM energy follows the fundamental inverse square law. Consequently, energy, radiated by the isotropic antenna to a distance r, is uniformly distributed over a sphere of area $4\pi r^2$, as shown in Figure 1.8a. Note that the energy is independent of the frequency of the radiation. The power received per unit area is simply $p_0/4\pi r^2$, where p_0 is the amount of power at the input of the isotropic source. This value is defined as pfd:

$$\text{pfd} = \frac{p_0}{4\pi r^2} \text{ W/m}^2 \tag{1.10}$$

1.4.2.4 Power Radiated and Received by Antennas in a Communication System

Consider the power transmitted and received via radio transmission in an ideal communication channel. Let

p_T be the power of the transmitter in watts

g_T be the transmit antenna gain

g_R be the receive antenna gain

A_{eT}, A_{eR} be the effective apertures of the transmit and receive antennas in meters

r be the distance between the transmitter and receiver in meters

Then power transmitted $= p_T \times g_T$ W. The pfd in the desired direction at distance r is given by

$$\text{pfd} = \frac{p_T g_T}{4\pi r^2} \text{ W/m}^2 \tag{1.11a}$$

The product $p_T g_T$ is called the *equivalent isotropic radiated power* (EIRP), an important factor in link calculations. The pfd represents the received power that is intercepted by an ideal antenna with an aperture of 1 m^2, expressed in decibels

$$\begin{aligned}\text{PFD} &= P_T + G_T - 10\log 4\pi r^2 \\ &= \text{EIRP} - 10\log 4\pi r^2 \text{ dBW/m}^2\end{aligned} \tag{1.11b}$$

where $P_T = 10\log p_T$, $G_T = 10\log g_T$, and $\text{PFD} = 10\log(\text{pfd})$. This equation relates that no matter what, in a radio transmission, a very large portion of the transmitted energy, represented by the term $10\log 4\pi r^2$, is not seen by the receiver. Also, this loss is independent of the frequency, a point often lost if the equation is manipulated in terms of the gains of the transmit and receive antennas. To clarify this point, equation (1.11) can be expressed in terms of power p_r received at distance r as follows:

$$\begin{aligned}p_r &= (\text{pfd})_t \times A_{eR} = \frac{p_T g_T}{4\pi r^2} g_R \frac{\lambda^2}{4\pi} \\ &= p_T g_T g_R \left(\frac{\lambda}{4\pi r}\right)^2 \text{ W}\end{aligned} \tag{1.12a}$$

In decibels, this is

$$P_R = P_T + G_T + G_R - 20\log\frac{4\pi r}{\lambda} \text{ dBW} \tag{1.12b}$$

This equation indicates that the power received has frequency and distance dependent terms. The combined dependence arises because of the receive antenna, required to focus the received energy to a point source. In other words, the term $20\log(4\pi r/\lambda)$, often referred to as the *range loss*, represents the power loss between two isotropic antennas within a particular range and at a particular frequency. As far as spreading of the radio energy from a point source is concerned, the energy follows the classic inverse square law, independent of the frequency, described by equation (1.11). Regardless, either equation (1.11) or (1.12) can be used for link calculations.

Example 1.1

In a microwave radio relay link, repeater stations are located 50 km apart. They are equipped with 10 W HPAs and antennas with a gain of 30 dB. Assuming the transmission line and filter losses to be 2 dB, compute (1) the EIRP of the transmitter and the pfd at the receive antenna and (2) given an antenna efficiency of 0.55, calculate the diameter of a circular parabolic antenna to realize the 30 dB gain in the 4 GHz frequency band.

Solution:

1) EIRP $= 10\ \text{dBW} + (-2\ \text{dB}) + 30\ \text{dB}$
   $= 38\ \text{dBW}$
   PFD $= \text{EIRP} - 10\log 4\pi r^2$
   $= 38 - 105$
   $= -67\ \text{dBW/m}^2\ (\text{or}\ 200\ \text{nW/m}^2)$
   Thus, the aperture of a square meter area, located 50 km apart from the transmitter, intercepts 200 nW of RF power, radiated by the transmit antenna.
2) The aperture to realize a specific antenna gain is derived from

$$A = \frac{G}{\eta}\frac{\lambda^2}{4\pi}\ \text{m}^2$$

   The power received by the aperture depends on the RF even though the pfd remains unchanged. For a circular parabolic antenna with a diameter D, the aperture area A equals $(\pi/4) \times D^2$. By expressing the antenna gain in decibels, and the diameter in meters, we obtain

$$G = 20\log D + 20\log f_{\text{MHz}} + 10\log\eta - 39.6\ (\text{dB})$$

For $f = 4000$ MHz, $G = 30$ dB, and $\eta = 0.55$, D is calculated to be 1 m.

PART II Noise in a Communication Channel

1.5 Noise in Communication Systems

Noise in its broadest definition consists of any undesired signal in a communication circuit. Noise represents the fundamental constraint in the transmission capacity of a communication system. It is also of interest to the ITU and domestic radio regulation agencies. An essential element of radio regulations is the specification of allowable levels of radiation to other existing or proposed systems. Such a restriction to control the interference between communication systems is essential in a multisystem environment. Without the restriction, it is not possible to design a reliable communication system. Typically, the regulating agencies specify the allowable level of radiation outside the intended geographic area and allocated frequency spectrum. Also, the regulations include the interference guidelines of competing services and systems. The noise sources external to a communication system are as follows:

1) Cosmic radiation, including that from the sun
2) Anthropogenic (synthetic, caused by humans) noise such as power lines, electrical machinery, consumer electronics, and other terrestrial sources
3) Interference from other communications systems.

There is little to be done about the cosmic noise except to ensure that it is considered in the system design. Generally, anthropogenic noise occurs at low frequencies and is rarely a problem for communication systems operating above 1 GHz. Interference from other communication systems is strictly controlled by ITU and domestic agencies. Typically, this requires control of transmitter powers, antenna patterns, generation, and suppression of frequencies outside the

allocated spectrum. The regulations ensure that these sources of unwanted radiation are kept to a minimum and significantly below the levels of noise inherent in the design of a communication system. The dominant sources of noise in a communication system include the following:

- Interference from adjacent copolarized channels
- Interference from adjacent cross-polarized channels
- Multipath interference
- Thermal noise
- Intermodulation (IM) noise
- Noise due to channel imperfections.

These noise sources are now reviewed.

1.5.1 Adjacent Copolarized Channel Interference

One thing that is common to all communication systems is the channelization of the frequency spectrum. Channelization allows flexibility in meeting multidestination traffic requirements, as well as in maximizing the communication capacity of the system. In a subsequent section, modulation schemes employed for long-haul transmission are described. The most commonly used modulation scheme is frequency modulation (FM). One property of FM is that it produces sidebands of decreasing magnitudes extending to infinity. Consequently, there is no escape from some energy spilling over in the adjacent bands causing distortion. To a degree, this spillage can be controlled by the choice of channelization filters, highlighting the criticality of filter networks in a communication system. Figure 1.9 is a pictorial representation of adjacent channel interference.

1.5.2 Adjacent Cross-Polarized Channel Interference

The nature of EM radiation allows the polarization of energy in a given direction. This property is exploited in communication systems by employing orthogonal polarization of antenna beams. It allows frequency reuse that doubles the available bandwidth. Polarization can be linear or circular. The primary limitation is the level of cross-polarization isolation that is achievable in practical antenna networks. Therefore, this interference is totally dependent on and controlled by the antenna design. Typical cross-polarization isolation in practical systems is of the order 27–30 dB. It should be noted that polarization can be altered when EM radiation propagates through the atmosphere. This also needs to be accounted for in the design.

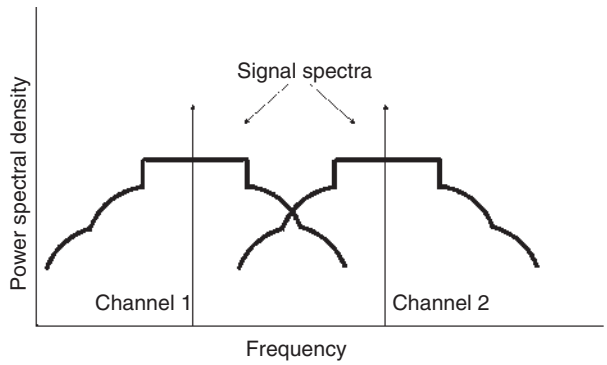


Figure 1.9 Adjacent channel interference.

1.5.3 Multipath Interference

This distortion mechanism is caused by obstacles in the spread energy field of the transmitter where the energy is reflected in the direction of the receiver. Such obstacles can be tall buildings or trees or foliage in an urban environment. Also, distortion can stem from refractive index gradients and other atmospheric effects. Interference occurs when the reflected rays from obstacles or ground are received by the receiver. These interfering signals are time-displaced echos of the original transmission. Figure 1.10 is a pictorial description of multipath interference. In a fixed LOS environment, the radio paths are optimized by taking such obstructions into account. This is not the case when it comes to mobile communication systems. The mobile terminal can be fixed or in motion. Even a handheld terminal might have micromotion. When a terminal is in motion, the path characteristics are constantly changing and multipath propagation tends to become the limiting factor. Mobile units experience multipath scattering, reflection, and diffraction by various obstructions and buildings in the vicinity. Although constructive and destructive fading can become quite complex, there are ways to deal with this problem, including frequency and space diversity and forward error correction. Regardless of these compensation techniques, multipath interference continues to be the principal cause of missed calls, fading, and disruptions in mobile communications. The limitation of the available bandwidth for this service further exacerbates the problem. This topic and other issues related to radio propagation are described by Freeman [7].

1.5.4 Thermal Noise

Thermal noise permeates all communication systems and is the ultimate limit to their performance. For this reason, thermal noise is addressed in more depth. Thermal noise is the name given to the electrical noise arising from the constant agitation of molecules and its constituents in a conductor. This agitation at the atomic level is a universal characteristic of all matter. Molecules consist of a nucleus and a cloud of electrons around it. The nucleus consists of neutrons and protons, and the number of protons is equal to the number of electrons. Electrons

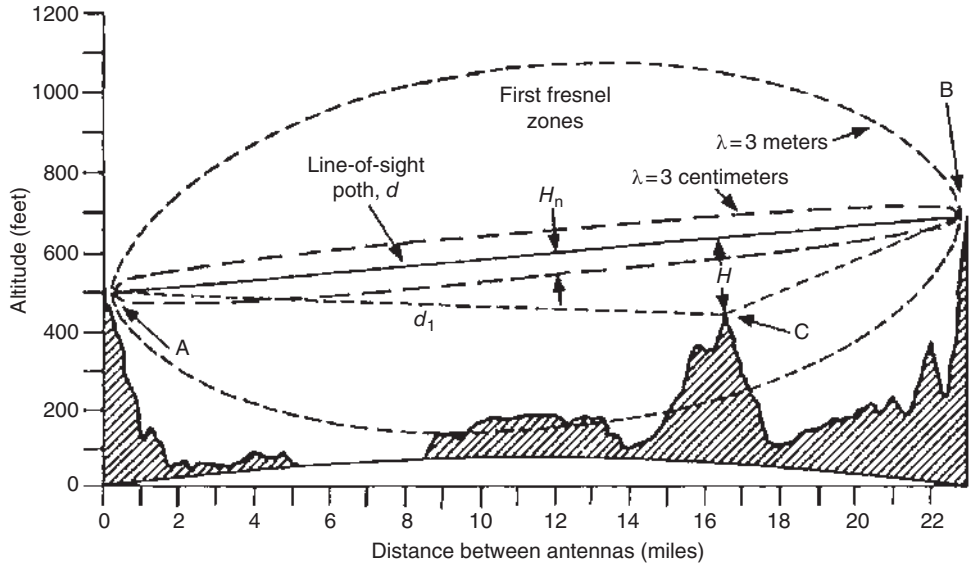


Figure 1.10 Multipath interference. (*Source*: Reproduced with the permission of Bell Telephone Laboratories, Inc.)

and positive ions (atoms with one or more electrons removed) are, on an average, distributed uniformly in a conductor, rendering the structure electrically neutral. The electrons in a conductor are in continual random motion, and in thermal equilibrium with the molecules. There is kinetic energy associated with this motion, and it increases with temperature. Since each electron carries a unit negative electric charge, each flight of an electron between collisions with molecules constitutes a short pulse of current. The random agitation of the large number of electrons creates statistical fluctuations from the neutrality that translates into electrical noise. The mean square velocity of electrons is directly proportional to the absolute temperature. The equipartition law due to Boltzmann and Maxwell (and the works of Johnson and Nyquist) states that for a thermal noise source, the available power in a 1 Hz bandwidth is given by

$$p_n(f) = kT \text{ W/Hz} \tag{1.13}$$

where k is the Boltzmann constant $= 1.3805 \times 10^{-23}$ J/K and T is the absolute temperature of the thermal noise source in degrees Kelvin. At a room temperature of 17°C or 290 K, the available power is $p_n(f) = 4.0(10^{-21})$ W/Hz or −174.0 dBm/Hz. The result given by the equipartition theory is one of a constant power density spectrum versus frequency. Because of this property, a thermal noise source is referred to as a *white noise source*, an analogy to white light that contains all the visible wavelengths of light. In all the reported measurements, the available power of a thermal noise source has been found to be proportional to the bandwidth over any range, from direct current to the highest microwave frequencies. If the bandwidths are unlimited, the results of the equipartition theory states that the available power of a thermal noise source should also be unlimited. Obviously, this is not possible. The reason is that the equipartition theory is based on classical mechanics, which tends to break down as the higher frequencies are approached. If the principles of quantum mechanics are applied to the problem, kT must be replaced by $hf/[\exp(hf/kT) - 1]$, where h is Planck's constant $= 6.626 \times 10^{-34}$ J s. By applying this result to the expression for available power of a thermal noise source [1], we obtain

$$p_n(f) = \frac{hf}{\exp(hf/kT) - 1} \text{ W/Hz} \tag{1.14}$$

This relationship demonstrates that at arbitrarily high frequencies, the thermal noise spectrum eventually drops to zero, but this does not mean that noiseless devices can be built at these frequencies. A quantum noise term equal to hf needs to be added to equation (1.14) in this case. Figure 1.11 offers a plot of the noise density as a function of frequency. The transitional region occurs at about 40 GHz for $T = 3$ K, at 400 GHz for $T = 30$ K, and at 4000 GHz at room temperature. For most practical purposes, the available noise power of a thermal noise source can be assumed to be directly proportional to the product of the bandwidth of the system or detector and the absolute temperature of the source.

Thus

$$p_a = kTB \text{ W} \tag{1.15a}$$

where B is the noise bandwidth of the system or detector in hertz and p_a is the available noise power in watts. By assuming an ambient temperature of 290 K and expressing the available noise power in dBm gives

$$p_a = -174 + 10 \log B \text{ dBm} \tag{1.15b}$$

This represents the minimum amount of noise power that must ultimately limit the SNR of a signal. This relationship represents an average value; it does not tell us anything about the statistical distribution. As stated earlier, thermal noise is attributed to the random motions of electrons in the conductors. Therefore, thermal noise might be regarded as the superposition

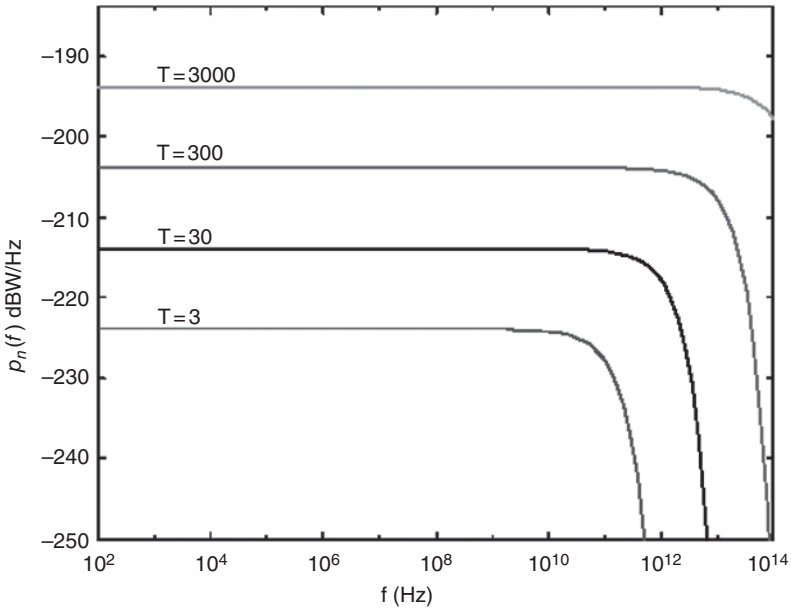


Figure 1.11 Thermal noise power densities as a function of frequency.

of an exceedingly large number of independent electronic contributions. It is well known in the field of statistics that the limiting form for the distribution function of the sum of a large number of independent quantities that can have various distributions is a Gaussian function. This result is known in statistics as the *central limit theorem*. Consequently, thermal noise satisfies the theoretical conditions for a Gaussian function and must follow this distribution. The Gaussian probability density function for the zero mean is reflected in Figure 1.12a and its equation is

$$p(V) = \frac{1}{\sigma_n \sqrt{2\pi}} \exp\left(\frac{-V^2}{2\sigma_n^2}\right) \tag{1.16}$$

where V represents the instantaneous voltage and σ_n the standard deviation. The Gaussian distribution function is shown in Figure 1.12b and is given by the integral of equation (1.16)

$$P(V) = \frac{1}{\sigma_n \sqrt{2\pi}} \int_{-\infty}^{V} \exp\left(\frac{-x^2}{2\sigma_n^2}\right) dx \tag{1.17}$$

It can be readily shown that the mean square voltage (the expected value of V^2) is equal to the variance, σ_n^2. Thus, the rms (root mean square) voltage of the Gaussian distributed noise source is given by σ_n, the standard deviation.

Gaussian noise has a probability greater than zero that exceeds any finite magnitude no matter how large it is. As a result, the peak factor, given by the ratio of peak to rms voltage, does not exist for a thermal noise signal. For this particular case, it is convenient to modify the definition of the peak factor to be the ratio of the value, exceeded by the noise a certain percentage of the time to the rms noise value. This percentage of time is commonly chosen to be 0.01%. A table of the normal distribution indicates that signal magnitudes greater than $3.89\sigma_n$ (i.e., $|V| > 3.89\sigma_n$) occur less than 0.01% of the time.

Since σ_n is the rms value of the noise signal, the peak factor for a thermal noise signal is 3.89, or 11.80 dB. Inclusion of 0.001% peaks increases the peak factor by 1.1 dB, raising it to a value of 12.9 dB. The fact that thermal noise is white, as well as Gaussian, has led many engineers into

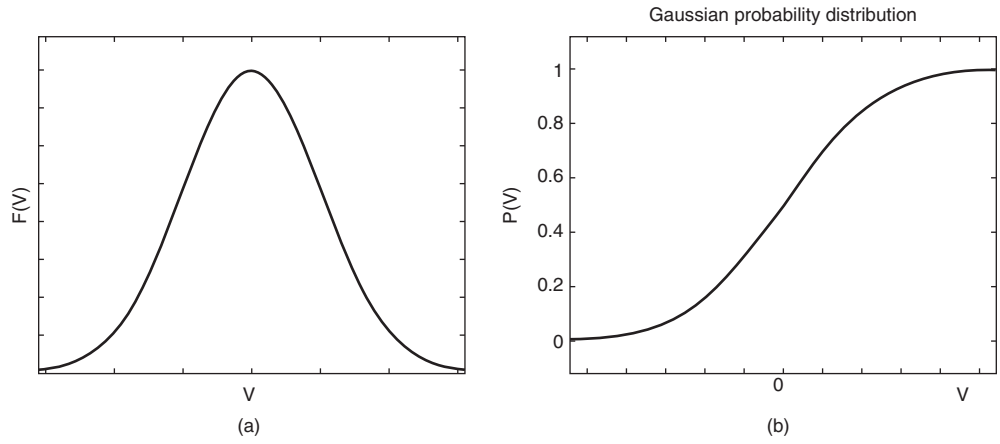


Figure 1.12 (a) Gaussian density and (b) distribution functions.

carelessly viewing white and Gaussian noise as synonymous, which is not always the case. For example, passing Gaussian noise through a linear network such as a filter will leave it Gaussian but can drastically change the frequency spectrum. A single impulse does not exhibit a Gaussian amplitude distribution but has a flat or white frequency spectrum.

1.5.4.1 Effective Input Noise Temperature

Since the available noise power of a thermal noise source is directly proportional to the absolute temperature of the source, an equivalent noise temperature can be attributed to it [1]. For resistive elements, the noise temperature is equal to the physical temperature of the resistor, that is, if a given noise source produces an available power of p_a watts in a small frequency interval of df hertz, the noise temperature of the noise source is given by $T = p_a/k\ df$. It should be emphasized that the concept of noise temperature does not have to be restricted to the noise sources alone and that the noise temperature does not have to equal the physical temperature of the source. For example, consider an antenna. The output noise is simply the noise collected by the aperture due to the radiating elements in the antenna's field of view. The physical temperature of the antenna has no bearing on this, and the noise temperature can still be used to define the noise power from the antenna.

Consider a two-port network with an available gain of $g_a(f)$. When it is connected to a noise source having a noise temperature of T, the available noise power in a small band df at the output of the network is $p_{\text{no}} = g_a(f)kT\ df + p_{\text{ne}}$. This power consists of two components: (1) power due to the external noise source, $g_a(f)kT\,df$ and (2) the power due to the internal noise sources of the network p_{ne}, which is the available noise power of the network when the input of the network is connected to a noise-free source. The effective noise temperature T_e of this equivalent representation of the internal noise source of the noisy network is then given by

$$T_e = \frac{p_{\text{ne}}}{g_a(f)k\ df} \tag{1.18}$$

The available noise power at the output of the network in terms of the effective input noise temperature now becomes

$$p_{\text{no}} = g_a(f)k(T + T_e)df \tag{1.19}$$

The effective input noise temperature T_e can vary as a function of the frequency, depending on how g_e and p_{ne} vary. The concept of the effective input noise temperature is useful when the source noise temperature differs from that of the standard temperature. It has a distinct advantage when the noise performance of a complete communication system is being evaluated. Another concept that is useful in analyzing active devices and communication systems is the noise figure.

1.5.4.2 Noise Figure

The IRE (Institute of Radio Engineers, precursor to IEEE) definition of the noise factor for a two-port network is as follows: "The noise figure (noise factor) at a specified input frequency is the ratio of the total noise power per unit bandwidth at a corresponding output frequency, available at the output when the noise temperature of the input source is standard (290 K); to that portion of this output power engendered at the input frequency by the input source." In terms of this definition

$$\text{Noise figure} = n_F = \frac{p_{\text{no}}}{g_a(f)kT_0df} \tag{1.20}$$

where $T_0 = 290\,\text{K}$ and is referred to as the standard temperature. A noise figure such as that described for a narrowband df is called a *spot noise figure*, which can vary as a function of frequency. The noise figure can also be related to the effective noise temperature. In equation (1.19), if T is replaced by T_0, as required in the definition of the noise figure, the output noise power p_{no} is then given by $g_a(f)k(T_0 + T_e)df$. The output noise in terms of the noise figure is given by equation (1.20). Equating these two expressions yields the relationship between the noise figure and the effective noise temperature as

$$n_F = 1 + \frac{T_e}{T_0} \tag{1.21}$$

and

$$T_e = T_0(n_F - 1) \tag{1.22}$$

The concept of the noise figure is most useful when the input source has a noise temperature approximately equal to the standard temperature. Consider the expression for the noise power at the output of the network as given by equation (1.20).

By rewriting this equation in terms of dBm (reference 1 mW), we have

$$P_{\text{no}} = N_F + G_a + 10\log df - 174\ \text{dBm} \tag{1.23}$$

The symbols are defined as follows:

$$P_{\text{no}} = \frac{10\log p_{\text{no}}}{10^{-3}}$$

$$N_F = 10\log n_F$$

$$G_a = 10\log g_a \tag{1.24}$$

Thus, the available noise power of the two-port network in dBm can be written as the sum of the noise power of the thermal noise source in dBm, the available gain of the network in dB, and the noise figure of the network in dB. As a result, the effects of internal noise sources of a two-port network are accounted for by adding the noise figure in dB to the available noise power of the source in dBm.

1.5.4.3 Noise of a Lossy Element

Any signal is attenuated by the lossy elements in its path. Lossy elements absorb energy, thereby raising the level of agitation of the molecules in the element resulting in additional noise. By adopting arguments similar to the preceding sections, the effective input noise temperature of a lossy element is derived as [1]

$$T_e = T(l_a - 1) \tag{1.25}$$

where l_a is the loss ratio of the element. The loss L_a of the element in dB is given by $L_a = 10 \log l_a$. The noise figure of the lossy element is then given by

$$n_F = 1 + \frac{T}{T_0}(l_a - 1) \tag{1.26}$$

If the lossy element is at the standard temperature T_0, then

$$n_F = l_a \quad \text{and} \quad N_F = L_a \text{ dB} \tag{1.27}$$

As an example, for a transmission line with a loss of 1 dB at room temperature, the effective input noise temperature is given by $T_e = 75$ K, and the noise figure is simply 1 dB.

1.5.4.4 Attenuators

Attenuators are used in communication systems to control power levels of different channels or transponders. Such networks can be constructed using lossy elements or reactive elements or a combination of both. If composed of ideal and entirely reactive elements (implies zero resistive component), an attenuator will not contribute any noise and will have an effective input noise temperature of zero. However, loss due to such attenuators must be included when dealing with a chain of network elements as described subsequently in equation (1.29). Practical reactive elements always have an associated resistive component, however small, and that contributes noise in the channel. The key point to remember is that it is the resistive loss associated with any device that contributes to system noise temperature.

Example 1.2

What is the amount of thermal noise power radiated by a source at a room temperature of 290 K in a bandwidth of 50 MHz? What are the peak values of this noise that do not exceed 0.01% and 0.001% of the time?

Solution:

Using equation (1.14), the thermal noise is given by

$$\begin{aligned} N_T &= kTB \\ &= -228.6 + 10 \log T + 10 \log B \quad \text{dBW} \end{aligned}$$

For $T = 290$ K and $B = 50 \times 10^6$ Hz, we have

$$\begin{aligned} N_T &= -127 \text{ dBW} \\ &= 0.2 \text{ pW} \end{aligned}$$

As a result, an antenna that is pointed at a 290 K source, where the source fills the antenna beamwidth, receives 0.2 pW of noise power in a bandwidth of 50 MHz. As described in Section 1.5.4, the peak factors for thermal noise not exceeding 0.01% and 0.001% of the time are 11.8 and 12.9 dB, respectively. As a consequence, the peak value of the noise can be as high as—115.2 dBW (3.0 pW) for 0.01% of the time and—114.1 dBW (3.9 pW) for 0.001% of the time.

Example 1.3
What is the noise power at the output of a LNA with a noise figure of 2 dB and a gain of 30 dB over a bandwidth of 500 MHz? What is the equivalent noise temperature of the LNA?

Solution:
Equations (1.20)–(1.22) lead to

$$\begin{aligned} N_{\text{LNA}} &= N_F + G_a + 10\log df - 174 \text{ dBm} \\ &= (2 \text{ dB}) + (30 \text{ dB}) + 10\log 500 \times 10^6 - 174 \\ &= -55 \text{ dBm} \end{aligned}$$

The equivalent noise temperature is given by

$$\begin{aligned} T_{\text{LNA}} &= T_0(n_F - 1) \\ &= 290(1.585 - 1) \\ &= 169.6 \text{ K} \end{aligned}$$

1.5.5 Noise in Cascaded Networks

The two networks connected in tandem in Figure 1.13 have effective input temperatures T_{e1} and T_{e2} and available gains g_1 and g_2.

Suppose that these two tandem networks are connected to a noise source with a noise temperature of T. In a small frequency band df, the noise power at the output due to the noise source alone is $g_1 g_2 kT\,df$. The noise power due to noise sources in the first network is $g_1 g_2 kT_{e1}\,df$ and in the second network is $g_2 kT_{e2}\,df$. The total noise appearing at the output of the second network is $kg_2(g_1 T + g_1 T_{e1} + T_{e2})df$. Since the portion of this noise due to the noise sources internal to the two networks is $kg_2(g_1 T_{e1} + T_{e2})\,df$, the effective input temperature T_e of the two networks in tandem is therefore given by

$$\begin{aligned} T_e &= \frac{kg_2(g_1 T_{e1} + T_{e2})df}{g_1 g_2 k\,df} \\ &= T_{e1} + \frac{T_{e1}}{g_1} \end{aligned} \tag{1.28}$$

This result can be easily generalized to n networks in tandem. The resulting effective input noise temperature is

$$(T_e)_1 = T_{e1} + \frac{T_{e2}}{g_1} + \frac{T_{e3}}{g_1 g_2} + \cdots + \frac{T_{en}}{g_1 g_2 \cdots g_{n-1}} \tag{1.29a}$$

It should be noted in equation (1.29a) that the reference point for the effective noise temperature is the input terminal of the first element in the chain. This relationship can be readily modified to compute the effective noise temperature with reference to any terminal within the chain. For example, if the chosen reference is the input port of element 3(T_{e3}), then the effective

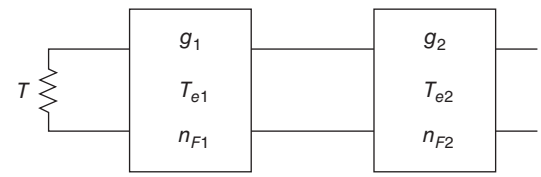


Figure 1.13 Noise in cascaded networks.

noise temperature is given by

$$(T_e)_3 = T_{e1}g_1g_2 + T_{e2}g_2 + T_{e3} + \frac{T_{e4}}{g_3} + \frac{T_{e5}}{g_3g_4} + \cdots \tag{1.29b}$$

Equation (1.29) is known as *Friis* formula, named in honor of H.T. Friis.

This relationship is advantageous in system calculations where the selected reference is typically the input terminal of the LNA. Such a reference provides the evaluation of the figure of merit defined by the ratio of the effective antenna gain to the noise temperature (G/T).

From the relationship between the noise figure and effective input noise temperature, it can be easily demonstrated that the resulting noise figure of n stages in tandem is

$$(n_F)_1 = n_{F1} + \frac{n_{F2} - 1}{g_1} + \cdots + \frac{n_{Fn} - 1}{g_1g_2 \cdots g_{n-1}} \tag{1.30}$$

The significance of these relationships lies in the fact that noise contribution after an amplifier in the chain is reduced by the gain of the amplifier. Typical amplifiers have gain in excess of 20 dB; this implies that the noise contribution by the elements after the amplifier in the chain will be reduced by a factor of 100. This is an important consideration in the design of a communication channel as well as multistage amplifiers. Example 1.4 illustrates the significance of these relationships.

Example 1.4

Calculate the noise temperature contributions in a 6 GHz, 500 MHz bandwidth receive section as denoted in Figure 1.14. Assume that the receive antenna has a noise temperature of 70 K.

Solution:

The noise temperatures and associated gain/loss ratios for the various elements are computed as follows:

Feed Network:

$$l_1 = 1.5849$$

$$g_1 = \frac{l}{l_1} = 0.631$$

$$T_{e1} = 290(l_1 - 1) = 169.6 \text{ K}$$

Bandpass Filter:

$$l_2 = 1.1885$$

$$g_2 = \frac{l}{l_2} = 0.8414$$

$$T_{e2} = 290(l_2 - 1) = 54.7 \text{ K}$$

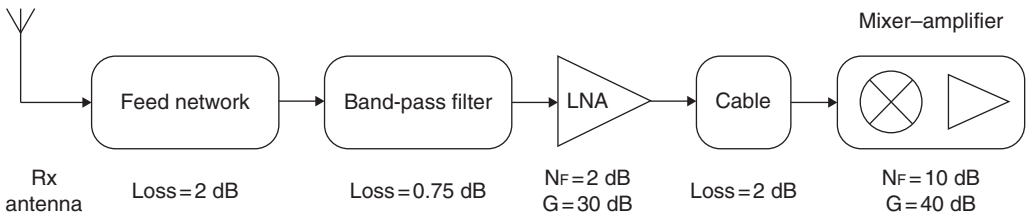


Figure 1.14 Example of noise calculations of a receive network.

LNA:

$$n_{F3} = 1.5849$$

$$g_3 = 1000$$

$$T_{e3} = 290(n_{F3} - 1) = 169.6 \text{ K}$$

Cable:

$$l_4 = 1.5849$$

$$g_4 = \frac{1}{14} = 0.631$$

$$T_{e4} = 290(l_4 - 1) = 169.6 \text{ K}$$

Mixer Amplifier:

$$n_{F5} = 10$$

$$g_5 = 10{,}000$$

$$T_{e5} = 290(n_{F5} - 1) = 2610 \text{ K}$$

The total system noise temperature referred to the input of the LNA is

$$\begin{aligned}(T_e)_{\text{sys}} &= (T_{\text{ant}} + T_{e1})g_1g_2 + T_{e2}g_2 + T_{e3} + \frac{T_{e4}}{g_3} + \frac{T_{e5}}{g_3g_4} \\ &= 127.2 + 46.0 + 169.6 + 0.17 + 4.14 \\ &= 347.1 \text{ K}\end{aligned}$$

It is interesting to note the relatively large contribution to the overall system noise temperature by the losses prior to the LNA and the relatively insignificant noise contributions by components after the LNA. These are important considerations in the design of communication systems, emphasizing the need to minimize losses prior to the LNAs in the receive section of a communication channel.

1.5.6 Intermodulation (IM) Noise

IM noise is generated by the nonlinearities present in communication systems. Similar to thermal noise, nonlinearities are present to some degree in all electrical networks. Nonlinearities can limit the useful signal levels in a system and thus become an important design consideration. The device that is the primary source of IM is the nonlinear HPA. It is an essential element of a communication system. The HPA's efficiency and the degree of nonlinearity are inversely related; thus, the HPA characteristics and operating power levels are important design parameters.

Consider the voltage transfer characteristics of a general two-port, which can be a device, network, or system, as depicted in Figure 1.15. For a two-port memory-less nonlinear network, the transfer function is described by the Taylor series expansion:

$$e_0 = a_1e_i + a_2e_i^2 + a_3e_i^3 + \cdots \tag{1.31}$$

For a single-frequency sinusoid input, $e_1 = A\cos(ax)$, we obtain

$$\begin{aligned}e_0 &= a_1A\cos ax + a_2A^2\cos^2 ax + a_3A^3\cos^3 ax + \cdots \\ &= K_0 + K_1\cos(ax) + K_2\cos(2ax) + K_3\cos(3ax) + \cdots\end{aligned} \tag{1.32}$$

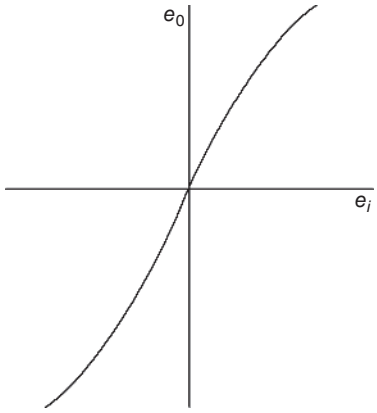


Figure 1.15 Transfer characteristics of a nonlinear two-port network.

where the K values are constants related to $a_1, a_2, a_3, \ldots$. Therefore, a single sinusoid input leads to an output containing the fundamental frequency and its harmonics. Similarly, if $e_i = A\cos\omega_{1i} + B\cos\omega_{2i} + \cos\omega_{3i} + \cdots$, then by using trigonometric identities, we obtain

$$e_0 = K_0 + K_{1i}f(\omega_i) + K_{2i}f(2\omega_i) + K_{3i}f(3\omega_i) + \cdots \tag{1.33}$$

where

$$\begin{aligned} f(\omega_i) &= \text{collection of first-order terms containing } \omega \\ f(2\omega_i) &= \text{second-order terms such as } 2\omega_1, \omega_1 + \omega_2, \omega_1 - \omega_2 \\ f(3\omega_i) &= \text{third-order terms such as } 3\omega_1, 2\omega_1 \pm \omega_2, 2\omega_2 \pm \omega_3, \omega_1 + \omega_2 + \omega_3, \\ &\quad \omega_1 + \omega_2 - \omega_3, \ldots \\ K &= \text{a distinct constant associated with each term} \end{aligned}$$

Consequently, the output contains harmonics and all possible combinations of the sum and difference frequencies of the input signals. The dc (direct-current) term is of no interest and is filtered out. The desired output corresponds to the linear case, given by the first-order products K_{1i} due to term a_1. All the other outputs are spurious and contribute to the objectionable noise and interference. As the number of input signals increases, the number of IM products grows rapidly. These IM products fall either inside or outside of the RF channel, depending on the order of the product and the separations between the carrier frequencies. In the classic case of a single channel per carrier (SCPC), frequency-division multiplex (FDM-FM) system, IM products are so widespread that they resemble flat Gaussian or thermal noise. As a result, for RF channels employing a large number of carrier frequencies, there is a trade-off between thermal noise and IM noise. As the number of carriers increases, thermal noise per carrier is smaller, whereas the IM noise level is larger. There is some point of optimum loading of the wideband FM transmitter where the thermal noise and the IM noise in the channel when combined represent a minimum of the total system noise, as described in Figure 1.16. This provides guidance for the best operating power level for the amplifier. Typically, the third-order IM product is dominant and often forms part of the specification. Amplifiers are designed to have a carrier to third-order product values exceeding 20 dB. This topic is discussed further in Sections 1.8.3 and 1.8.4.

1.5.7 Distortion Due to Channel Imperfections

An ideal transmission channel transmits all signals without distortion over a certain bandwidth and completely attenuates all signals outside this bandwidth. Such a performance is characterized by a channel that provides a constant loss (ideally zero) and linear phase (i.e., constant

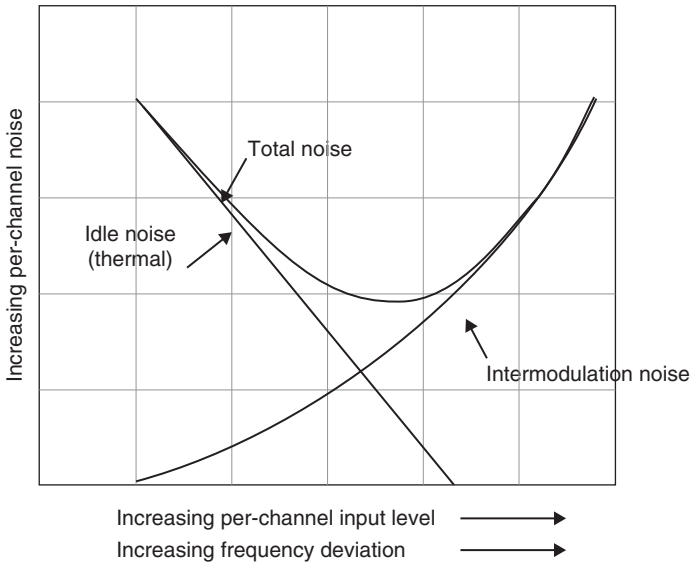
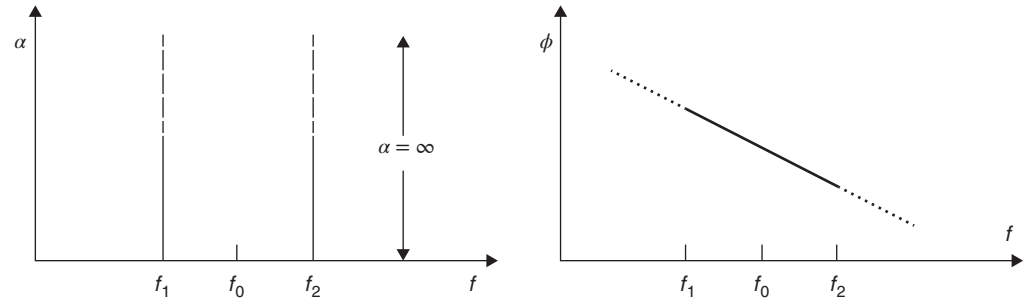


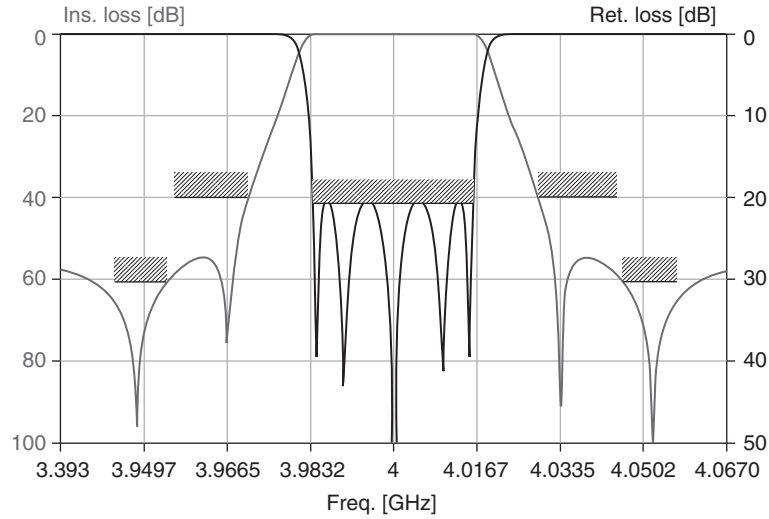
Figure 1.16 Effect of loading and thermal noise in a communication channel.

delay) over the passband and infinite attenuation outside it, as shown in Figure 1.17a and b. Such a filter performance is not feasible. The unit impulse response of such a filter exists for negative values of time, violating the condition of causality [11].

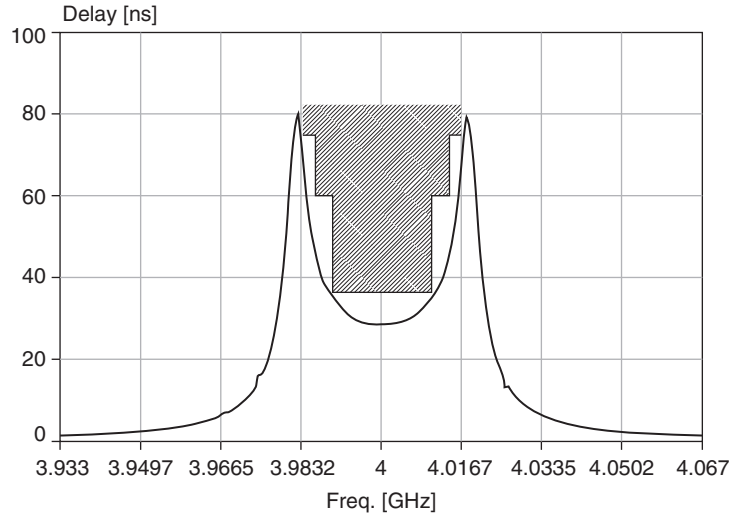
Although ideal filter characteristics are not feasible, it is possible to approach these characteristics as closely as desired. The characteristics of a practical filter, typical of narrowband channels in communication systems, are plotted in Figure 1.17c and d. There exists a trade-off between the design complexity and its departure from ideal filter characteristics. Other components in the communication system, including amplifiers, frequency converters, modulators, cables, and waveguides, are wideband devices, exhibiting a minimal deviation from the flat amplitude and group delays over the narrow channel bandwidths. In other words, filter networks represent the controlling features of amplitude and phase characteristics of a transmission channel. This has been a key driver for filter research and development for many decades.

All practical filters exhibit transmission deviations. Filters are passive and linear devices. Their response is time-invariant and has amplitude and phase shapes that are prescribed functions of frequency. Unlike nonlinear devices such as HPAs, no new frequencies are produced when an FM signal is passed through the filter. However, filters do change the relative amplitude and phase of the carrier and sidebands, which, in turn, is interpreted by the demodulator as additional modulation, causing distortion in the received signal. For insight into this process, consider an FM signal with many sidebands applied to a network with ideal transmission characteristics except at the frequency of one of the sideband components. The amplitude of this component gets altered and is equivalent to adding an extraneous signal to the applied FM signal. As a result, the demodulated output signal consists of the desired signal, which is proportional to the input modulating signal and the undesired signal. When many carriers are present, it can be demonstrated that the transmission deviations in an FM system can introduce baseband frequency components at the output that do not exist at the input. In this sense, the transmission deviations in an FM system have an effect similar to that of an amplifier nonlinearity. For that reason, the distortion introduced by transmission deviations is often called *IM noise*. There is no exact solution to the problem of determining IM noise produced by transmission deviations. An approximate approach for analog systems that has proved successful





(c)



(d)

Figure 1.17 Transmission channel characteristics: (a) an ideal amplitude response; (b) an ideal phase response; (c) a practical amplitude response; (d) a practical group delay response.

is described in Ref. [1] and is based on previous works [12–14]. The analysis assumes that the transmission characteristics of the channel are sufficiently smooth so that the gain and phase as a function of frequency can be represented by

$$Y_n(\omega) = [1 + g_1(\omega - \omega_c) + g_2(\omega - \omega_c)^2 + g_3(\omega - \omega_c)^3 + g_4(\omega - \omega_c)^4] \times e^{i[b_2(\omega-\omega_c)^2 + b_3(\omega-\omega_c)^3 + b_4(\omega-\omega_c)^4]} \tag{1.34}$$

where

ω_c = carrier frequency in radians per second (rad/s)

g_1, g_2, g_3, g_4 = linear, parabolic, cubic, and quartic gain coefficients, respectively

b_2, b_3, b_4 = parabolic, cubic, and quartic phase coefficients, respectively

This assumption is consistent with the frequency response that can be readily achieved by microwave filter networks. The second assumption is that the FM signal has a sufficiently low modulation index and that the bandwidth of the channel is much smaller than the carrier frequency. This is indeed the case for most communication systems. From these assumptions, it is possible to compute the distortion due to transmission deviations. Another related source of distortion is introduced by the HPAs that follow filter networks. The nonlinearity of amplifiers converts the amplitude variations introduced by filters to phase variations, causing distortion of the FM signal. If a limiter can successfully remove the amplitude variations before the signal reaches the device, it can be dismissed. A summary of these distortion terms and noise power (included in Appendix 1.A) is described in Ref. [15]. These values are applicable for analog FM transmission.

For digital systems, the impact of transmission deviations is relatively small, unless the data rate is very high. For advanced digital modulation schemes, sophisticated simulation tools are needed to compute distortion as a function of the amplitude and phase deviations in the RF channel. It leads to trade-offs between in-band response (transmission deviations) and out-of-band attenuation, consistent with microwave filter technology and other system design parameters [16]. Such trade-offs characterize the RF channels. In effect, RF channel filters control the amplitude and phase characteristics of a communication channel, that is, the channel filters define the effective usage of the available channel bandwidths.

It should be noted that for analog FM transmission, variations in the amplitude slope of a filter followed by a nonlinear amplifier cause AM-to-PM conversion, resulting in intelligible (coherent) cross talk between FM carriers. With digital modulation, cross talk is unintelligible (noncoherent), but there is still a modulation transfer, causing an increase in the required E_b/N_0 (ratio of energy per bit to noise density) in the link. Digital transmission is relatively insensitive to variations in group delay unless, of course, it is very large and the data rate is high (short symbol duration). In most cases, the effect of the group delay slope is small on the link E_b/N_0.

1.5.8 RF Link Design

A communication link is characterized by a number of RF links in tandem. In this section, we describe the carrier-to-noise ratio (CNR) for a single link and the impact of cascading a number of such links.

1.5.8.1 Carrier-to-Noise Ratio (*C*/*N*)

The carrier-to-noise ratio (C/N) of a signal is simply given by C/N, where C is the carrier power and N is the total noise present in a given bandwidth. If N_0 is the noise density as defined by the noise power in a bandwidth of 1 Hz, then, by virtue of equation (1.15), we have

$$N_0 = kT_s \quad \text{and} \quad N = N_0 B = kT_s B \tag{1.35}$$

where k is the Boltzmann constant, T_s is the total effective noise temperature, and B is the bandwidth of the carrier frequency. In a communication system, T_s is chosen as the total effective temperature referenced to the input of the LNA of the receive system. It includes the receive antenna noise temperature, lossy transmission lines, bandpass filter prior to the LNA, and the noise temperature of the LNA itself. If G_R is the receive antenna gain at the same reference (i.e., including the losses up to LNA), the carrier power at the reference point is [see equation (1.12b)]

$$C = \text{EIRP} + G_R - P_L \tag{1.36}$$

where $P_L = 20\log(4\pi r/\lambda)$ is referred to as the *path loss* between the transmit and receive antennas. Therefore,

$$\frac{C}{N} = \text{EIRP} + G_R - P_L + 228.6 - 10\log T_s - 10\log B$$

$$= \text{EIRP} + \frac{G}{T_s} - P_L + 228.6 - 10\log B \tag{1.37}$$

This represents the link equation for the thermal noise of the system. The factor $G_R - 10\log T_s$ or G/T_s represents the figure of merit of the receiving system. The link equation can also be expressed in the alternative form

$$\frac{C}{N_0} = \text{EIRP} + \frac{G}{T_s} - P_L + 228.6 \tag{1.38}$$

$$\frac{C}{T_s} = \text{EIRP} + \frac{G}{T_s} - P_L \tag{1.39}$$

All the terms in the link equations (1.37–1.39) are in decibels.

1.5.8.2 Multiple RF Links in Tandem

A communication channel, consisting of a number of RF links, is described in Figure 1.18.

Since the n links in tandem are physically different, the noise power generated in each of the links is not coherent with those generated in the other links. Thus, the noise power of all the links can be summed up as independent power sources to attain the total end-to-end noise power contribution and the overall C/N. The noise power of each link can be referenced to the carrier level, which can be normalized to unity. Therefore, by adding all the noise floors N_i, the overall carrier-to-noise ratio $(C/N)^*$ is given by

$$\left[\left(\frac{C}{N}\right)_T^*\right]^{-1} = \left[\left(\frac{C}{N}\right)_1^*\right]^{-1} + \left[\left(\frac{C}{N}\right)_2^*\right]^{-1} + \cdots + \left[\left(\frac{C}{N}\right)_n^*\right]^{-1} \tag{1.40}$$

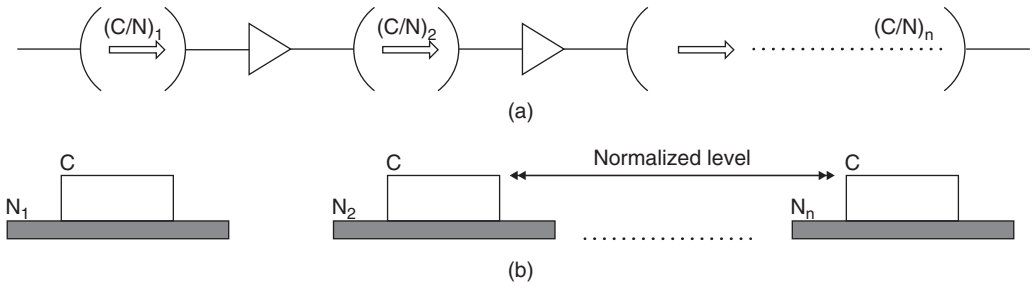


Figure 1.18 A communication channel depicting (a) RF Links in tandem and (b) normalized thermal noise of each link.

The terms $(C/N)_1^*$, $(C/N)_2^*$ signify that C/N is expressed in terms of ratios, as opposed to decibels.

Example 1.5

A transmit earth station has a 1 kW output power and an antenna with a gain of 55 dB. The transmission line and filter losses between the amplifier and antenna amount to 2 dB. Using Example 1.4 as reference for the receive network at the satellite

1) Compute the value of EIRP, the pfd, and the C/N_0 for the 6 GHz uplink for a satellite located at 40,000 km from the surface of the earth. Assume that the receive antenna at the satellite has a gain of 25 dB and include a margin of 3 dB due to losses through the atmosphere and antenna pointing accuracy.
2) Compute the value of the carrier to thermal noise at the satellite for a 36 MHz RF channel.
3) Compute the value of the combined uplink–downlink thermal noise, assuming that the power radiated by the satellite to the earth station in the 4 GHz downlink beam results in a CNR of 20 dB. What is the impact on the total noise if the uplink power is reduced by 10 dB?

Solution:

1) $$\begin{aligned}\text{Uplink EIRP} &= 30\ \text{dBW} + (-2\ \text{dB}) + 55\\ &= 83\ \text{dBW}\\ \text{PFD} &= \text{EIRP} - 10\log 4\pi r^2\\ &= 83 - 163\\ &= -80\ \text{dBW/m}^2\end{aligned}$$

This represents a power density of 10 nW per square meter at the satellite located in the geostationary orbit. From equation (1.38), we have

$$\begin{aligned}\frac{C}{N_0} &= \text{EIRP} - 3 - 20\log\frac{4\pi r}{\lambda} + G_R - L + 228.6 - 10\log T_s\\ &= 83 - 3 - 20\log\frac{4\pi\times 40\times 10^6}{(3\times 10^8)/(6\times 10^9)} + 25 - 2.75 + 228.6 - 10\log 347.1\\ &= 83 - 3 - 200 + 25 - 2.75 + 228.6 - 25.4\\ &= 105.45\ \text{dB Hz}\end{aligned}$$

2) $$\frac{C}{N} = \frac{C}{N_0} - 10\log 36\times 10^6 = 29.9\ \text{dB}$$

3) The combined value of the carrier to thermal noise, in terms of ratios, is given by

$$\left[\left(\frac{C}{N}\right)_T^*\right]^{-1} = \left[\left(\frac{C}{N}\right)_{\text{UL}}^*\right]^{-1} + \left[\left(\frac{C}{N}\right)_{\text{DL}}^*\right]^{-1}$$

$$\left(\frac{C}{N}\right)_{\text{UL}}^* = 10^{2.99} \quad \text{and} \quad \left(\frac{C}{N}\right)_{\text{DL}}^* = 10^2$$

The total thermal noise ratio is, therefore, given by

$$\begin{aligned}\left[\left(\frac{C}{N}\right)_T^*\right]^{-1} &= 10^{-2.99} + 10^{-2}\\ &= 0.001 + 0.01\\ &\approx 0.01\end{aligned}$$

For the ratio in decibels (dB), we obtain $(C/N)_T = 20$ dB.

If the uplink power is reduced by 10 dB, the $(C/N)_{UL}$ is reduced to 19.9 dB, and the carrier to total thermal noise power is

$$\left[\left(\frac{C}{N}\right)_T^*\right]^{-1} = 10^{-1.99} + 10^{-2}$$
$$= 0.102 + 0.01$$
$$= 0.0202$$

Since this represents a CNR of 16.95 dB, the impact on the total noise is no longer negligible.

1.6 Modulation–Demodulation Schemes in a Communication System

In a communication system, the baseband signal, consisting of a large number of individual message signals, needs to be transmitted over the communication medium, separating the transmitter from the receiver. The efficiency of the transmission requires that this information be processed in some manner before it is transmitted. Modulation is the process whereby the baseband signal is suitably impressed on a carrier signal to increase its efficiency for transmission over the medium. Modulation can shift the signal frequencies to facilitate transmission or change the bandwidth occupancy, or it can alter the form of the signal to optimize the noise or distortion performance. At the receiver, a demodulation scheme reverses this process. Comprehensive coverage of this topic is described in references [2, 3, 6, 11].

Modulation techniques are categorized as linear or nonlinear, depending on whether the modulated signal varies linearly (i.e., the superposition holds) or varies nonlinearly with the message.

There are two forms of modulation: amplitude modulation and angle (phase or frequency) modulation. The process of modulation is represented by

$$M(t) = a(t)\cos[(\omega_c t + \phi(t)] \tag{1.41}$$

Here $a(t)$ represents the amplitude of the sinusoidal carrier and $\omega_c t + \phi(t)$ is the phase angle. Although both the amplitude and angle modulation can be present simultaneously, an amplitude-modulated system is one in which $\phi(t)$ is a constant and $a(t)$ is made proportional to the modulating signal. Similarly, in an angle-modulated system, $a(t)$ is held constant and $\phi(t)$ is rendered proportional to the modulating signal.

1.6.1 Amplitude Modulation

For the amplitude modulated wave, we have

$$M(t) = a(t)\cos\omega_c t \tag{1.42}$$

where the carrier is at frequency f_c and $a(t)$ is the modulating time function. If $a(t)$ is a single-frequency sinusoid of unit amplitude at frequency f_m, then $a(t) = \cos\omega_m t$, and the modulated wave is

$$M(t) = \cos\omega_m t \cos\omega_c t$$

This can be expanded to

$$M(t) = \frac{1}{2}\cos(\omega_c - \omega_m)t + \frac{1}{2}\cos(\omega_c + \omega_m)t \tag{1.43}$$

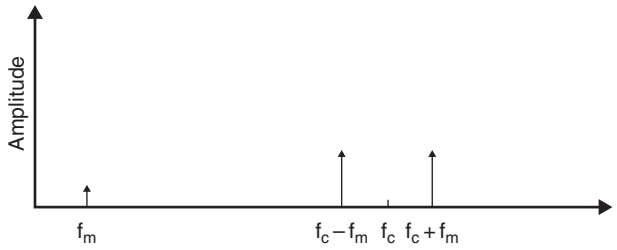


Figure 1.19 Amplitude-modulated carrier with a single sinusoid modulating frequency.

The modulated wave contains no component at the original carrier and only a sideband on either side of the carrier, spaced f_m hertz from the carrier, as reflected in Figure 1.19.

The effect of the product modulation is to translate $a(t)$ in the frequency domain so that it is reflected symmetrically about f_c. It can be shown that this is true for a complex waveform as well. If either sideband is rejected by a filter, the result is a single-sideband (SSB) signal, representing a pure frequency translation. A more general representation of the amplitude modulation is

$$M(t) = [1 + ma(t)] \cos \omega_c t \tag{1.44}$$

This expression is equivalent to adding a dc term of magnitude unity. Also, it is imperative that

$$|ma(t)| < 1 \tag{1.45}$$

so that the envelope of modulated wave remains undistorted, as denoted in Figure 1.20.

Here, m is defined as the modulation index with a maximum value of unity, representing 100% modulation. It represents the relative magnitude of the modulating wave for a carrier frequency of unity magnitude. The modulated wave is derived as

$$M(t) = \cos \omega_c t + \frac{m}{2} \cos(\omega_c - \omega_m)t + \frac{m}{2} \cos(\omega_c + \omega_m)t \tag{1.46}$$

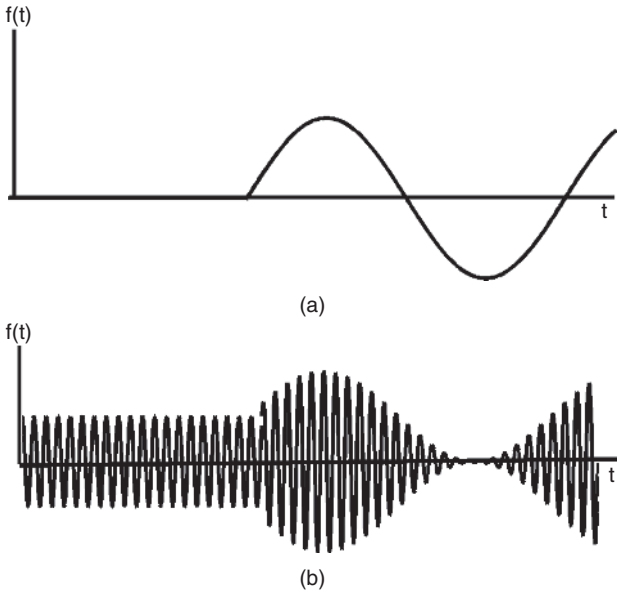


Figure 1.20 Amplitude modulation of a carrier: (a) modulating signal; (b) amplitude-modulated carrier.

The average power in each sideband is $m^2/4$ or a total sideband power of $m^2/2$ W. Thus, for 100% modulation, only one-third of the total power is in the information bearing sideband. For complex signals, the power in the sidebands is considerably less, amounting to a few percent of the total power. A second drawback of the AM wave is its sensitivity to amplitude deviations in the signal path. Any such deviation can result in distortion of the signal. Furthermore, for AM signals, linear HPAs are necessary for signal amplification to avoid excessive distortion of the signal. There are limits to obtaining practical linear amplifiers with an adequate gain and power, rendering AM impractical for long-haul transmissions. However, one substantial advantage of AM is the preservation of the bandwidth. As a consequence, it finds application for frequency translation and in multiplexing individual message channels by translating them to higher frequencies. A good example of amplitude modulation is in the formation of a baseband signal with a large number of voice channels.

1.6.2 Formation of a Baseband Signal

A message channel is composed of a large number of independent signals multiplexed to form a composite signal that occupies a continuous spectral range, referred to as the *bandwidth* of the signal. A hierarchy has been established for North American telephony communication systems to allow standardization and a high degree of commonality. The basic message channel, although originally intended for voice transmission, has also been adopted for data transmission. The basic group consists of 12 channels, each 4 kHz wide extending over the band 60–108 kHz. The equipment can be considered as a series of modulators with distinct carrier frequencies, followed by the appropriate bandpass filters and then multiplexed to form the composite signal as shown in Figure 1.21.

Here, modulation is amplitude modulation, and the single sideband is extracted via the bandpass filters. Amplitude modulation is a linear process, although the mixer to modulate the carrier is inherently a nonlinear device. Here, the mixer is used in a quasi-linear mode and simply translates the modulation frequency symmetrically about the carrier frequency. This group of 12 channels, occupying a 48 kHz bandwidth, is a basic building block in the Bell system. The next step in the FDM hierarchy is the combination of five groups in a 60-channel supergroup, for a bandwidth of 240 kHz, occupying a 312–552 kHz frequency band. Modern broadband transmission systems are capable of larger and larger systems, extending to groups of up to 3600 message channels. Although the grouping described here is generally used for long-haul Bell systems, it is not a universal standard. Along with telephony channels, television channels occupy bandwidths of 4–6 MHz or data channels ranging from kilobits to several megabits per second range, or Internet traffic. Therefore, a typical composite signal, referred to as a *baseband signal*, can vary from a simple basic group to a combination of groups of voice

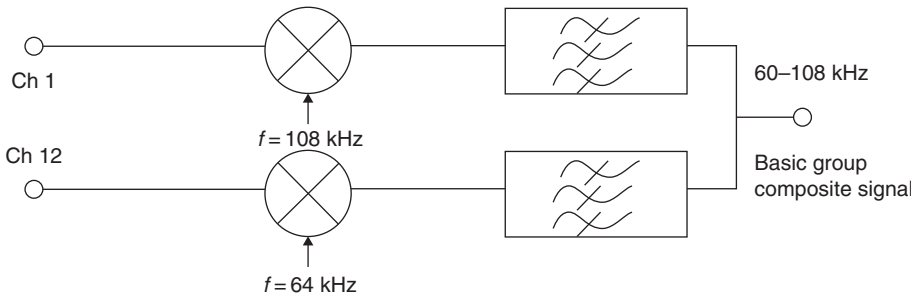


Figure 1.21 Formation of a message signal.

channels, television signals, and data channels. The signal composition is determined by the traffic requirements of the system. The hierarchy of data rates exists all the way up to optical fiber rates of 40 Gbps (gigabits per second).

1.6.3 Angle-Modulated Signals

The modulated wave for angle-modulated signals is represented by

$$M(t) = A\cos[\omega_c t + \phi(t)] \tag{1.47}$$

Phase modulation (PM) is defined as angle modulation in which the instantaneous phase deviation $\phi(t)$ is proportional to the modulating signal voltage. Frequency modulation is angle modulation (FM) in which the carrier varies with the integral of the modulating signal. The average power for a PM or an FM wave is proportional to the square of the voltage and results in

$$\begin{aligned} P(t) &= A_c^2\cos^2[\omega_c t + \phi(t)] \\ &= A_c^2\left[\frac{1}{2} + \frac{1}{2}\cos(2\omega_c t + 2\phi(t))\right] \end{aligned} \tag{1.48}$$

The second term is assumed to consist of a large number of sinusoids about the carrier frequency $2f_c$, with the average value of zero and

$$P_{\text{av}} = \frac{A_c^2}{2} \tag{1.49}$$

Thus, the average power of a FM wave is the same as the average power in the absence of modulation. This represents a major advantage, compared with the AM, where the average power of information bearing sidebands is a third or less than that of the average power of the carrier frequency. However, there are no free lunches, and this power advantage comes at the expense of increased bandwidths. FM frequency analysis is quite complicated and beyond the scope of this book. A brief account of the key results and underlying assumptions are now described.

1.6.3.1 Spectra of Analog Angle-Modulated Signals

Narrowband FM Phase and FM are special cases of angle modulation, and one form is easily derived from the other. The discussion is confined to FM, owing to its wider range of applications for both analog and digital communication systems. A frequency-modulated carrier is shown in Figure 1.22.

The modulating signal is assumed to be a repetitive sawtooth of period T, where $(2\pi/T) \ll \omega_c$. As the sawtooth modulating signal increases in magnitude, the FM signal oscillates more rapidly, resulting in the widening of the frequency spectrum. Note that its amplitude remains unchanged.

Analysis of the FM process is inherently much more complicated than that for AM, due to the nonlinearity of the FM process. Assume a sinusoidal modulating signal $v(t)$ at frequency f_m:

$$v(t) = a\cos\omega_m t \tag{1.50}$$

The instantaneous radian frequency ω_i is then

$$\omega_i = \omega_c + \Delta\omega\cos\omega_m t, \quad \Delta\omega \ll \omega_c \tag{1.51}$$

where $\Delta\omega$ is a constant, depending on the amplitude a. As a result, the instantaneous radian frequency varies about the unmodulated carrier frequency ω_c at the rate ω_m of the modulating signal and with a maximum deviation of $\Delta\omega$ radians. The phase variation $\theta(t)$ for this special

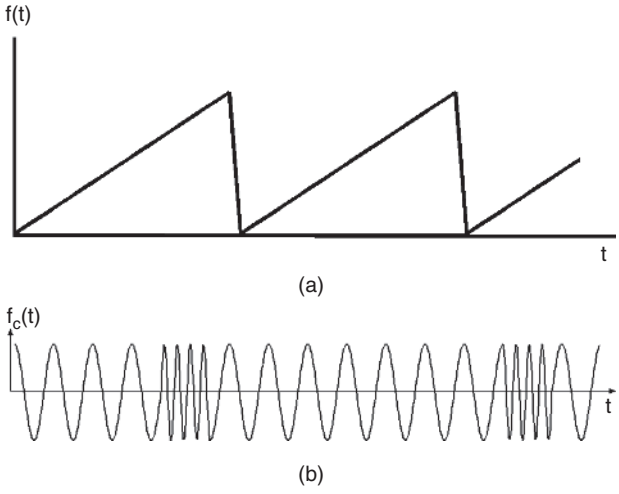


Figure 1.22 Frequency modulation: (a) modulating wave; (b) modulated FM carrier.

case is computed by

$$\theta(t) = \int \omega_i \, dt = \omega_c t + \frac{\Delta\omega}{\omega_m} \sin \omega_m t + \theta_0 \tag{1.52}$$

where θ_0 is a constant, providing a reference for the phase. By choosing it to be zero, we see that the frequency-modulated carrier is

$$M(t) = \cos(\omega_c t + m \sin \omega_m t) \tag{1.53}$$

with

$$m = \frac{\Delta\omega}{\omega_m} = \frac{\Delta f}{f_m}$$

where m is the modulation index and is given by the ratio of the frequency deviation to the baseband bandwidth. The modulated carrier frequency is written in its expanded form as $M(t) = \cos \omega_c t \cos(m \sin \omega_m t) - \sin \omega_c t \sin(m \sin \omega_m t)$. For $m \ll \pi/2$, we have

$$\begin{aligned} M(t) &\approx \cos \omega_c t - m \sin \omega_m t \sin \omega_c t \\ &\approx \cos \omega_c t - \frac{m}{2}[\cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t] \end{aligned} \tag{1.54}$$

Under this condition, the system is referred to as *narrowband FM* and has a form similar to that of AM carriers. It contains the original unmodulated carrier and two sideband frequencies, displaced $\pm\omega_m$ radians from ω_c. The bandwidth of the narrowband FM signal is thus $2f_m$, like that of an AM signal. Despite this similarity, there is a clear distinction between AM and the narrowband FM; in the FM case, the carrier amplitude is constant whereas for the AM, the amplitude varies in accordance with the modulating signal. The carrier and sideband terms are in phase for the AM, whereas the sidebands are in phase quadrature with the carrier for narrowband FM case.

Wideband FM The case where $m > (\pi/2)$ is referred to as *wideband FM*. Its analysis requires the expansion of $M(t)$ given by equation (1.53). In the equation, the terms $\cos(m \sin \omega_m t)$ and $\sin(m \sin \omega_m t)$ are periodic functions of ω_m and are expanded in a Fourier series of period

$2\pi/\omega_m$. This expansion can be expressed in terms of the Bessel functions [3]:

$$\begin{aligned} M(t) = {} & J_0(m)\cos\omega_c t - J_1(m)[\cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t] \\ & + J_2(m)[\cos(\omega_c - 2\omega_m)t + \cos(\omega_c + 2\omega_m)t] \\ & - J_3(m)[\cos(\omega_c - 3\omega_m)t - \cos(\omega_c + 3\omega_m)t] \\ & + \cdots \end{aligned} \tag{1.55}$$

Equation (1.55) represents a time function, consisting of a carrier and an infinite number of sidebands, spaced at frequencies, $\omega_c \pm \omega_m, \omega_c \pm 2\omega_m$, and so on.

These successive sets of sidebands are called *first-order sideband*, *second-order sideband*, and so on, the magnitudes of which are determined by the coefficients $J_1(m), J_2(m), \ldots$ respectively.

When two or more sinusoids are present in the modulating signal, the spectrum contains not only multiples of individual modulating frequencies but also all the possible sums and differences of multiples of the modulation frequencies. As more and more modulating signals are added, the complexity of the solution increases rapidly. Eventually, the baseband signal is represented by random noise, extending uniformly across the baseband from 0 to f_m hertz. The frequency spectrum of the corresponding FM signal appears as a continuum of sidebands. The magnitudes of the carrier and sideband terms depend on m, the modulation index, expressed by the appropriate Bessel function. From a theoretical viewpoint, the bandwidth required for 100% of the signal energy is infinite. For practical systems, only significant sidebands with a magnitude of at least 1% of the magnitude of the unmodulated carrier are considered. The number of significant sidebands varies with m and can be determined from the tabulated values of the Bessel function. For the large value of $m(> 10)$, the minimum bandwidth is given by $2\Delta f$, where Δf is the peak deviation. A general rule, postulated by J. R. Carson in 1939 for the minimum bandwidths of an FM signal, is

$$B_T \approx 2[f_m + \Delta f] \tag{1.56}$$

This is an approximate rule (referred to as *a rule of thumb*) and is appropriate for most applications. The actual bandwidth required is, to some extent, a function of the waveform of the modulating signal and the quality of transmission desired. It can be seen from this expression that for $m < 1$, the minimum bandwidth is given by $2f_m$, and for $m > 10$, the minimum bandwidth is $2\Delta f$. A more accurate assessment of bandwidth can be made by computing the coefficients of the Bessel function to any desired magnitude for a given modulation index. For modulation indices between 1 and 10, bandwidth given by equation (1.56) corresponds to the case of including sidebands with at least 10% of the magnitude of the unmodulated carrier frequency. For large values of m (>10), the number of significant sidebands in FM wave is equal to m. A key point is that wideband FM systems require greater bandwidth compared with that of AM systems, the extent of which is determined by the modulation index.

1.6.4 Comparison of FM and AM Systems

So far, only ideal AM and FM systems have been investigated. It was shown that an AM system is a linear process and conserves the bandwidth, but the modulation process transfers only one third of the input power to the information bearing sidebands, the remainder power remains at the carrier frequency. An FM system, however, is a nonlinear process. Modulation produces new frequencies, and the modulated signal requires much larger bandwidths. From a theoretical standpoint, the energy of an FM signal is dispersed over an infinite bandwidth. However, the bulk of the energy is contained in the first few sidebands.

For most FM signals, 99% of the energy is contained in a bandwidth of $2(f_m + \Delta f)$, where Δf is the peak deviation and f_m is the baseband frequency. The advantage of the FM system is that all

of the input power is transferred to the information-bearing sidebands. The average power at the carrier frequency, after modulation, is zero. Another advantage is that the amplitude of the FM envelope is nearly constant and, thus, the signals are amplified with a minimum of distortion by the practical nonlinear amplifiers. The critical parameter for any modulation scheme is the S/N ratio under different traffic conditions. For an AM system, this ratio is expressed by [3]

$$\frac{S}{N} = \frac{A_c^4}{8N^2 + 8NA_c^2} \tag{1.57}$$

where A_c is the voltage amplitude of the unmodulated carrier and N is the mean noise power. The CNR is

$$\frac{C}{N} = \frac{A_c^2}{2N} \tag{1.58}$$

Therefore

$$\frac{S}{N} = \frac{1}{2}\ \frac{(C/N)^2}{1 + 2(C/N)} \tag{1.59}$$

For $C/N \ll 1$, the output SNR drops as the square of the CNR. This is the suppression that is characteristic of envelope detection. For CNR $\gg 1$

$$\frac{S}{N} = \frac{1}{4}\frac{C}{N} \tag{1.60}$$

The output SNR is then linearly dependent on the C/N, another characteristic of envelope detectors. In addition, the relation shows that no SNR improvement is possible for AM systems. An increase in the transmission bandwidth $2f_m$, needed to pass the AM signals, serves only to increase the noise N, decreasing the output SNR.

For a FM system, the SNR is given by [3],

$$\frac{S}{N} = 3\left(\frac{\Delta f}{B}\right)^2 \frac{C}{2N_0 B} \tag{1.61}$$

Here, C is the average power of the FM carrier and $\Delta f/B$ is the modulation index m. Denoting $2N_0B = N$, the average noise power in the sidebands, we obtain

$$\frac{S}{N} = 3m^2\frac{C}{N} \tag{1.62}$$

Compare the FM and AM systems by assuming the same unmodulated carrier power and noise spectral density n_0 for both. For a 100% modulated FM signal, the C/N corresponds to the CNR of an AM system as described by equation (1.49), that is, $(S/N)_{\text{AM}} = C/N$. Equation (1.62) can be modified to read as

$$\left(\frac{S}{N}\right)_{\text{FM}} = 3m^2\left(\frac{S}{N}\right)_{\text{AM}} \tag{1.63}$$

For a large modulation index (corresponding to a wide transmission bandwidth, with $m \gg 1$), the SNR can be increased significantly over that of the AM case. For example, if $m = 5$, the FM output SNR is 75 times that of an equivalent AM system. Alternatively, for the same SNR at the output in both receivers, the power of the FM carrier can be reduced 75 times, but this requires an increase in the transmission bandwidth from $2B$ (AM case) to $16B$ (FM case). Note that this bandwidth is greater than that given by the rule of thumb criteria, as it is based on using sidebands with at least 1% magnitude of the carrier. Frequency modulation provides a substantial improvement in the SNR, but at the expense of increased bandwidth. This, of course, is a characteristic of all noise improvement systems.

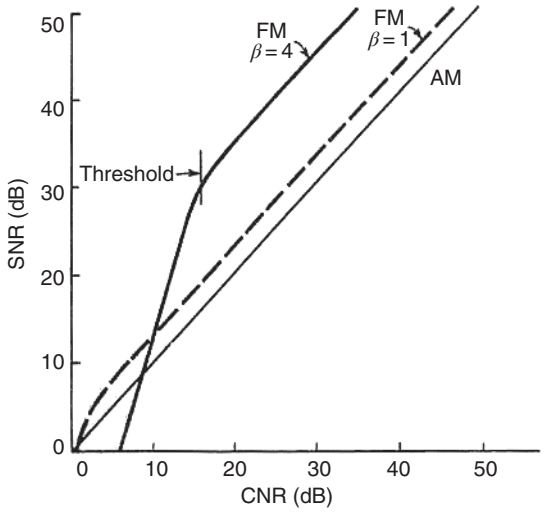


Figure 1.23 FM threshold effect. (*Source*: Schwartz 1980 [3]. Reproduced with the permission of McGraw-Hill Professional.)

Is it possible to continue increasing the output SNR indefinitely by increasing the frequency deviation, and the corresponding bandwidth? If the transmitter power is fixed, increasing the frequency deviation increases the required bandwidth with it, incurring more noise. Eventually, the noise power at the limiter becomes comparable with the signal power and the noise is found to "take over" the system. The output SNR falls off much more sharply than the input CNR. This effect is called the threshold effect, as described in Figure 1.23. For proper operation of FM systems, the CNR must be kept above the threshold value, typically >13 dB.

1.7 Digital Transmission

The widespread use of digital communication systems is the result of many factors. They include the relative simplicity of digital circuit design, the ease with which integrated circuit techniques can be applied to digital circuitry and the rapid advances in digital signal processing (DSP) techniques. The information content of a digital signal consists of discrete states, such as the presence or absence of a voltage, characterized as 1 or 0. This implies that a simple decision circuit can be employed as a regenerator, that is, a corrupted digital signal enters on one side and a clean perfect signal comes out the other side. Accumulated noise on the corrupted signal stops at the regenerator and does not accumulate as the signal passes through different stages of the communication channel. Using coding techniques (at the expense of small increase in bandwidth) to minimize the effect of noise in the detection of 1s and 0s at the regenerating stations (repeaters) at appropriate intervals allows almost error-free digital transmission over long distances. Digital techniques allow development of efficient compression techniques by eliminating statistical redundancies, removing unnecessary information such as pauses in voice communication, or removing small invisible parts of a picture or redundancies between images in video transmission. Compression techniques increase transmission capacity and more efficient utilization of the frequency spectrum (conservation of bandwidth). However, compression of information implies additional hardware and latency. Compression techniques have

matured to the point where digital systems now require less bandwidth than analog signals. In the early era of satellite communication, one 36 MHz transponder carried one analog TV channel; same transponder can now carry 10 digitally compressed channels. Not only that, the 10 digital channels can be combined into 1 signal to allow the transmitters to operate near saturation. It is apparent that digital communication systems require considerable amount of electronic circuitry. Such circuitry used to be expensive but nowadays electronics circuitry in the form of very-large-scale integrated (VLSI) circuits has become inexpensive. Although cost used to be a major factor in selecting analog communication over digital communications in the past that is no longer the case. Digital networks have therefore become the overwhelming choice for communication systems.

1.7.1 Sampling

The well-known Nyquist criterion states that "If a message that is magnitude–time function is sampled instantaneously at regular intervals, and at a rate at least twice the highest significant message frequency, then the samples contain all of the information of the original message."

This rather amazing result is fundamental to the ability to digitize analog signals without the loss of information. As an example, a message band limited to f_m hertz is completely specified by its amplitudes at any set of points in time spaced T seconds apart, where $T = \frac{1}{2}f_m$, as shown in Figure 1.24. The result implies that the bandwidth required to convert an analog signal into a digital representation is at least twice the bandwidth of the highest frequency in the signal. The amplitude-modulated pulse signal can then be transmitted to a receiver in any form that is suitable from a transmission standpoint. At the receiving end, a reverse process is followed to recreate the original pulse amplitude-modulated signal. To recover the original message, it is necessary to pass the impulses through an ideal low-pass filter with a cutoff of f_m. The output of this filter is a replica of the original message, delayed in time. Since information can be digitized without loss of accuracy, the challenge is then to exploit the potential for DSP for communications systems. The first step in this process is the quantization of the sampled signals.

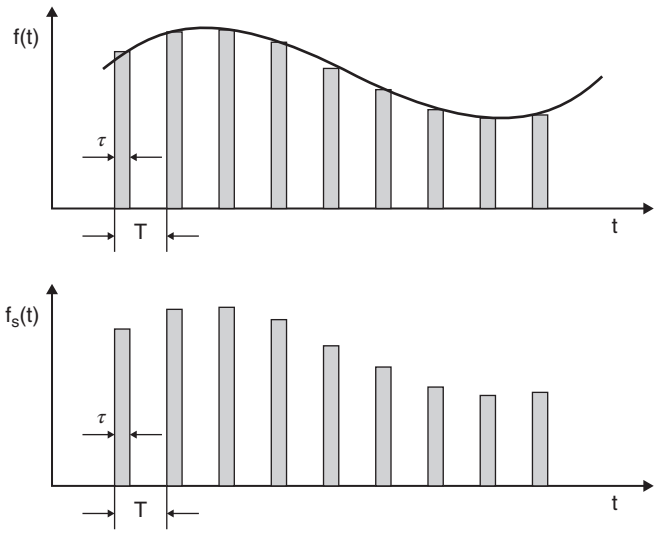


Figure 1.24 The sampling process (τ = sampling time; $T = 1/f_c$: sampling interval): (a) input function $f(t)$ and (b) sampled output $f_s(t)$.

1.7.2 Quantization

The process of quantization consists of breaking the amplitudes of the signals into a prescribed number of discrete amplitude levels. Then, when the message is sampled in a pulse amplitude-modulated (PAM) system, the discrete amplitude nearest the true amplitude is sent. Consequently, the process of quantization introduces an error in the representation of the amplitudes of the sampled signals. Moreover, this error is irretrievable. This seemingly purposeful distortion of the signal is kept below the noise introduced during transmission and at the receiver. In essence, the uncertainty introduced by the fundamental thermal noise, and noise due to imperfections in circuits and devices, limits the ability to distinguish between all the possible amplitude levels, thus making quantization possible. The advantage of quantization is that once the number of discrete amplitude levels is established, each level can then be coded in some arbitrary form before transmission. Thus, quantization makes it possible to deploy the full potential of DSP techniques to optimize the information flow in a communication system.

1.7.3 PCM Systems

Systems that embody the transmission of digitized and coded signals are commonly called *pulse-code-modulated* (PCM) systems. Binary digital systems constitute the most frequently encountered form of PCM systems. A quantized sample can be sent as a simple pulse with certain possible discrete amplitudes. However, if many discrete samples are required, design circuits become quite complicated and uneconomical. On the contrary, if several pulses are used as a code group to describe the amplitude of a sample, then each pulse can have only two states. In a binary system, a code group of m on/off pulses can be used to represent 2^m amplitudes. For example, eight pulses yield 2^8 or 256 amplitude levels. The m pulses must be transmitted in the general sampling interval allotted for the quantized sample. This constraint necessitates an increase in the bandwidth by a factor of m.

For the digital transmission of a 4 kHz voice channel, using 256 levels of quantization and binary code requires a bandwidth of $4 \times 2 \times 8$ or 64 kHz, a factor of 16 times greater than that of an analog system. The required bandwidth can be reduced by choosing a smaller number of amplitude levels for sampling or by using a code where the pulses can be represented by more than two amplitude levels. A quantized signal sample can be coded into a group of m pulses, each with n possible amplitude levels. Thus, if the signal is quantized into M possible amplitude levels, then, $M = n^m$. Each combination of n^m must correspond to one of the M levels. Consider an example with four amplitude levels ($n = 4$) to represent a pulse. The 256 amplitude levels of the sampled signal in the previous example can now be represented by four pulses per sample ($m = 4$). This requires a bandwidth of $2 \times n$ or eight times the bandwidth of the analog system. Similarly, if a smaller value is selected for M, the bandwidth is again reduced. For all such schemes, there is a trade-off involved in terms of system noise and signal power. This ability to trade the bandwidth and SNR, back and forth via coding and signal processing, is a characteristic of all PCM systems.

1.7.4 Quantization Noise in PCM Systems

The process of quantization introduces noise at the transmitting terminal. This noise depends on the number of amplitude intervals chosen to represent a signal. For a signal quantized with a uniform interval, the peak signal to rms noise ratio is given by [2, 3]

$$\left(\frac{S}{N}\right)^* = 3M^2$$
$$= 3n^{2m}$$

Table 1.1 SNR versus relative bandwidth for binary transmission.

Number of quantizing levels	Binary digits for coding	Relative bandwidth	Peak SNR (dB)
8	3	6	22.8
16	4	8	28.8
32	5	10	34.8
64	6	12	40.8
128	7	14	46.8
256	8	16	52.8
512	9	18	58.8
1024	10	20	64.8

For the binary case, $n = 2$ and the SNR in dB is

$$\left(\frac{S}{N}\right) = 4.8 + 6m \text{ dB} \tag{1.64}$$

This relation provides the trade-offs between the SNR and the relative bandwidth, summarized in Table 1.1.

The frequency spectrum of the quantizing noise with a uniform interval is essentially flat over the range of interest. Quantization does not need to be uniform. In fact, few message signals possess uniform amplitude distribution, and most have a large dynamic range. To overcome such constraints, quantization, typically, has nonuniform spacing and is optimized to achieve a relatively uniform signal-to-distortion ratio over a wide dynamic range. Such nonuniform quantization is referred to as *companding*. Quantization noise can be reduced to any desired degree by choosing increasing finer quantization levels. However, the larger the number of quantizing steps, the greater the required bandwidth is. It is, therefore, desirable to choose as few steps as necessary to meet the objectives of the transmission. Needless to say, many subjective tests have been carried out to determine the acceptable number of levels for voice, video, and data signals. High-quality speech transmission is readily achieved with 128 levels or a 7 bit PCM. Good-quality television requires 9 or 10 bit PCM.

1.7.5 Error Rates in Binary Transmission

The deliberate quantization error or noise imparted to the PCM signal is a major source of signal impairment and originates only at the transmitting (or coding) end of the system. This noise can be made arbitrarily small at the expense of increased bandwidth. The other type of noise that is always present is the thermal noise generated by dissipative elements in the path of the signals and the noise generated by active devices. These noises are random and follow a Gaussian distribution function. They are added to the incoming group of pulses at the receiver. The noise density and distribution function of such noise are the same as those given in equations (1.19) and (1.20).

To detect the presence or absence of a pulse in a binary system, a minimum SNR is required on the digital line. If the pulse power is too low compared with the noise power, the detector will make occasional errors, indicating a pulse where there is none or vice versa. However, if the signal power is increased, the error can be rendered arbitrarily small. To determine the probability of error quantitatively, assume the amplitude of the pulse to be V_p when it is present and zero when it is absent, designated by 1 and 0, respectively. The composite sequence of the

binary symbols and the received noise is sampled once every binary interval and decision must be made as to whether a 1 or 0 is present. A simple and perhaps obvious way to decide is to say that if the voltage plus noise sample exceeds $V_p/2$, it is a 1 and if it is less than $V_p/2$, it is a zero. Error then occurs if, with a pulse present, the composite voltage sample is less than $V_p/2$ or, with a pulse absent, if the noise alone exceeds $V_p/2$. To calculate the probability of error, let us assume that a 0 is sent. The probability of error is then just the probability that noise will exceed $V_p/2$ volts and be mistaken for a 1. The probability of error is therefore the voltage that appears somewhere between $V_p/2$ and ∞. Assuming noise to be Gaussian with an rms value of σ, the probability of error is given by [3]

$$P_{e0} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{V_{p/2}}^{\infty} e^{-v^2/2\sigma^2}\, dv \tag{1.65}$$

In a similar manner, the probability of error when a pulse is sent and interpreted as zero:

$$P_{e1} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\infty}^{V_{p/2}} e^{-(v-V_p)^2/2\sigma^2}\, dv \tag{1.66}$$

The two types of errors are mutually exclusive and are equivalent. If it is further assumed that the two binary signals are equally likely, then the system probability P_e is the same as P_{e0} or P_{e1}: $P_e = P_{e0} = P_{e1}$.

The probability function P_e is well known and available in various mathematical tables. It is plotted in decibels as a function of V_p/σ in Figure 1.25. It should be noted that P_e depends solely on V_p/σ, the ratio of peak signal to rms noise ratio. It is interesting that P_e has a maximum value of 1/2. Thus, even if the signal is entirely lost in the noise, the receiver cannot be wrong by more

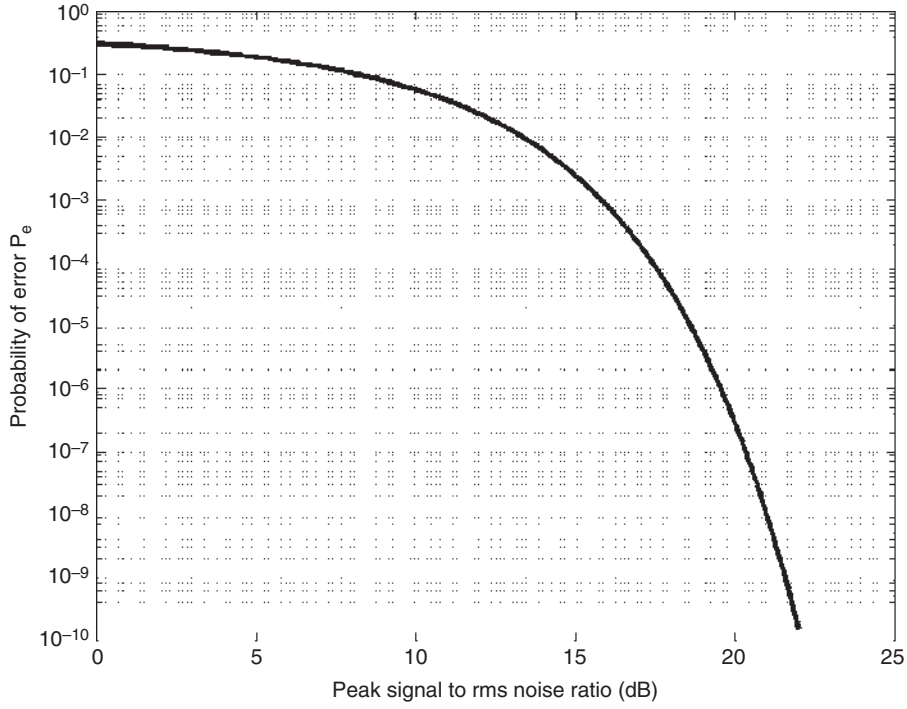


Figure 1.25 Probability of error versus peak signal to rms noise voltage ratio.

than half the time on the average. The probability curve indicates a sharp drop around the 16 dB level. Below this level, the error rate increases sharply and is called the *threshold effect*. For this reason, for the transmission of binary digits, the threshold level is chosen to be somewhere between 16 and 18 dB.

The implicit assumptions in this probability curve are (1) the statistics of the receive signal plus noise are Gaussian and (2) the transmission system is transparent, with no effect on the signal statistics or noise prior to detection. These assumptions allow the chosen decision level to be midway within a pulse amplitude.

1.7.6 Digital Modulation and Demodulation Schemes

Advances in data compression, digital modulation, and coding techniques coupled with the spectacular reduction in the costs of digital circuits are diverting more and more traffic to the digital domain. It is a matter of time before most, if not all, traffic is carried by using digital communication systems. A brief overview of digital modulation schemes, aspiring to achieve ever greater power and bandwidth efficiencies, approaching the Shannon limit, is outlined. The frequency spectrum of various modulation schemes has an impact on the desired filter characteristics, required to extract and process information in the communication channel.

Digital baseband signals can be modulated on a sinusoidal carrier by modulating one or more of its three basic parameters: amplitude, frequency, and phase. Accordingly, there are three basic modulation schemes: amplitude shift keying (ASK), frequency shift keying (FSK), and phase shift keying (PSK).

1.7.6.1 Amplitude Shift Keying (ASK)

This type of modulation is characterized by the amplitude of the carrier wave switching between zero (off state) and some predetermined amplitude level (on state). The amplitude-modulated ASK signal is therefore given by

$$M(t) = Af(t)\cos\omega_c t \tag{1.67}$$

where $f(t) = 1$ or 0, over intervals T seconds long. This is analogous to the amplitude modulation in analog systems. The Fourier transform of the ASK signal is then given by

$$F(\omega) = \frac{A}{2}[F(\omega - \omega_c) + F(\omega + \omega_c)] \tag{1.68}$$

The binary signal simply shifts the frequency spectrum to the carrier frequency f_c with the signal energy distributed between the upper and lower sidebands. The required transmission bandwidth is twice the baseband bandwidth. For a pulse of amplitude A and width T (the binary interval), the spectrum is given by

$$A\frac{T}{2}\left[\frac{\sin(\omega - \omega_c)T/2}{(\omega - \omega_c)T/2} + \frac{\sin(\omega + \omega_c)T/2}{(\omega + \omega_c)T/2}\right] \tag{1.69}$$

This is the well-known $\sin(x)/x$ response of a pulse with a finite width, as illustrated in Figure 1.26.

1.7.6.2 Frequency Shift Keying (FSK)

The frequency-modulated signal by a binary pulse is represented by

$$M(t) = A\cos\omega_1 t \quad \text{or} \quad M(t) = A\cos\omega_2 t \tag{1.70}$$

where $-T/2 \le t \le T/2$.

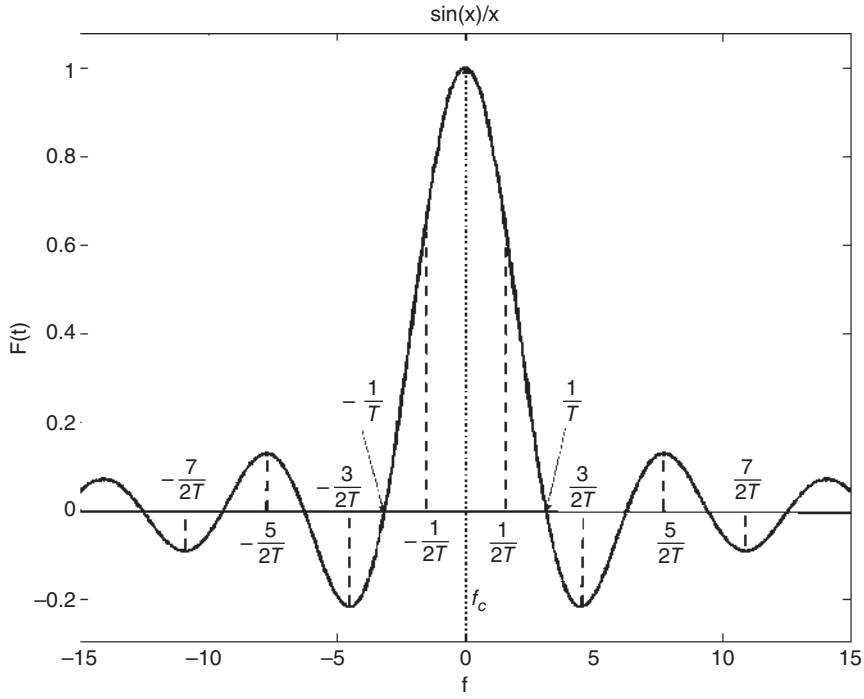


Figure 1.26 Spectrum of a periodic ASK signal.

In this scheme, one of the frequencies, say, f_1, can represent a one and the other frequency f_2, a zero. An alternative representation of FSK is, by letting $f_1 = f_c - \Delta f$ and $f_2 = f_c + \Delta f$:

$$M(t) = A\cos(\omega_c \pm \Delta\omega)t \tag{1.71}$$

The frequency deviates $\pm\Delta f$ about f_c, and Δf represents the frequency deviation. The frequency spectrum of the FSK is complex and has a form similar to that for analog FM.

1.7.6.3 Phase Shift Keying (PSK)

This form of modulation is characterized by a change of phase of the carrier frequency. Owing to the binary nature of the baseband signal, this simply implies a change of polarity. A PSK signal is represented in a form similar to ASK

$$M(t) = f(t)\cos\omega_c t, \quad \frac{-T}{2} < t < \frac{T}{2} \tag{1.72}$$

where $f(t) = \pm 1$. Here a 1 in the baseband binary stream corresponds to positive polarity and a 0 to negative polarity. The PSK signal has the same double-sideband characteristic as the ASK transmission. This result is similar to low index phase modulated analog systems. A comparison of the basic ASK, FSK, and PSK modulations is provided in Figure 1.27.

1.7.7 Advanced Modulation Schemes

As stated in Section 1.4, the two primary resources in a communication system are the signal power and the available transmission bandwidth. Advanced modulation schemes are geared to achieving higher efficiencies in either or both of these resources. Before proceeding further, let us examine the bandwidth and power efficiencies of a digital system.

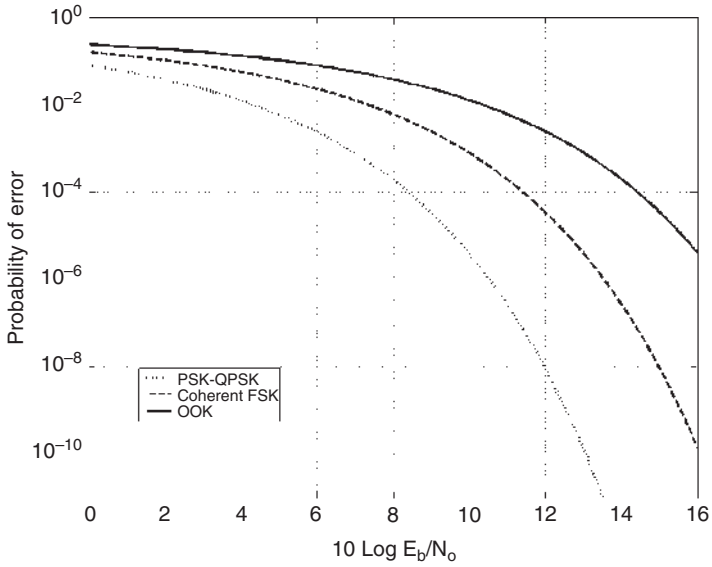


Figure 1.27 Comparison of ASK, FSK, and PSK modulation schemes. (*Source*: Schwartz 1980 [3]. Reproduced with the permission of McGraw-Hill Professional.)

1.7.7.1 Bandwidth Efficiency

As described in Section 1.3, Shannon's Theorem States

$$C = B\log_2\left(1 + \frac{S}{N}\right)$$

If the information rate R is equal to C, then

$$\frac{R}{B} = \log_2\left(1 + \frac{S}{N}\right) \tag{1.73}$$

defines the ultimate limit of the bandwidth efficiency, or the so-called Shannon limit. As described in Section 1.7.3, the quantized state of a signal can be represented by pulses of varying amplitudes or phases. Each state of a signal sample may be represented by [17]

$$M = 2^m$$
$$m = \log_2 M \tag{1.74}$$

Each state of M is referred to as a *symbol* and consists of m bits. Each symbol is transmitted as an electric voltage or current waveform. If the duration of the transmission of a symbol is T_s, then the data rate R is given by

$$R = \frac{m}{T_s} = \frac{\log_2 M}{T_s} \tag{1.75}$$

If T_b represents the duration of a bit ($= T_s/m$) and B is the allocated bandwidth, then the transmission bandwidth efficiency is expressed as

$$\frac{R}{B} = \frac{\log_2 M}{BT_s} = \frac{1}{BT_b} \tag{1.76}$$

The smaller the BT_b product, the higher is the bandwidth efficiency of the communication system. By assuming ideal Nyquist filtering, bandwidth B is simply given by $1/T_s$ and

$$\frac{R}{B} = \log_2 M \text{ bps/Hz} \tag{1.77}$$

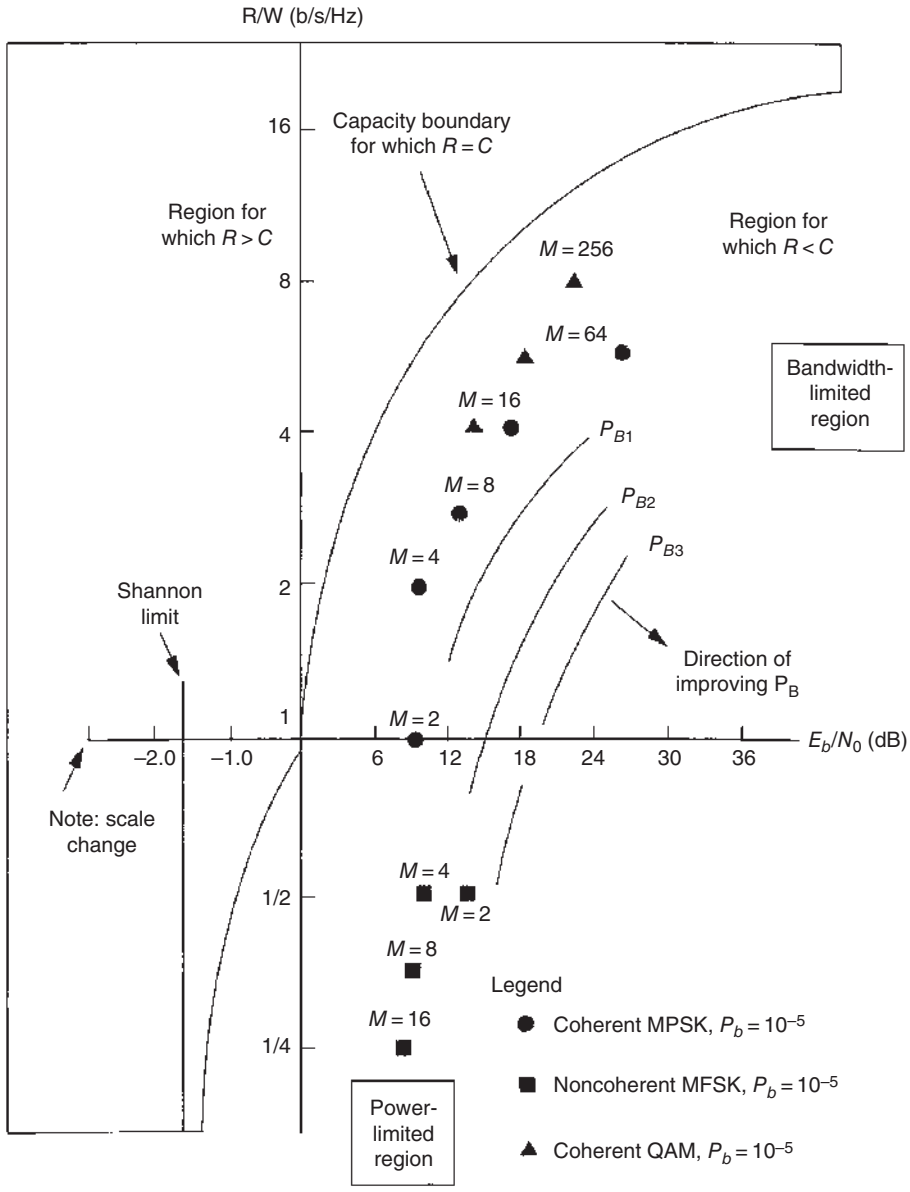


Figure 1.28 Bandwidth efficiency of digital modulation schemes. (*Source*: Xiong 1994 [18]. Reproduced with the permission of IEEE.)

where bps = bits per second. This represents the Shannon limit of bandwidth efficiency. As M increases, so does R/B. However, this comes at the cost of increased E_b/N_0. Figure 1.28 depicts the trade-off between the bandwidth and required E_b/N_0 for multiphase or M-ary (MPSK) modulation schemes. As expected, each modulation scheme leads to its own unique frequency spectrum. For example, Figure 1.29 traces the frequency spectrum for a binary (BPSK) and quadrature (QPSK), and offset QPSK modulation schemes. QPSK and offset QPSK are the most widely used modulation schemes in satellite communication systems.

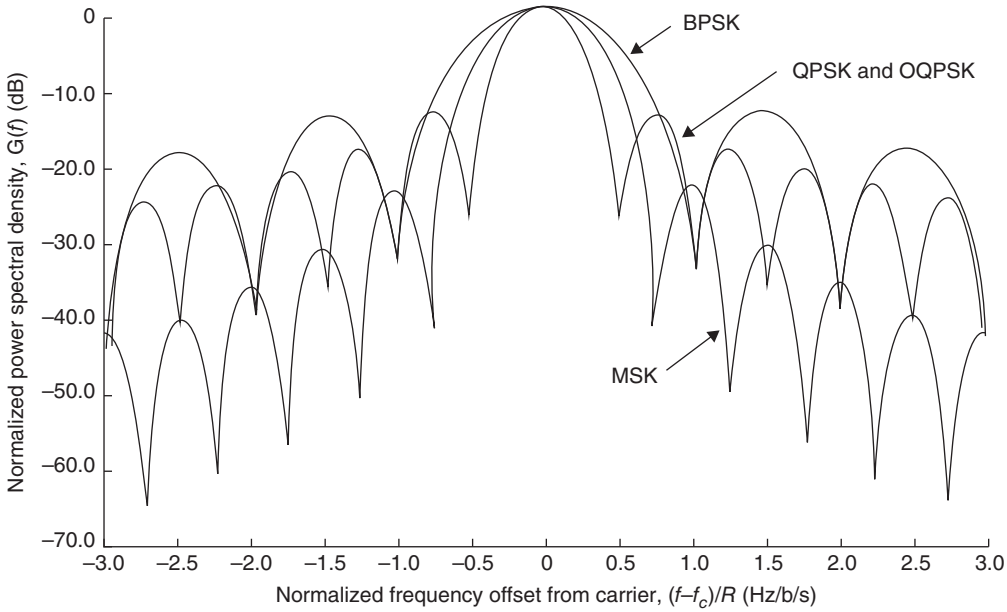


Figure 1.29 Normalized power spectral densities for PSK modulation schemes. (*Source*: Xiong 1994 [18]. Reproduced with the permission of IEEE.)

1.7.7.2 Power Efficiency

Power efficient modulation schemes are best suited for FSK modulation systems. In the binary FSK, the required bandwidth is twice the symbol rate and has a bandwidth efficiency of 0.5 bps/Hz. This is similar to the case of a narrowband analog FM that requires twice the bandwidth of the baseband signal. Again, analogous to the wideband analog FM, the power efficiency is improved by trading off the bandwidth. For MFSK scheme, the minimum bandwidth, required in accordance with the Nyquist criterion, is provided in [17]

$$B = \frac{M}{T_s} = MR_s \tag{1.78}$$

where $R_s(= 1/T_s)$ is the symbol rate.

By using M different orthogonal waveforms, each requiring a bandwidth of $1/T_s$, the bandwidth efficiency of incoherent orthogonal MFSK signals with Nyquist filtering is given by

$$\frac{R}{B} = \frac{\log_2 M}{M} \quad \text{bps/Hz} \tag{1.79}$$

Bandwidth–power trade-offs for MFSK modulation schemes are depicted in Figure 1.28, which shows how it is possible to achieve lower E_b/N_0 at the expense of increased bandwidth.

1.7.7.3 Bandwidth and Power Efficient Modulation Schemes

Table 1.2 describes a wide range of advanced digital modulation schemes [8]. They can be classified into two large categories: constant envelope and non-constant envelope. Generally, the constant envelope class is considered as the most suitable, where the effects of nonlinear amplification in HPAs are an important system consideration.

The PSK schemes (see Figure 1.29) have a constant envelope but discontinuous phase transitions from symbol to symbol. Classic PSK techniques include the BPSK and QPSK. More

Table 1.2 Advanced digital modulation schemes.

Abbreviation	Alternate abbreviation	Definition
ASK	—	Amplitude shift keying
FSK	—	Frequency shift keying (generic name)
BFSK	FSK	Binary frequency shift keying
MFSK	—	*M*-ary frequency shift keying
PSK	—	Phase shift keying (generic name)
BPSK	2PSK	Binary phase shift keying
DPSK	—	Differential BPSK
QPSK	4PSK	Quadrature phase shift keying
DPQSK		Differential QPSK (with differential demodulation)
DEQPSK		Differential QPSK (with coherent demodulation)
OQPSK	SQPSK	Offset QPSK, staggered QPSK
λ/4-QPSK	—	λ/4-quadrature phase shift keying
λ/4-DQPSK	—	λ/4-differential QPSK
CTPSK	—	λ/4-controlled transition PSK
MPSK	—	*M*-ary phase shift keying
CPM	—	Continuous phase modulation
SHPM	—	Single-*h* (modulation index) phase modulation
MHPM	—	Multi-*h* phase modulation
LREC	—	Rectangular pulse of length *L*
CPFSK	—	Continuous phase frequency shift keying
MSK	FFSK	Minimum shift keying, fast frequency shift keying
DMSK	—	Differential MSK
MSK	—	Gaussian MSK
SMSK	—	Serial MSK
TFM	—	Timed frequency shift keying
CORPSK	—	Correlative PSK
QAM	—	Quadrature amplitude modulation
SQAM	—	Superposed QAM
Q^2PSK	—	Quadrature phase shift keying
QPSK	—	Differential Q^2PSK
IJF OQPSK	—	Intersymbol-interference jitter-free OQPSK
SQORC		Staggered quadrature-overlapped raised-cosine modulation

Source: Reproduced from [8] with permission of John Wiley and Sons.

generally, modulation schemes with *M*-ary PSK (MPSK) and *M*-ary FSK (MFSK) signals can be used with a variety of trade-offs between power and bandwidth efficiencies.

The continuous phase modulation (CPM) schemes have not only a constant envelope but also continuous phase transitions from symbol to symbol. They have less side lobe energy in their spectra in comparison with the PSK schemes. A great variety of CPM schemes can be obtained by varying the modulation index and pulse frequency [8].

1.7.8 Quality of Service and *S/N* Ratio

The quality of a wireless communication link depends not only on its design but also on random effects of the propagation environment, such as rain attenuation, tropospheric and ionospheric scintillation, Faraday rotation, Doppler effect, and antenna pointing errors. As a consequence, transmission performance is defined probabilistically, in terms of the specific quality of the signal over a percentage of time. This has given rise to a variety of standards, some agreed on and others still open for discussion [8]. Conventionally, the quality of the signal is expressed in terms of SNR in the case of analog transmission and bit error rate (BER) in the case of digital transmission. For most applications, the specification calls for SNR to remain above a certain value or BER to remain below a certain value over 99% and 99.9% of the time, averaged over a year. Over these time periods, typical SNR for analog TV is specified to be 53 and 45 dB, respectively. At the present time, no consensus has emerged on the standards for digital transmission, although more and more traffic is moving toward it. With the widespread applications and advances in coding technology, it is becoming possible to have virtually error-free transmission for digital signals. For digital TV, satellite systems in the 14/11 GHz band are targeting a performance objective of one uncorrected error per transmission hour, specifically, BER $\leq 10^{-10}$ or 10^{-11}, depending on user bit rate [8].

It should be noted that SNR depends on two parameters, CNR and signal modulation. CNR is a measure of the efficiency of radio transmission at RF, whereas modulation provides the conversion of CNR into SNR. The variety of modulation and coding techniques provide the trade-offs between CNR and SNR. Systems are always designed to ensure that CNR levels are significantly larger than the threshold value of FM of digital demodulators.

PART III Impact of System Design on the Requirements of Filter Networks

1.8 Communication Channels in a Satellite System

Communication satellites (Figure 1.30) are radio relay stations in space. They serve much the same purpose as the microwave towers one sees spread over populated landmass areas. Satellite communication has evolved since the 1970s and represents a mature niche in the field of telecommunications [5–8, 19].

The satellites receive radio signals transmitted from the ground, amplify them, translate them in frequency, and retransmit them back to the ground. Since the satellites are at high altitude, they can see all the microwave transmitters and receivers (earth stations) on almost one-third of the earth. Thus, they can connect any pair of stations or provide point-to-multipoint services, such as television. The coverage can be extended to any part of the earth via intersatellite links or interconnection with long-distance fiber networks, giving satellite systems an inherent advantage of being insensitive to distance. Satellites are unique in their capability to provide global, seamless, and ubiquitous coverage, including mobile services via handheld units. The frequency plans for commercial satellite systems are listed in Tables 1.3 and 1.4.

It should be noted that frequency allocations are periodically addressed and revised by domestic and international regulatory agencies in order to accommodate new services. To determine the characteristics of communication channels, let us examine the typical block

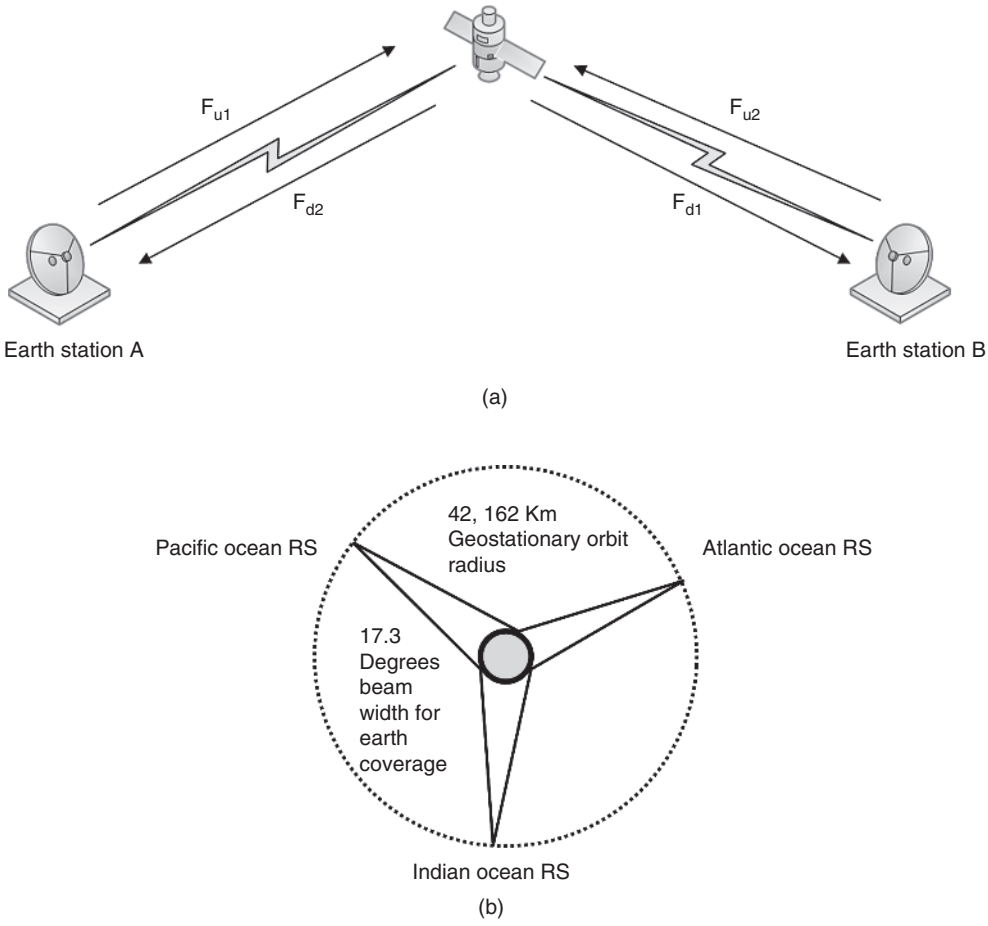


Figure 1.30 Satellite communications: (a) typical satellite link; (b) worldwide coverage using a three-geostationary satellite system.

Table 1.3 Frequency allocations for satellite systems.

Approximate frequency range (GHz)	Letter	Typical usage
1.5–1.6	L	Mobile satellite service (MSS)
2.0–2.7	S	Broadcasting satellite service (BSS)
3.7–7.25	C	Fixed satellite service (FSS)
7.25–8.4	X	Government satellites
10.7–18	Ku	Fixed satellite service (FSS)
18–31	Ka	Fixed satellite service (FSS)
44	Q	Government satellites

Table 1.4 Intersatellite frequency allocations.

Allocation frequency (GHz)	Total bandwidth (MHz)	Satellite services
22.55–23.55	1000	Fixed, mobile, broadcasting
59–64	5000	Fixed, mobile, radio location
126–34	8000	Fixed, mobile, radio location

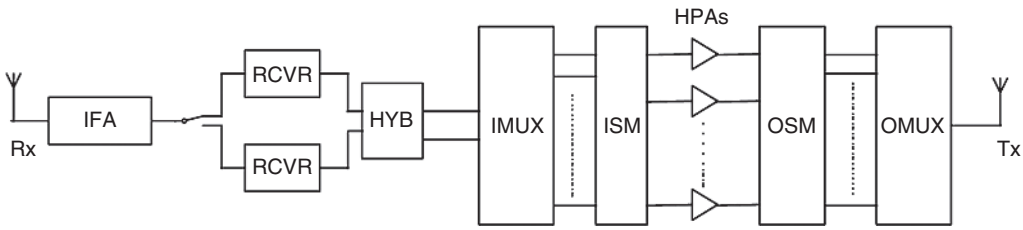


Figure 1.31 Communication subsystem block diagram.

diagram of a communication subsystem of a satellite repeater, as depicted in Figure 1.31. Most commercial satellite systems employ dual orthogonal polarization (linear or circular) providing twofold increase in the available bandwidth. Advance satellite systems employ multiple beams, which allow further reuse of the available bandwidth. However, such advance architectures come at the cost of higher complexity for the spacecraft.

Regardless of the satellite architectures, the block diagram of transponders on a given beam and polarization is essentially the same as the one in Figure 1.31. A receive antenna (Rx) is connected to a wideband filter (IFA - Input Filter Assembly), followed by a low-noise receiver (RCVR). After that, the signal is channelized into its various transponders via an input multiplexing network (IMUX). The allocated frequency band in a satellite system is divided into a number of RF channels, often referred to as *transponders*. Typical channel separation and bandwidths in Ku- and C-band satellite systems are listed in Table 1.5.

Each RF channel is amplified separately and recombined by an output multiplexer (OMUX) network into a composite wideband signal that feeds into the transmit antennas (Tx). Input and output switch matrices (ISM and OSM) are there to provide onboard reconfiguration of traffic flow from one transponder to another, from one beam to another, or various combinations thereof. Often, the ISM and OSM consist of mechanical switches. For ISM, loss is not a

Table 1.5 Typical channel separations and bandwidths in C and Ku bands for fixed satellite services.

Channel separation (MHz)	Desired usable bandwidth (MHz)
27	24
40	36
61	54
80	72

constraint, and, as a consequence, solid-state switches can be utilized to conserve mass and volume. For the OSM, loss is critical, and there is no substitute for mechanical switches by virtue of their very low loss. From the standpoint of the filtering requirements, the block diagram can be separated into three distinct groups: the front end, the channelizer section, and the high-power output circuits. Such a breakdown is typical of most communication repeaters. We now deal with each of the three subgroups separately.

1.8.1 Receive Section

The receive section consists of a wideband input receive filter, the LNA, frequency downconverter, and driver amplifier, as shown in Figure 1.32. For high reliability, satellites invariably employ a redundant receiver that requires a switch prior to the receiver. At the receive antenna, the signal strength is at its lowest level, and therefore, it is imperative to minimize the energy loss prior to the LNA. The wideband receive filter is required to ensure that only the signals in the *allocated frequency spectrum*, *typically comprising* 500 MHz bandwidth, are fed to the LNA and all other signals outside this range are attenuated. For the filters and transmission lines prior to the LNA, the designs are dictated by a need for low insertion loss in the passband. Typical requirement calls for an insertion loss of no more than a few tenths of a decibel.

The frequency downconverter consists of a mixer–local oscillator (LO) assembly. The separation between the transmit and receive frequencies is required to minimize the interference between them. A driver amplifier (DAMP) is used to attain the required power levels prior to channelization and final amplification of the signals. It also allows LNA to operate at low enough power level consistent with the lowest possible noise figure. The hybrid is a type of 3 dB power divider that provides two separate paths for the channelization of the transponders.

1.8.2 The Channelizer Section

A detailed block diagram of the channelizer or input multiplexer (IMUX) section is shown in Figure 1.33.

Once the signal has been amplified by the LNA, the loss in the subsequent equipment is no longer critical. This is due to the fact that the receive system noise temperature is reduced by the gain of the LNA, and the post-LNA losses have less impact (as described in Section 1.5.5). Instead, the design driver is the efficient channelization of the composite signal into its various RF channels or transponders. The criteria for channelization is that the signals incur a minimum

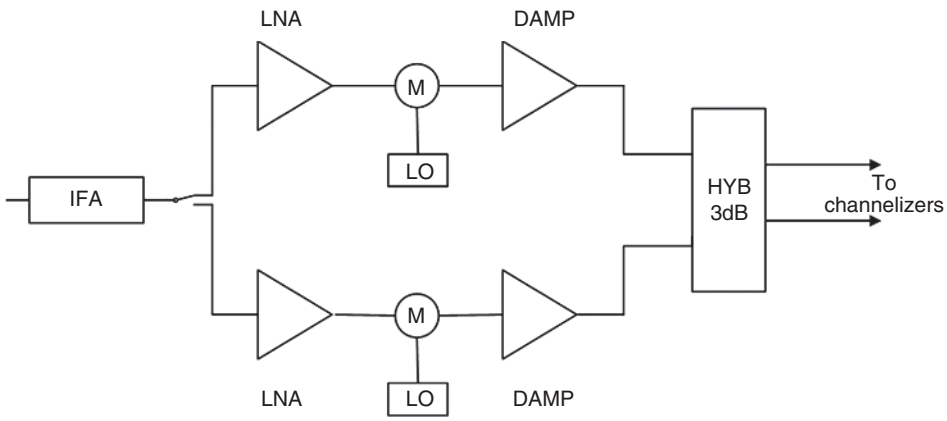


Figure 1.32 Satellite receiver block diagram.

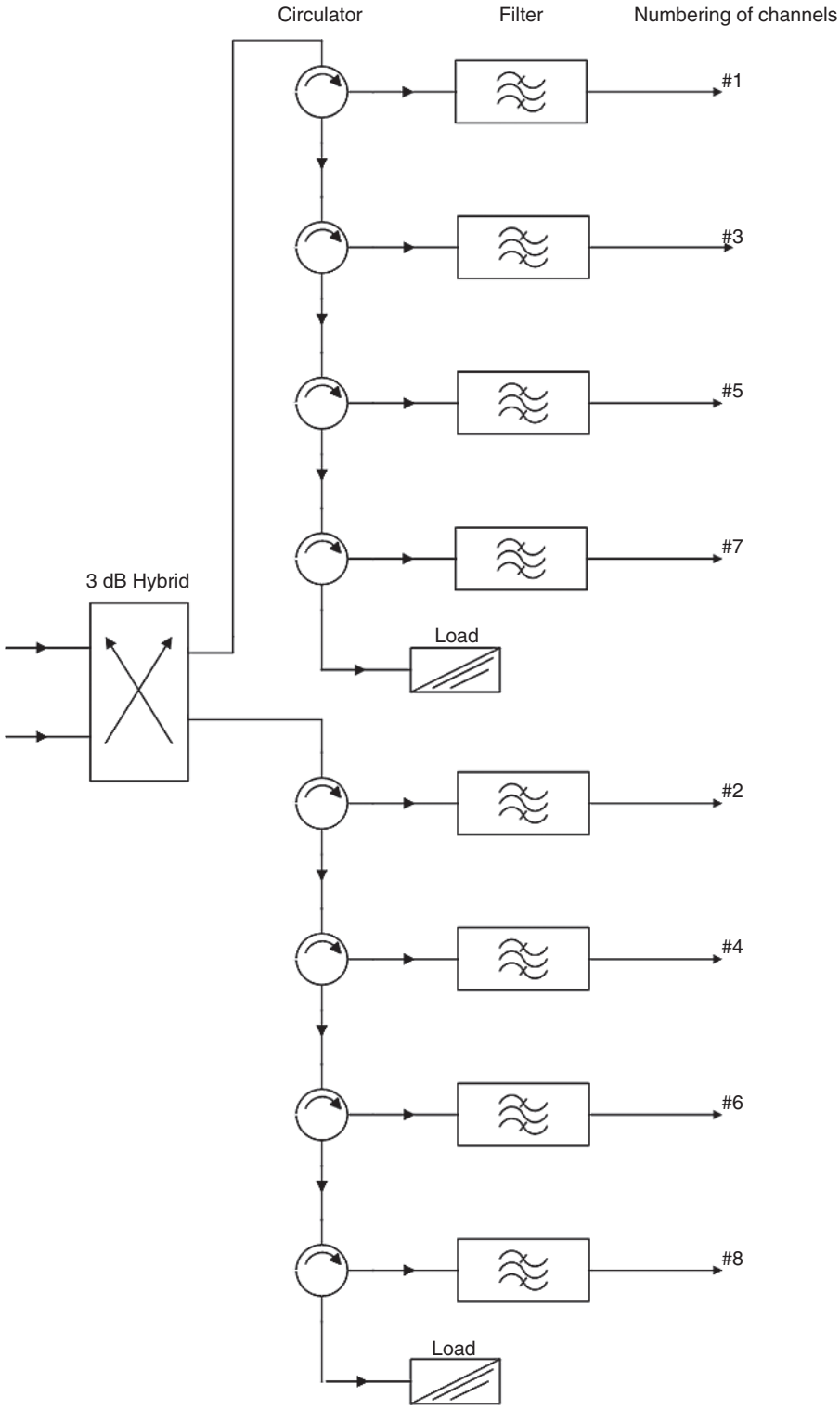


Figure 1.33 Input multiplexing network in a satellite.

of distortion in the passband and, at the same time, provide enough isolation (typically >20 dB) to control interference from other channels. This implies filter characteristics that emulate close to an ideal response as described in Figure 1.17a and b. Any departure from the ideal response means signals would incur distortion, especially at band edges where the amplitude and group delay tends to depart from its desired flat response. A guard band between channels (typically comprising 10% of channel bandwidth) is allocated to allow margin for the nonideal characteristics of practical filters. In essence, channelizing filters largely determine the effective usable bandwidth of each RF channel with a minimum of guard band. For communication channels, typical bandwidth requirements range from 0.3% to 2%. Such narrow band filters incur relatively large transmission deviations and often require a degree of phase and group delay equalization at the expense of additional hardware. As an example of a 6/4 GHz satellite system, the typical specification for a channelizing filter is described in Table 1.6.

The defining feature of the channelizing filter networks is the very stringent out-of-band rejection requirement. It necessitates the use of filters with transmission zeros just outside the passband to achieve a steep isolation response. Although the insertion loss at the band center is not critical, the amplitude variation across the passband is and can be minimized only by employing filters with high unloaded Q. An interesting feature of the channelization scheme is the use of channel dropping ferrite circulators and 3 dB power dividers with coaxial cables to ease the layout of the equipment. Circulators are three-port nonreciprocal devices, constructed from ferrite materials, biased with external magnets, to achieve the desired property of the unidirectional flow of energy. Such devices incur a few tenths of a decibel loss in coaxial or planar structures, and less than a tenth of a decibel loss in the waveguide realization. The reverse isolation is 30–40 dB. Circulators are inherently wideband devices, whose bandwidths range from 10% to 20%. It is possible to optimize the performance over narrower bandwidths. Circulators are widely used in radar and communication systems. As shown in Figure 1.33, use of 3 dB hybrid and channel dropping circulators enable design simplicity and flexibility for the layout of the multiplexing networks. However, this design simplicity comes at the expense of increased loss, which, as described earlier, is of little consequence for this portion of the network. Consequently, the narrowband channelizing filters need to be high-Q structures, whereas the wideband ancillary equipment, namely, circulators, isolators, and hybrid, use the conventional lower-Q but more compact coax designs.

Table 1.6 Typical specifications for channelizing filters in a 6/4 GHz satellite system.

Frequency band	3.7–4.2 GHz
Number of RF channels	12
Channel separation	40 MHz
Passband bandwidth	36 MHz
Rejection over narrowband	10–15 dB at band edge of adjacent copolarized channels, rising to 30–40 dB within 10–15% of channel bandwidth (2–3 MHz) beyond band edge over frequency band 3.7–4.2 GHz
Rejection over wideband	>40–45 dB over receive 5.925–6.425 GHz band
Insertion loss in passband	Not critical
Passband loss variation	<1 dB
Passband relative group delay	<1–2 ns over middle 70%, rising to 20–30 ns toward band edges
Operating temperature range	0–50°C

Narrowband filters incur relatively high group delay variation across its passband. Group delay varies inversely with the absolute bandwidth and is further constrained by the steepness of the required amplitude response of the filter. Group delay can be equalized using all-pass external equalizer network or by using higher-order self-equalized linear phase filters. Either way, group delay equalization increases the design complexity and cost. Most satellite systems employ some degree of equalization. An added benefit of group delay equalization is that it also reduces the amplitude variation across the passband at the expense of a small increase in mid-band loss. Trade-offs for the channelizing filter are critical in establishing the usable passband bandwidth and hence the efficiency in the utilization of the frequency spectrum.

1.8.3 High-Power Amplifiers (HPAs)

HPAs are required to raise the power levels of RF signals prior to their transmission back to earth. System level trade-offs determine the gain and maximum RF power of HPAs to ensure the desired communication capacity for a transponder. The traveling wave tube amplifier (TWTA) is the dominant power amplifier for communications satellites, although many satellites use solid-state power amplifiers (SSPA) as well. Over the years, the available power, reliability, and efficiency of TWTAs have significantly improved, allowing its continued dominance of the satellite market. However, SSPAs tend to exhibit better nonlinear characteristics for moderate power levels for C- and Ku-band satellite systems. As always, system-level trade-offs are required to select the appropriate HPA for a given system. HPAs are power-hungry, requiring its operation at a high efficiency. HPAs are inherently nonlinear devices and their characteristics involve trade-off between efficiency, output power level, and the amount of nonlinearity. Typical characteristics of a TWTA are depicted in Figure 1.34.

To obtain maximum RF power, the amplifier must be operated under the condition of saturation. However, at this power level, the amplifier is highly nonlinear and is suitable only for amplification of a single carrier. For multicarrier operation, it is essential to operate the TWTA in a backoff mode to keep the carrier-to-IM (C/I) ratio to an acceptable level. As described in Section 1.5.6, as the number of carriers in an RF channel increases, the number of IM products grows rapidly. These products fall inside as well as outside the RF channel. It is usual to specify the IM performance by the intercept point defined as the theoretical output power $\overline{IP}$, where

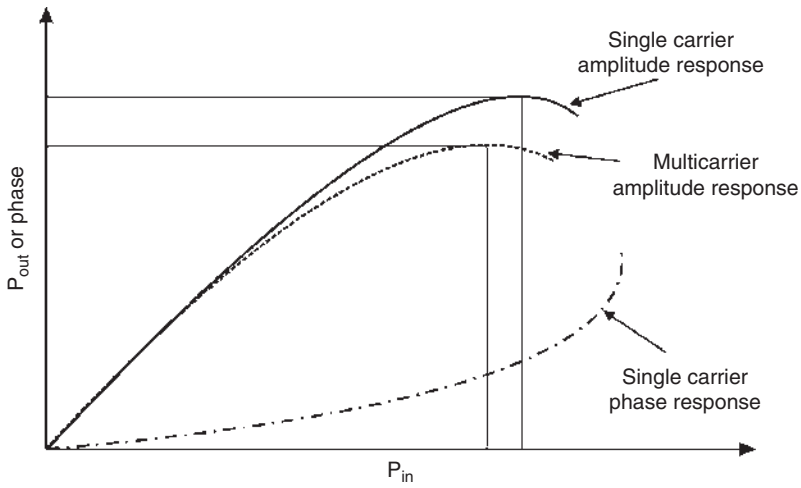


Figure 1.34 Typical characteristics of a high-power amplifier.

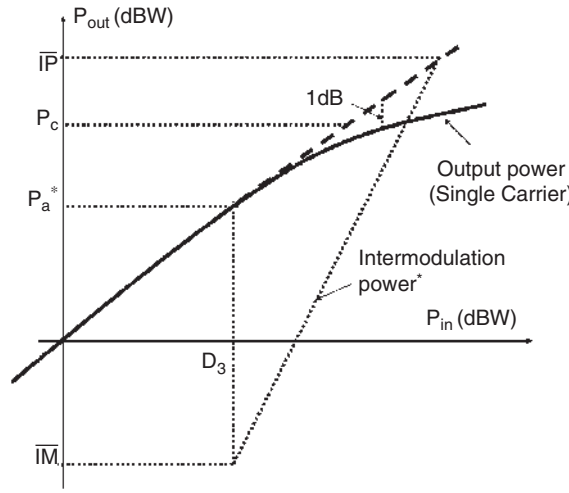


Figure 1.35 Third-order intermodulation (IM) intercept point in amplifiers. (*Source*: Reproduced with the permission of John Wiley and Sons.)

the extrapolated linear single carrier output power is equal to the extrapolated two-carrier IM power, as represented in Figure 1.35. The carrier to third order ($2f_1 \pm f_2$ or $2f_2 \pm f_1$) intermodulation (IM_3) is then given by [8, 19]

$$\frac{C}{IM_3} = 2(\overline{IP} - P_0) \text{ dB}$$

where P_0 is the single-carrier output power level. The value of the intercept point is specified by the tube manufacturer.

This relationship shows that carrier to third-order IM decreases 2 dB for every dB backoff of the output power of the HPA. Also, the greater the intercept point, the larger the value of *C*/*IM*. The intercept point is a characteristic of the amplifier and provides a measure of its linearity. There are good reasons for specifying HPA performance with respect to two carriers. It represents the minimum number of carriers to generate an IM product. In addition, third-order products have the highest power level and are critical in determining *C*/*IM* performance. The level of odd-order IM products tends to decrease between 10 and 20 dB with increasing order. IM products that fall inside the RF channel can be controlled only by the operating power level and linearity of the HPAs. There exists a trade-off between the efficiency and linearity of HPAs. Highly linear HPAs tend to have lower efficiency and vice versa.

The typical way to control IM is to operate HPA in the linear region of the power curve, which is usually 2–3 dB lower than its saturated power level. For two-carrier operation, HPAs are backed off by 2–3 dB below saturation power level. For multicarrier operation, the backoff can be as much as 10–12 dB to achieve acceptable performance. This represents a significant penalty in the output power to ensure acceptable IM performance. Obviously, linearity of amplifiers is a major issue. Often a "linearizer" is employed to improve the linearity of amplifiers, incurring extra hardware and cost. It should be noted that the prime reason for operating HPAs in a backoff mode for multicarrier operation is to control those IM products that fall inside the RF channel; those outside the RF channel are of no consequence, unless they lie within the transmit or receive frequency bands. IM products falling in the transmit band are attenuated by the high-power output filter or multiplexing networks that follow HPAs. IM products pertaining to the receive band are only possible if we consider products of much higher order. This is due to the fact that we are dealing with carriers in a single channel, not the whole transmit band. Such higher order products are not normally generated in the HPAs, even if they did, the power of such products would be negligible. However, high-power multiplexer and antenna equipment

handling all the channels can generate IM products, referred to as *passive inter-modulation* products or PIM, of lower order that can fall in the receive band. This topic is dealt with in a subsequent section on PIM. The other frequency band of significance corresponds to the second and third harmonics of carriers generated by HPAs. These harmonics, if not suppressed, can interfere with ground-based systems as well as with military- and science-based space systems. The harmonic suppression must be achieved prior to the transmit antenna. Harmonic suppression is part of the output multiplexing subsystem.

1.8.4 Transmitter Section Architecture

The transmit section of the satellite combines the outputs of various high-power channel amplifiers in an output multiplexing network for transmission via a common antenna. Once the signals have been amplified by the final amplifier, conservation of power becomes critical. Consequently, the challenge is how to achieve the lowest loss for each RF channel, preserve the useable bandwidth, and, at the same time, keep the design of multiplexer and antenna subsystems simple. This trade-off has given rise to two alternative multiplexing schemes for satellite systems.

Early satellite systems employed the noncontiguous multiplexing scheme. In such an architecture, alternate channels are combined in the so-called noncontiguous multiplexer, as depicted in Figure 1.36a. The output power of each noncontiguous multiplexer is then combined in a 3 dB hybrid as shown in Figure 1.36b. Power from each input port of the hybrid is divided equally between the two output ports with a phase difference of 90°. Thus, each output port of the hybrid contains all the RF channels at half the power. Such a scheme requires an antenna feed network with two input ports and is called *a dual-mode feed network*. This architecture was employed extensively for satellites in the 1970s and 1980s. Its advantage is the design simplicity of the multiplexer network. The disadvantage is that the architecture requires a more complex beam forming network, incurring a penalty in achievable antenna gain resulting in some loss of EIRP. The alternative is to combine the powers of all the transponders on a given polarization in a single device, the contiguous band multiplexer as shown in Figure 1.36c. Its main advantage is that it allows a simpler beamforming network and it is easier to optimize the antenna gain. Furthermore, owing to its inherent sharper amplitude characteristics, it reduces the effect of multipath in the satellite, resulting in improved RF channel characteristics [16]. Its disadvantage is that the design of such a multiplexer is more complex and it incurs slightly higher loss and group delay variation across the passband of RF channels. Over the years, technology advances have overcome the disadvantage of design complexity. Higher antenna gain achievable with this scheme more than compensates for the disadvantage of slightly higher loss in the multiplexer. As a result, most modern satellite systems employ contiguous multiplexing scheme since it yields better performance in terms of the overall transponder channel characteristics and EIRP. For some applications, especially for systems with narrower-band transponders (~20 MHz), a noncontiguous scheme offers a better design. Comparison of satellite architectures employing alternate multiplexing schemes is described in more depth in Ref. [16].

1.8.4.1 Output Multiplexer

The output multiplexer (OMUX) performs the function that is reverse to that of channelizing section. It combines the power of the individual RF channels to form a single composite signal for transmission back to earth via a transmit antenna. OMUX consists of a number of bandpass filters whose outputs are connected to a common manifold. Each filter corresponds to a particular transponder channel and is optimized to accept the amplified signal within the passband of that channel and reject frequencies corresponding to other transponders. In addition, the

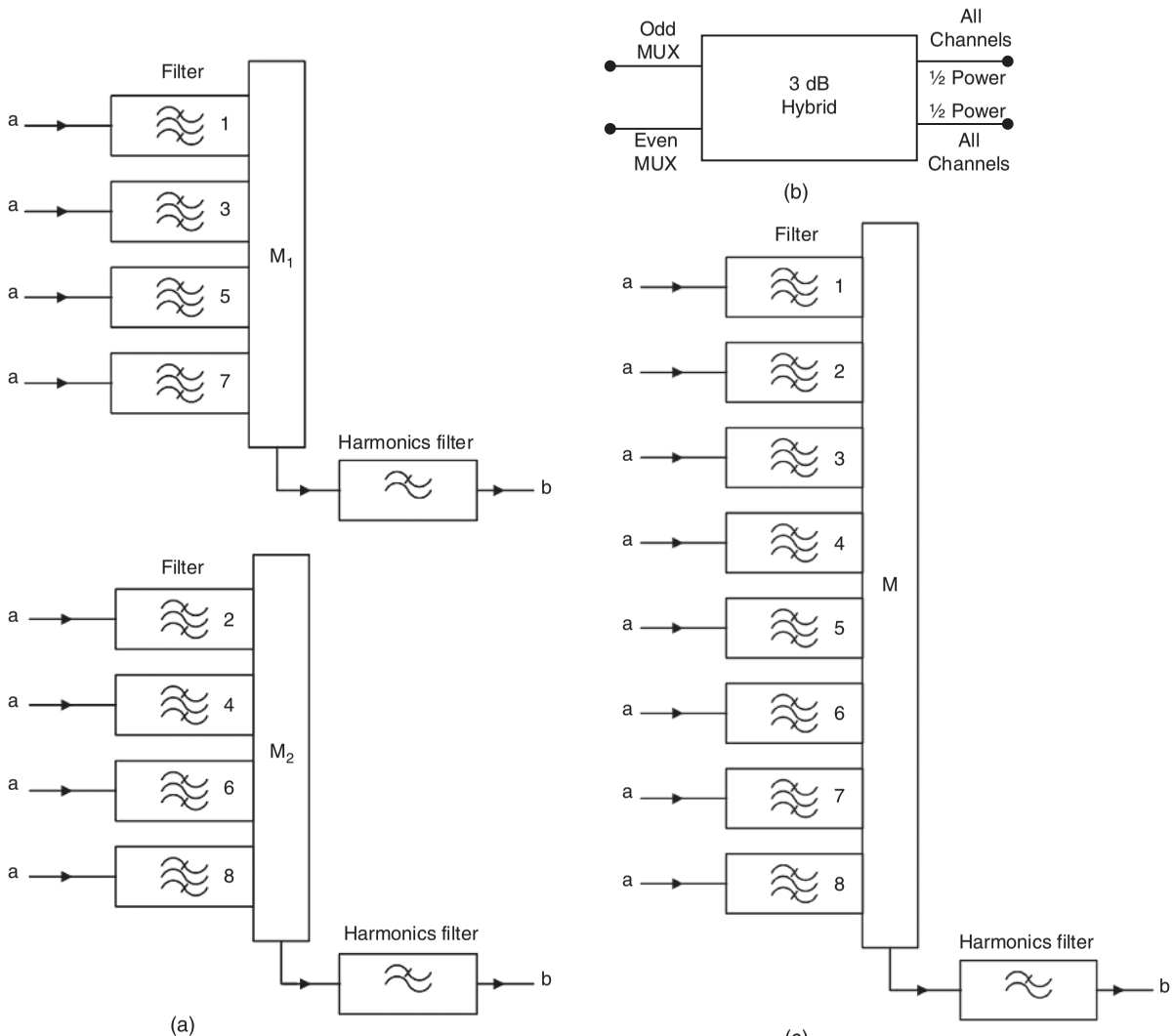


Figure 1.36 Alternative multiplexing schemes in a satellite system: (a) noncontiguous multiplexing scheme; (b) hybrid combiner; (c) contiguous multiplexing scheme.

Table 1.7 Typical specification of a channel in the high-power combining network in a 6/4 GHz satellite system.

Frequency band	3.7–4.2 GHz
Number of RF channels	12
Channel separation	40 MHz
Passband bandwidth	36 MHz
Passband insertion loss	Minimum (typically < 0.2–0.3 dB)
Rejection over narrowband[a)]	Shaped isolation response > 5–10 dB at band edges of adjacent copolarized channels, rising to 20–30 dB within 15–20% of channel bandwidth (3–4 MHz) beyond the band edge and over the frequency band 3.7–4.2 GHz
Rejection over wideband	>30–35 dB over the receive 5.925–6.425 GHz band
Passband relative group delay	<1–2 ns over middle 70%, rising to 10–20 ns toward band edges
Power handling	10–100 W per channel
Operating temperature range	0–50°C

a) The isolation response shape is dictated by the filter and multiplexer technology and the requirements of low-loss and high-power handling capacity.

performance of channel combining filters and overall multiplexer design is further optimized to achieve low loss and provide rejection over the receive band. TWTAs produce not only the desired amplified signals but also IM products and harmonics that must be suppressed. The OMUX provides part of that suppression. The typical requirements imposed for a channel in the OMUX of a 6/4 GHz satellite system are described in Table 1.7.

1.8.4.2 Harmonic Rejection Filters

The purpose of such filters is to provide a high rejection at the second and third harmonic while incurring a minimum of loss in the passband of the transponder. This is accomplished by a low-loss low-pass filter design. The power handling capability of such filters is a critical requirement. Two alternative designs are feasible to meet the power constraints as shown in Figure 1.37. One way is to use a single-harmonic filter connected after the multiplexer as shown in Figure 1.37a. The advantage is that it uses a single filter, incurring a lower mass and volume. However, it must be capable of handling the combined power of all the RF channels on the multiplexer, representing a major disadvantage. The alternative is to use a harmonic filter on a channel by channel basis, as depicted in Figure 1.37b. In this scheme, the harmonic filter is required to handle the power of a single channel and not the combined power of all the channels. Its disadvantage is the increase in the number of harmonic filters required.

Early satellite systems employed a single harmonic filter in OMUX. As the power of satellites has increased, there seems to be a shift toward employing harmonic filters for each individual channel. The particular choice is dictated by system level trade-offs.

1.8.4.3 PIM Requirements for High-Power Output Circuits

All active devices are inherently nonlinear and therefore generate IM products. What is not well understood is that all devices, including passive components such as filters, multiplexers, and antennas, have a degree of nonlinearity and are thus capable of generating IM products [20]. Such IM is known as passive inter-modulation or PIM. It arises owing to imperfections in the structure of materials. For most applications, the level of PIM is sufficiently low, and hence not an issue in system design. It can be an issue if a common antenna is used for the transmission and the reception of signals in a communication system. PIM is also problematic

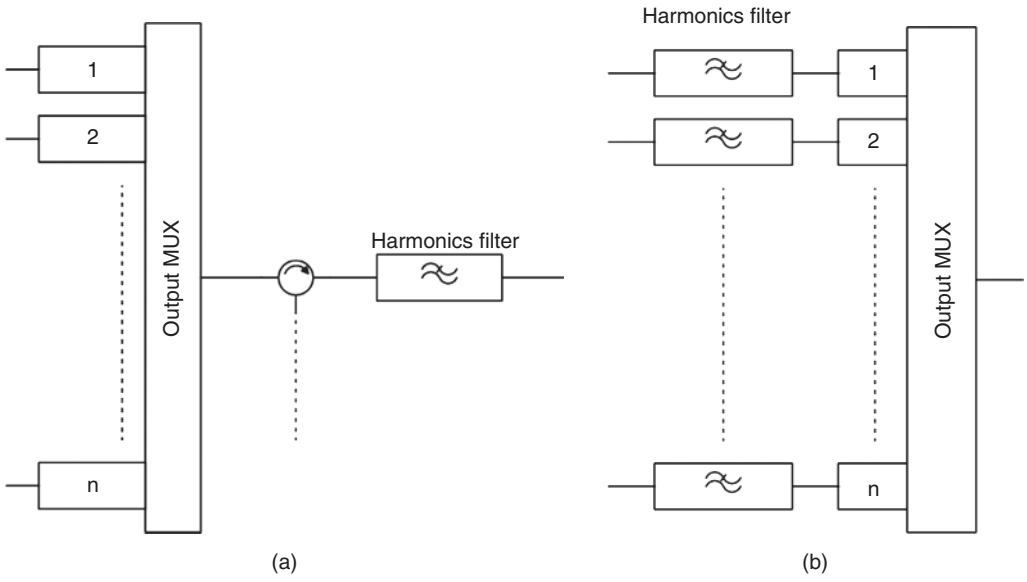


Figure 1.37 High-power output multiplexing network in a satellite, using (a) a single-harmonic filter and (b) individual channel harmonic filters.

if the high-power transmitter and low-level receivers are in close proximity, allowing coupling between the two. For example, the difference in the power levels between the transmit and receive section in a satellite repeater can be as high as 130–140 dB. This implies that the level of PIM should be between 160 and 170 dB or more below the power level of the transmitter to ensure that there is no interference with the low-level received signals. It has implications for the system design, as well as for the quality of materials and work standards for high-power equipment. One important impact on the system design pertains to the allocation of the transmit and receive frequency bands. As described in Section 1.8.3, the third-order IM is the dominant product. Consequently, the separation between the transmit and receive frequency bands should be such so as to avoid at least the third-order IM from falling into the receive band, preferably the fifth-order IM as well. If F_1 and F_2 represent the lower and upper band edge of the transmit frequency band, then

$$\text{Frequency location of IM products} = |m\ F_2 \pm nF_1|$$

$$\text{IM order} = m + n, m \text{ and } n \text{ are integers}$$

The frequency spectrum of IM products generated by two carriers is illustrated in Figure 1.38.

It is evident that even-order IM products are of no consequence since they lie well outside the bands of interest. Also, the odd-order IM products that lie below F_1 are of no consequence since the receive frequency band is invariably chosen to be higher than the transmit band. The reason for this choice is the higher efficiency of equipment at the lower transmit frequency band. The IM products that are of interest are the ones that are above F_2. The frequency locations of the upper odd-order IM products are given by $2F_2 - F_1$ for the third order, $3F_2 - 2F_1$ for the fifth order, and so on. As an example, transmit and receive bands for a 6/4 GHz satellite system are 3.7–4.2 and 5.925–6.425 GHz, respectively. In this case, the minimum odd-order IM product that falls in the receive band is the ninth-order product. This represents a relatively safe separation between the transmit and receive frequency bands. Some of the earlier military satellite systems had much closer separation that allowed third-order PIM to fall in the receive

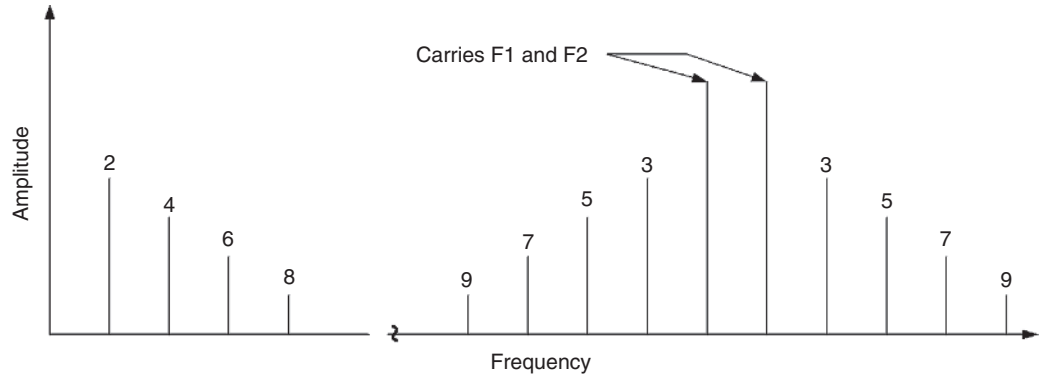


Figure 1.38 Frequency spectrum of IM products generated by two carriers.

band and that proved to be very problematic. It was only then that the system designers were made aware of the PIM phenomenon.

High-power output circuits are often subjected to a specification on PIM. It depends on the antenna design for the communication system. If a common antenna is used for transmission and reception, then any PIM products generated in the receive band can couple via the antenna feed networks into the receive section and interfere with the received signals. This can also occur for systems with independent transmit and receive antennas located in close proximity. A most conservative specification of PIM is to assume antenna Tx–Rx isolation to be zero and the PIM level to be equivalent to that of third-order IM, regardless of the order of the PIM that is permissible to fall into the receive band. Such a specification can put excessive constraints on hardware designers, and seldom can be met or demonstrated with a high degree of confidence. In practical systems, the Tx–Rx isolation is of the order of 20–30 dB for systems with a common antenna. It is much higher (60–100 dB) if separate transmit and receive antennas are used. If the permissible PIM falling into the receive band is of fifth order, its value is typically 10–20 dB lower than that of third-order PIM. Higher-order PIMs go down in level by 10–20 dB with each increasing order. This has been demonstrated by experiments for a number of satellite systems [21]. A typical PIM specification for the 6/4 GHz satellite system with a common antenna is −120 dBm for the third-order IM product.

1.9 RF Filters in Cellular Systems

Cellular radio evolved from mobile communication systems. The earliest application of mobile radio communications was to ocean-going vessels. This was followed by radio services to aircraft, and subsequently, on land for public safety services such as police, fire departments, and hospital ambulances. However, it was not until the early 1990s that mobile communications took hold as a means for personal communications, extending the reach of fixed telephony networks to people just about anywhere: in their backyards, in automobiles, urban centers, or remote regions. It is indeed possible to remain connected anywhere on the planet with access to open space. The conceptual layout of a cellular radio system is shown in Figure 1.39

The geographic area served by a cellular system is divided into small geographic cells, ideally in a hexagonal arrangement. Typically, the cell centers are spaced between 6 and 12 km, depending on the terrain and local climate. The system consists of cell sites, a switching center, and mobile units. Each cell represents a microwave radio repeater. The radio facility is housed

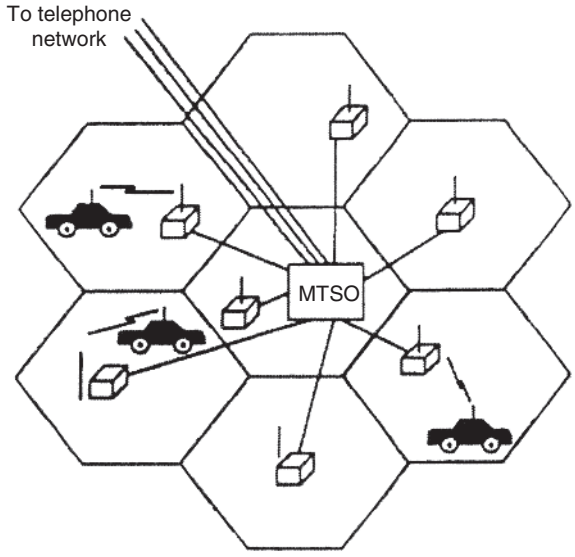


Figure 1.39 Conceptual layout of a cellular system. (*Source*: Freeman 1997 [7]. Reproduced with the permission of John Wiley and Sons.)

in a building or shelter and can connect and control any mobile unit within its own allocated area. The switching facility, referred to as the mobile telephone switching office (MTSO), provides switching and control functions for a group of cell sites. In addition, the MTSO provides connectivity with the public switched telecommunications network (PSTN), adding the distant reach of public telephony to mobile users.

Since the 1990s, cellular systems have seen explosive growth and this exponential growth is likely to continue in the foreseeable future. To meet this high demand, ITU, working with various domestic communication agencies (FCC, European Union, and others) has opened up increasing amount of frequency spectrum for mobile and fixed communication systems. The current state of frequency allocations is summarized in Table 1.8.

In addition, there is ongoing dialogue and meetings among ITU-R, mobile industry players, regulatory agencies, government, manufacturers, and incumbent service providers for further allocations of frequencies. Frequency allocation is a difficult process facing each nation, requiring coordination and harmonization with other systems and services not only within the national boundaries but with international use of the frequency spectrum. The availability of frequency spectrum does not imply that it is easy to get the license for its use. To obtain the license, the operator must successfully negotiate with any existing users of that spectrum, other applicants for the same spectrum, domestic regulatory agencies, and ITU-R. Sometimes it can take years to obtain licenses for the desired spectrum. Frequency allocations are reviewed on a regular basis. For mobile services, the biggest demand is for the frequency band 400–1000 MHz. This lower frequency band is most suitable for providing coverage with lower costs owing to propagation characteristics. L band is a good alternative if frequency spectrum is not available below 1 GHz. Wideband capacity is made available using higher frequency bands.

The architecture of cellular systems requires many trade-offs, especially with respect to the size (coverage area) of the cells, the number of channels allocated to each cell, the layout of the cell sites, and the traffic flow [7]. The total available (one-way) bandwidth is split up into N sets of channel groups. The channels are then allocated to cells, one channel set per cell in a regular pattern that repeats to fill the required number of cells. As N increases, the distance between channel sets (D) increases, reducing the level of interference. As the number of channels sets (N) increases, the number of channels per cell decreases; reducing the system capacity. Selecting

Table 1.8 Summary of frequency allocations from 300 MHz to 30 GHz for ITU region 2.

300–1000 MHz	1000–3000 MHz	3.0–6.0 GHz	6.0–30.0 GHz
300–328.6 (M,F,O)	1427–1525 (M,F,S)	3.3–3.4 (O,M,F)	6.0–7.075 (S,M,F)
335.4–399.9 (M,F)	1700–2690 (M,F,S)	3.4–4.2 (S,M,F)	7.075–7.25 (M,F)
406–430 (M,F,O)		4.4–4.5 (M,F)	7.25–7.85 (S,M,F)
440–470 (M,F,S)		4.5–5 (S,M,F)	7.85–7.9 (M,F)
470–608 (B,M,F)		5.15–5.35 (S,M)	7.9–8.5 (S,M,F)
614–790 (B,M,F)		5.47–5.725 (O,M)	10–10.6 (O,M,F)
790–960 (M,F,B)		5.85–6.0 (S,M,F)	10.7–13.25 (S,M,F)
			14.3–15.35 (S,M,F)
			17.3–19.7 (S, M,F)
			21.4–23.6 (S,M,F)
			24.25–24.64 (O,S,M)
			24.65–29.5 (S,M,F)

Note 1: Letters within brackets represent—M, mobile; F, fixed; S, satellite; B, broadcast; O, others.
The first letter indicates the incumbent service provider, followed by others in order of priority.
Others include radio location, radio astronomy, and radio navigation services
Note 2: Frequency allocations are shared with existing or proposed services requiring ongoing negotiations with ITU-R, (R refers to the radio communication sector) and existing service providers for specific frequency allocations.
Note 3: Spectrum allocations for ITU Regions 1 (Europe) and Region 3 (Asia) are very similar to Region 2 (Americas) allocations
Reference: ITU-R M.2024 (2000), "Summary of spectrum usage survey results."

the optimum number of channel sets is a compromise between capacity and quality. Note that only certain values of N lead to regular repeat patterns without gaps. These are $N = 3,4,7,9$, and 12, and then multiples thereof.

The use of directional antennas can improve performance, allowing smaller cells and higher operating capacity. Regardless of the architecture of the system, each cell site requires highly efficient RF filters to ensure the maximum use of the available frequency spectrum. Figure 1.40 shows the block diagram of a typical RF branching network in a base station using space diversity [22]. Only one transmitter is connected, the other is a redundant backup unit. In the receive direction, two antennas are used to provide space diversity. Use of space diversity is the norm in cellular systems. Thus, one transmit and two receive antennas are required in each direction, for a total of four antennas. Transmit and receive filters connected via a circulator can be replaced by a Diplexer as described in Section 18.5. The circulator coupled network represents the simplest structure but incurs added insertion loss of about 0.1 dB or less per path through the circulator. The diplexer represents the lowest loss structure at the expense of increased design complexity.

Cellular base stations combine the functions of a microwave repeater and a switching network. As a consequence, filtering requirements and constraints are very similar to those encountered in a satellite repeater. Receive and transmit filters must exhibit low loss; elsewhere, loss is not a constraint. Since the majority of cell sites are spread throughout the urban areas, real estate is expensive and, as a consequence, size of equipment is also a constraint. For extra high-reliability systems, both frequency diversity and space diversity are employed as described in Figure 1.41 [22]. It allows for frequency diversity and space diversity in each direction.

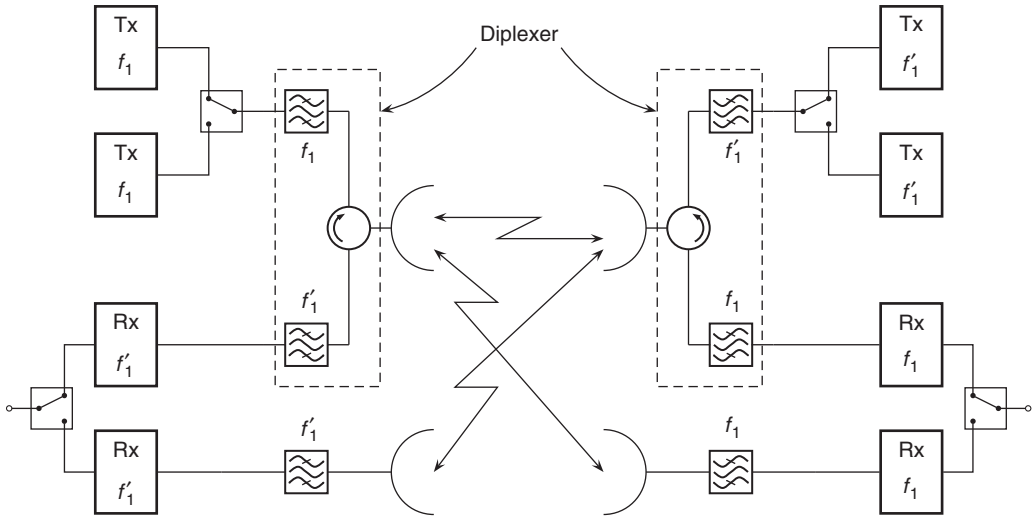


Figure 1.40 Block diagram of RF branching network with space diversity.

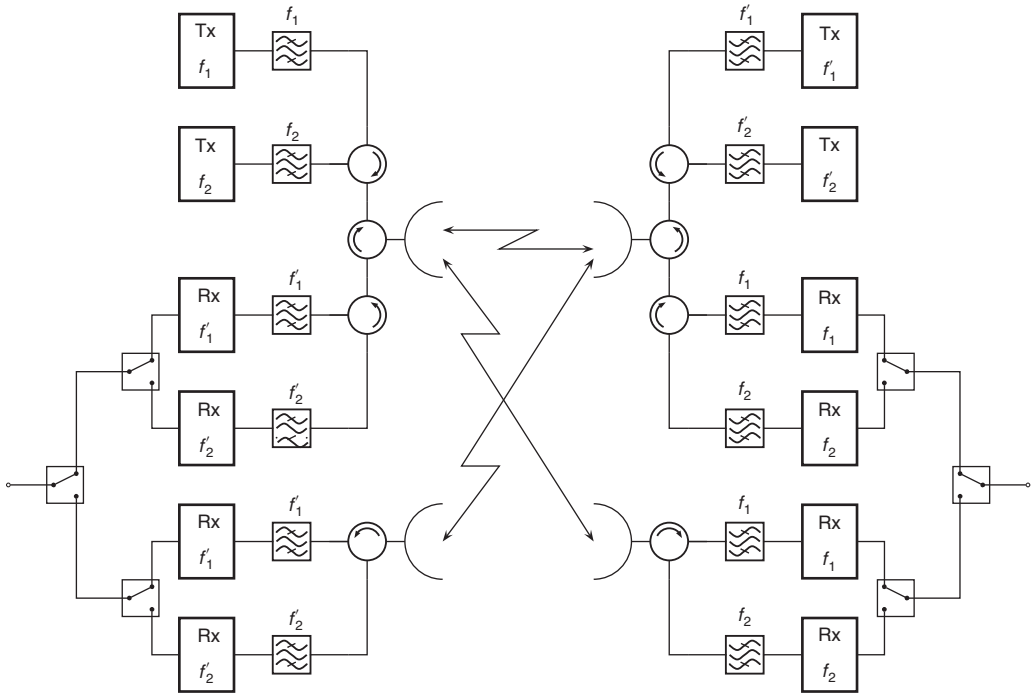


Figure 1.41 Block diagram of RF branching network with frequency and space diversity.

Cellular systems typically assign a single frequency band or channel to a base station to cover a defined geographical area requiring a single passband as shown in Figures 1.40 and 1.41. Owing to increasing demands for more capacity and services, additional frequency bands are being made available to meet this demand. Simple configurations of a two- and three-passband filter network using circulators are described in Figure 1.42. Such a configuration can be extended

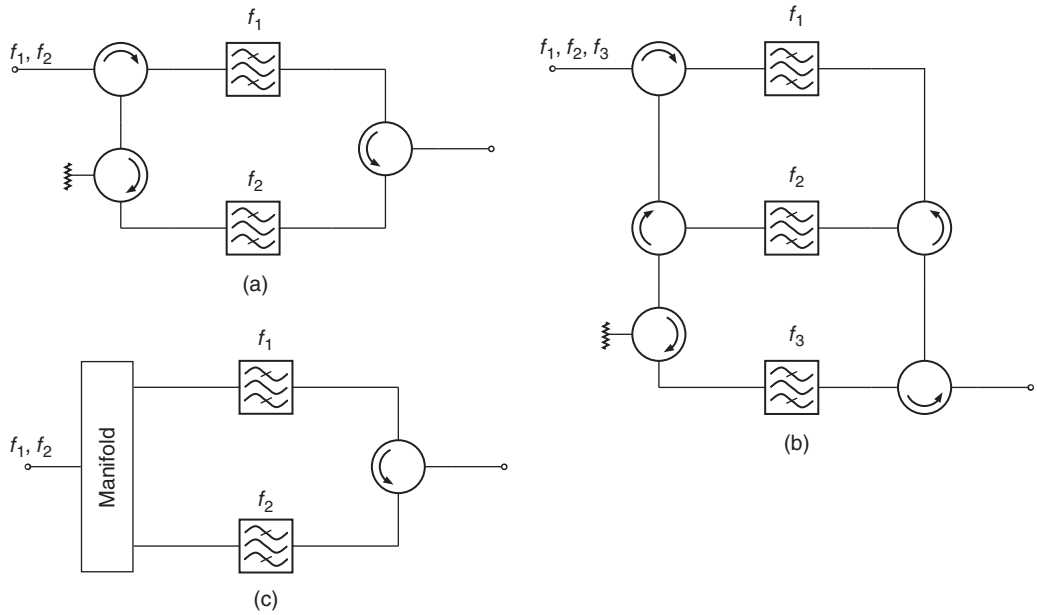


Figure 1.42 (a) Block diagram of circulator-coupled two-channel filter. (b) Three-channel filter. (c) Block diagram of a two-channel filter using manifold coupling at one end and circulator coupling at the other end.

readily to any number of passbands. Its key advantages are (1) simplicity, (2) modularity, and (3) ease of manufacture of single filters to desired performance. The disadvantages include additional hardware in terms of circulators (widely available for commercial applications), insertion loss penalty of approximately 0.3 dB (worst case) per additional passband, and slightly bigger footprint. This configuration is similar to the circulator-coupled multiplexers described in Chapter 18. The increased loss can be dealt with by increasing the power of transmitters by a corresponding amount; for example, to compensate for 0.3 dB loss, power needs to be increased by 7%. This is not an issue for the transmitters; however, it does imply increased usage of electricity that must be taken into account in the overall design of the base stations. The insertion loss penalty can be reduced by a factor of 2 by replacing the input circulators by a manifold-coupled multiplexer as described in Figure 1.42c at the expense of modularity.

Incorporation of additional bands in the base stations has spurred the development of multiband filters, the subject of the new Chapter 21 in this edition. Such filters are finding possible application in radio astronomy and military, suppressing narrowband jamming or interference signals that might occur within the wider usable passband width.

1.10 Ultra Wideband (UWB) Wireless Communication

Ultra Wideband (UWB) traces its origins in the evolution of spread spectrum techniques used for secure communications in military applications. UWB service occupies a very wide bandwidth and shares the frequency spectrum with many other existing services. UWB is a digital pulse wireless technology that can be used to transmit large amounts of digital data over a wide spectrum of frequency bands over short distances with very low power consumption [23–26]. In 1960s–1990s, UWB technology was restricted to defense applications under classified programs for highly secure communication. Since the year 2000, UWB spectrum has been

progressively released globally for short-range wireless communications. Major trends driving short-range wireless in general and UWB in particular are [24]:

- The growing demand for wireless data capability in portable devices
- Paucity of frequency spectrum, segmented and licensed by regulatory agencies
- The growth of high-speed wired access to the Internet in businesses, homes, and public places
- Shrinking costs of microprocessors and signal processing techniques

UWB enables high-speed data transfer over short-range wireless connection as well as applications in low data rate radar and imaging systems. Developments in high-speed microprocessors and fast switching techniques have made UWB technology commercially viable for short-range low-cost consumer communications. The key advantages of UWB can be summed up as [26]:

- Ability to share the frequency spectrum with many users
- Large communication capacity, excellent candidate for high data rate wireless applications
- Low-power density usually below environment noise ensures communication security
- Low sensitivity to multipath effects, ability to carry signals through doors and other obstacles that tend to reflect signals at more limited bandwidths
- Carrier-less UWB transmissions simplify transceiver architecture, thereby reducing cost and implementation time

In 2002, the FCC allowed UWB communication in the 3.1–10.6 GHz band for unlicensed operation having a −10 dB bandwidth greater than 500 MHz and a maximum EIRP spectral density of −41.3 dBm/MHz or 75 nW/MHz. The FCC defined UWB as any signal that occupies more than 500 MHz bandwidth in the 3.1–10.6 GHz band. In addition, FCC and subsequently ECC (Electronic Communications Committee in Europe) issued regulations for the radiated power for the UWB systems as described in Tables 1.9 and 1.10 [25, 26].

UWB is an overlay technology and is susceptible to in band interference from existing narrowband communication bands. Devices using UWB technology transmit ultra-low-power radio signals with short electrical pulses in the picosecond range across all frequencies simultaneously. UWB power levels are required to be below that of noise emissions allowed for electronic equipment. Due to low transmission power and large bandwidth, UWB systems are susceptible to interference from other existing systems. On the other hand, UWB systems are harmless to other communication systems because of their low transmit power.

A key component to remove the unwanted signals and noise from the UWB transmission systems is the bandpass filter. It must ensure that any interference from outside of UWB masks is kept well below the specified power levels for UWB transmissions. The existing communication

Table 1.9 FCC emission masks for indoor and outdoor UWB devices [26].

Frequency range (MHz)	Indoor EIRP (dBm/MHz)	Outdoor EIRP (dBm/MHz)
960–1610	−75.3	−75.3
1610–1990	−53.3	−63.3
1990–3100	−51.3	−61.3
3100–10600	−41.3	−41.3
Above 10600	−51.3	−51.3

Table 1.10 ECC emission masks for the UWB devices in Europe [26].

Frequency range (MHz)	EIRP (dBm/MHz)
Below 1600	−90
1600–2700	−85
2700–3400	−70
3400–3800	−80
3800–6000	−70
6000–8500	−41.3
8500–10600	−65
Above 10600	−85

systems employ narrowband channels (typically <2% bandwidth); for such channels, wide ranging microwave filter design and technologies have been developed over the last four decades. For UWB application, we require filters exceeding 100% bandwidth, a challenging proposition. Such wideband filters are realized using coupled transmission line structures. Many of the wideband filter design techniques are described in Ref. [26].

1.11 Impact of System Requirements on RF Filter Specifications

The emergence and widespread applications of wireless communication systems has stretched the required performance of all components and subsystems that constitute such systems. The scarcity of the available frequency spectrum for wireless communications has put new demands in signal processing and filtering in order to maximize the efficiency of the usage of the frequency spectrum. The factors affecting the specification of RF filters can be categorized as follows:

Frequency plan
Interference environment
Modulation–demodulation schemes
Operating environment
Location of filters in the communication link
HPA characteristics
Limitations of microwave filter technology
Each of these factors can be related to the appropriate area of system design.

1.11.1 Frequency Plan

As described in Section 1.2, the available frequency spectrum for commercially viable wireless communication is a limited natural resource. It is imperative that the communication systems make the most efficient use of the available bandwidth. RF filters perform the function of protecting the overall frequency band from outside interference, efficient channelization of the available bandwidth into various transponders to meet the traffic demands, and efficient power combining to form a common feed, minimizing the cost of the antennas. Consequently, the achievable performance of practical filters largely determines the effective usable bandwidth

of the allocated frequency spectrum. At higher frequencies, filter design is more complex and has higher sensitivity to manufacturing tolerances. Choice of narrower channel bandwidths implies higher loss, higher transmission deviations, and sensitivity. Typical allocation of guard band between RF channels is 10%. It implies an efficiency of 90% in the usage of the available frequency band. However, the allocation of 10% bandwidth to guard band is arbitrary; it merely provides a window to conduct trade-offs between passband performance within the constraint of providing a minimum of isolation at the band edges and beyond of the adjacent channels. For practical systems, performance can be optimized to within 1 dB of amplitude variation, and 2–3 ns of relative group delay over 70–80% of the passband while providing 20–30 dB (or greater) rejection over the adjacent channels. This implies minimal distortion of signals occupying the middle 80% of the passband, with distortion rising toward the band edges. System level trade-offs are critical in optimizing the filter design to maximize the usable bandwidth and hence the efficiency of the channel [16, 22]. The ultimate limit on the achievable bandwidth is imposed by the state of microwave filter technology and cost.

1.11.2 Interference Environment

The interference environment of an RF channel in a communication system is described in Section 1.3. The prime purpose of RF filters is to channelize (or combine) the allocated bandwidth with a minimum of guard band while maintaining a good passband performance. The near-out-of-band isolation requirements are dictated almost entirely by the interference from adjacent (copolarized) channels. The level of interference depends on the energy spectrum of signals in the adjacent channels. It is possible, therefore, to specify isolation in accordance with the shape of the interfering spectrum. This specification is the most severe constraint imposed on channel filters (see Sections 1.8.2 and 1.8.4). It drives the trade-off between the out-of-band amplitude response and the in-band amplitude and group delay variations. The level of such deviations depends on the filter technology, design complexity, and cost. Interference from other sources, such as cross-polarized channels, rain fades, multipath effects, and ionospheric scintillation, fall inside the RF channel and are therefore controlled by the system design. IM interference caused by nonlinear HPAs (Sections 1.5.6 and 1.8.3) is controlled by the operating power level of HPAs.

1.11.3 Modulation Schemes

Modulation schemes play a critical role in maximizing communication capacity of RF channels for different traffic requirements. They also provide a trade-off between power and bandwidth required for transmission of signals. Most systems require a high degree of flexibility in being able to handle different types of modulation schemes as well as a varying number of carriers in an RF channel. Such channel characteristics are best achieved by microwave filters exhibiting equiripple passbands and stopbands; such a response represents the optimal approach to a "brick wall" amplitude response. If necessary, a degree of phase and group delay equalization can be incorporated to achieve improved passband characteristics.

1.11.4 HPA Characteristics

HPAs are inherently wideband and nonlinear devices. In addition to amplifying the RF power, HPAs also generate harmonics of the fundamental frequency and wideband thermal noise. Another consequence of nonlinear characteristics is the generation of IM noise. As a result, HPAs control specification of harmonic suppression and wide out-of-band rejection response, as described in Sections 1.8.2 and 1.8.4.

1.11.5 Location of RF Filter in the Communication Link

A communication link consists of a cascade of transmitters and receivers. RF filters are required after HPAs, prior to LNAs and in between LNA and HPAs as described in Section 1.5. The impact on the specification of filters, based on their location in the communication link, is as follows:

Filter prior to LNA	Low loss
Filters after HPAs	Low loss and high power handling capability
Filters between LNA and HPAs	High selectivity

In addition to electrical requirements, there are always the constraints of size, mass, and cost. Low-loss and high-power handling capability imply larger size. However, if loss is not a constraint, then it is possible to trade-off smaller size at the expense of higher loss.

1.11.6 Operating Environment

For wireless communication, there are two operating environments: outer space or terrestrial space.

1.11.6.1 Space Environment

Key drivers for space applications are mass, size, reliability, and performance. In addition, the equipment must withstand the launch environment, radiation in outer space, and higher temperature range. These requirements have been drivers for the many innovations in filter technology over the last three decades. Typically, the microwave filters used in space are composed of air or dielectric-loaded metal structures and are therefore immune to the radiation environment. However, if high-temperature superconductors (HTS) are adopted to realize such filters, the radiation environment becomes a design constraint, which can be overcome by enclosing filter networks in a metal enclosure of appropriate thickness. Another phenomenon that is peculiar to this environment is multipactor breakdown, which can occur in the hard vacuum of outer space. It is dealt with in Chapter 20.

1.11.6.2 Terrestrial Environment

For terrestrial systems, cost is the main driver, along with the requirement of relatively large quantities of filters. This has spawned innovations in the area of volume production.

1.11.7 Limitations of Microwave Filter Technology

All electrical components dissipate energy; minimization of this energy loss is invariably a key design parameter. In communication systems, bulk of filter applications require narrow bandwidth filters. Energy loss in a filter depends upon two factors: (1) electrical conductivity of metal structures and loss tangent (reciprocal of unloaded Q) of dielectric materials used to realize the microwave filter and (2) the percentage bandwidth of the filter. The narrower the bandwidth of channels is, the higher are the losses incurred in the filter. Bandwidth requirements are a system constraint and is the raison d'etre for using filter networks. Silver has the highest electrical conductivity among metals and is employed as the plating material of microwave structures to achieve the lowest loss. Unplated aluminum or copper structures are often employed as well where specifications can be met without silver plating. For high-power applications, dissipated heat has to be conducted or radiated away to maintain the desired temperature environment, that implies that the materials must also have acceptable thermal

conductivity. All communication equipment has a typical specification of 0 to 50°C as the operating temperature environment. For the narrowband filters at microwave frequencies, this is critical since their performance is very sensitive with respect to any movement of the centre frequency. This implies that the materials must also exhibit excellent thermal stability, often in conflict with the requirements of high electrical and thermal conductivities. In other words, materials used for microwave filters must have low loss (high unloaded Q), good thermal conductivity, and excellent thermal stability.

For any microwave structure, especially narrow band devices such as filters, there is always a trade-off between size (and mass) and the dissipation of energy. The bigger the structure, the lower the loss and vice versa. Another factor that is crucial in determining the size of filters is the dielectric constant of materials. Advances in the quality of dielectric materials over the past few decades, namely, high dielectric constant, low loss (high Q), and excellent thermal stability has made dielectric loaded filters the preferred choice for many applications. Advances in materials are the key in realizing improved performance in ever more compact structures. Some of these trade-offs are discussed in more detail in Chapters 11 and 21.

Microwave filters are distributed structures. This allows realization of microwave filters in a large variety of topologies, as described in Chapters 9–14. Although most filters employ the dominant mode, it is possible to employ higher order propagation modes to realize high Qs at the cost of larger size; such features are well suited for high-power applications. Microwave filter design offers many choices and trade-offs and has given rise to a range of technologies in terms of its implementation. This is highlighted in Table 1.11.

Table 1.11 Microwave filter technologies for communications systems.

Application	Frequency band	Key requirements	Baseline filter technology	Competing technologies
IF Filters for Communication Systems at Large	100–600 MHz	Large number, small size, low power	Lumped Element Networks Microstrip	SAW, MMIC Active
Cellular Systems Handsets Infrastructure	800–900 MHz	Mass production/cost sensitive	Coax Dielectric Resonators Tunable Dielectric	SAW, MMIC, Active Tunable Filters, SIW
PCS/DCS Systems Handsets Infrastructure	2–5 GHz	Mass production/cost sensitive	Coax MIC	Dielectric, MMIC, SIW, Active Tunable Filters
Satellite Systems UHF	300 MHz	Low loss, high and low power, size and mass sensitive	Pressurized Coax	Dielectric
Mobile	1.5 GHz	Low loss, high and low power, size and mass sensitive	Coax Dielectric	Finline/SS/MIC SAW
FSS	4/6, 7/8, 12/14, 11/17 GHz	Low loss, high and low power, size and mass sensitive	Coax, WG, Dielectric	
Multimedia	20/30, 40/60 GHz	Low loss, high and low power, size and mass sensitive	WG, Dielectric	
LMDS	28, 38, 42 GHz	Large Volume, size and mass sensitive	Coax WG	Finline

1.12 Impact of Satellite and Cellular Communications on Filter Technology

In the 1970s and 1980s, the advent of satellite communications was the key R&D driver for microwave filter networks [27]. For space applications, the size and mass of onboard equipment have a critical impact on cost. In addition, power generation in space is costly, and thus, low loss for high-power equipment is equally critical. Lastly, the equipment must be ultra-reliable to operate in space, without failure, for the lifetime of the spacecraft. This led to many advances and innovations in the field of filters and multiplexing networks. These advances included the development of dual- and triple-mode waveguide filters with arbitrary amplitude and phase response, dielectric resonator filters, contiguous and non-contiguous multiplexing networks, surface acoustic wave (SAW) filters, HTS filters, and a variety of coaxial, finline, and microstrip filters. The evolution of widespread wireless cellular communication systems in the 1990s capitalized on this R&D and pushed the envelope further in developing materials and processes for large-scale and cost-effective production techniques for microwave filters for ground-based systems.

Computer-aided design and tuning by EM techniques have played a major role in this endeavor. In an interesting twist of circumstances, advances in production techniques are now being adopted for space microwave equipment to achieve lower costs. This is being driven by the new era of communication systems consisting of constellation of a large number (hundreds to thousands) of low earth orbiting satellites (LEOs). In addition, new frequency bands including millimeter waves are being made available to accommodate the ongoing demand for more and more services using wireless communication, be it satellite based or ground based cellular systems. This demand and new frequency bands will continue to fuel R & D for microwave filters. Some of the promising areas include continuing advances in dielectric materials, tunable filter technology and 3 D manufacturing techniques.

Summary

This chapter is devoted to an overview of communication systems, highlighting the relationship between the communication channel parameters and other elements of the system. The intent here is to provide the reader with sufficient background to be able to appreciate the critical role and requirements of RF filters in communication systems.

Starting with a description of a model of a communication system, the key questions addressed in this chapter include radio spectrum, concept of information, noise and the interference environment, and system considerations in the design of a communication channel.

The concept of information and its transmission over a medium is explored by reviewing the seminal work of Shannon on information theory, highlighting the fundamental trade-offs between signal power, noise in the system, and available bandwidth. This leads to a discussion of trade-offs between power, bandwidths, noise minimization, and antenna performance required to establish a radio link. The topic of noise is considered with emphasis on the thermal noise, and how it sets a fundamental noise floor for the communication channel. Brief overviews of analog and digital modulation are included. The chapter concludes with a discussion of system requirements and the specifications of microwave filter networks in satellite and cellular communication systems.

References

1 Members of Technical Staff (1971) *Transmission Systems for Communications*, 4th edn (revised), Bell Telephone Laboratories, Inc.

2 Taub, H. and Schilling, D.L. (1980) *Principles of Communications Systems*, 3rd edn, McGraw-Hill, New York.
3 Schwartz, M. (1980) *Information, Transmission, Modulation, and Noise*, 3rd edn, McGraw-Hill, New York.
4 Meyers, R.A. (ed.) (1989) *Encyclopedia of Telecommunications*, Academic Press, San Diego.
5 Gordon, G.D. and Morgan, W.L. (1993) *Principles of Communications Satellites*, Wiley, New York.
6 Haykin, S. (2001) *Communication Systems*, 4th edn, Wiley, New York.
7 Freeman, R.L. (1997) *Radio System Design for Telecommunications*, 2nd edn, Wiley, New York.
8 ITU (2002) *Handbook on Satellite Communications*, 3rd edn, Wiley, New York.
9 a Shannon, C.E. (1948) A mathematical theory of communications. *Bell System Technical Journal*, **27** (3), 379–623 and issue 4, 623–656, 1948; b Shannon, C.E. (1949) Communication in the presence of noise. *Proceedings of the IRE*, **37**, 10–21.
10 Krauss, J.N.D. (1950) *Antennas*, McGraw-Hill, New York.
11 Lathi, B.P. (1965) *Signals, Systems and Communication*, Wiley, New York.
12 Cross, T.G. (1966) Intermodulation noise in FM systems due to transmission deviations and AM/FM conversion. *Bell System Technical Journal*, **45**, 1749–1773.
13 Bennett, W.R., Curtis, H.E., and Rice, S.O. (1955) Interchannel interference in FM and PM systems under noise loading conditions. *Bell System Technical Journal*, **34**, 601–636.
14 Garrison, G.J. (1968) Intermodulation distortion in frequency-division multiplex FM systems—a tutorial summary. *IEEE Transactions on Communication Technology*, **16** (2), 289–303.
15 Kudsia, C.M. and O'Donovan, M.V. (1974) *Microwave Filters for Communications Systems*, Artech House, Norwood, MA.
16 Tong, R. and Kudsia, C. (1984) Enhanced performance and increased EIRP in communications satellites using contiguous multiplexers. Proceedings of the 10th AIAA Communication Satellite Systems Conference, Orlando, FL, March 19–22.
17 Sklar, B. (1993) Defining, designing, and evaluating digital communication systems. *IEEE Communications Magazine*, **11**, 91–101.
18 Xiong, F. (1994) Modem technologies in satellite communications. *IEEE Communications Magazine*, **8**, 84–98.
19 Maral, G. and Bousquet, M. (2002) *Satellite Communications Systems*, 4th edn, Wiley, New York.
20 Chapman, R.C., *et al.* (1976) Hidden threat: multicarrier passive component IM generation. Paper 76–296, AIAA/CASI 6th Communications Satellite Systems Conference, Montreal, April 5–8, 1976.
21 Kudsia, C. and Fiedzuisko, J. (1989) High power passive equipment for satellite applications. IEEE MTT-S Workshop Proceedings, Long Beach, CA, June 13–15, 1989.
22 Manning, T. (1999) *Microwave Radio Transmission Design Guide*, Artech House.
23 Roberto Aiello, G. and Rogerson, G.D. (2003) Ultra-wideband wireless systems. *IEEE Microwave Magazine*, **2**, 36–47.
24 Bedell, P. (2005) *Wireless Crash Course*, 2nd edn, McGraw Hill.
25 Wentzloff, D.D. *et al*. (2005) System design considerations for ultra-wideband communication. *IEEE Communication Magazine*, **8**, 114–121.
26 Zhu, L., Sun, S., and Li, R. (2012) *Microwave Bandpass Filters for Wideband Communications*, Wiley.
27 Kudsia, C., Cameron, R., and Tang, W.C. (1992) Innovations in microwave filters and multiplexing networks for communications satellite systems. *IEEE Transactions on Microwave Theory and Techniques*, **40**, 1133–1149.

Appendix 1A

Intermodulation Distortion Summary

Table 1A.1 Direct transmission deviations.

Transmission deviation	Order of distortion	NPR at top modulating frequency (without preemphasis)
Parabolic gain, A_2 (dB/MHz^2)	Third	$\frac{1.72 \times 10^4}{A_2^4 \sigma^4 f_m^4}$
Cubic gain, A_3 (dB/MHz^3)	Second	$\frac{33.6}{A_3^2 \sigma^2 f_m^4}$
Quartic gain, A_4 (dB/MHz^4)	Third	$\frac{6.32}{A_3^2 \sigma^2 f_m^4}$
Linear delay, B_1 (ns/MHz)	Second	$\frac{10^6}{\pi^2 B_1^2 \sigma^2 f_m^2}$
	Third	$\frac{7.5 \times 10^5}{\pi^4 B_1^4 \sigma^4 f_m^4}$
Parabolic delay, B_2 (ns/MHz^2)	Third	$\frac{7.5 \times 10^5}{\pi^2 B_2^2 \sigma^4 f_m^2}$
Cubic delay, B_3 (ns/MHz^3)	Second	$\frac{1.19 \times 10^6}{\pi^2 B_2^3 \sigma^2 f_m^6}$

Key: σ—multichannel RMS frequency deviation in MHz; f_m—top modulating frequency in MHz; A_n—is the nth-order amplitude coefficient in dB/(MHz)^n; B_n—is the nth-order group delay coefficient in ns/(MHz)^n; NPR—noise power ratio; a measure of intermodulation (IM) noise, given by ratio of white noise power spectral density to IM noise power spectral density in a communication channel.

Table 1A.2 Coupled transmission deviations—significant distortion terms only.

Transmission deviation	Order of distortion	NPR at top modulating frequency (without preemphasis)
Linear gain + linear AM/PM, A_1 (dB/MHz) + K_{P1}(•/dB/MHz)	Second	$\frac{3.28 \times 10^3}{K_{P1}^2 A_1^2 \sigma^2 f_m^2}$
Parabolic gain + constant AM/PM, A_2 (dB/MHz^2) + K_{P0}(•/dB)	Second	$\frac{3.28 \times 10^3}{K_{P0}^2 A_2^2 \sigma_m^2 f_m^2}$
Quartic gain + constant AM/PM, A_4 (dB/MHz^4) + K_{P0}(•/dB)	Second	$\frac{9.75 \times 10^2}{K_{P0}^2 A_4^2 \sigma^2 f_m^6}$
Linear delay + linear AM/PM, B_1 (ns/MHz) + K_{P1}(•/dB/MHz)	Second	$\frac{1.73 \times 10^3}{\pi^2 K_{P1}^2 B_1^2 \sigma^2 f_m^4}$
Parabolic delay + constant AM/PM, B_2 (ns/MHz^2) + K_{P0}(•/dB)	Second	$\frac{4.33 \times 10^7}{\pi^2 K_{P0}^2 B_2^2 \sigma^2 f_m^4}$