CHAPTER 1

Fundamental Concepts and Background

INTRODUCTION

In this chapter, we would like to illustrate a few fundamental concepts related to communication systems, circuits, devices, and electromagnetics to serve as a background for the materials to be illustrated in the later chapters. Any integrated system solution is a combination of the following functionalities: (1) data acquisition (sensor/analog interface), (2) signal processing, (3) communication (wireless or wired), and (4) power management. Irrespective of whether the end prototype is intended for wired or wireless communication applications, these four broad functionalities would be present in some form. Although each of these domains is diverse in nature, we illustrate only the fundamental concepts that are used in development of integrated communication microsystems. We start with communication systems, with an illustration of mathematical and physical tools that are necessary for understanding the principles of communication systems. Such tools can be used for design and analysis of systems architecture, circuits, and so on in an analytical, as well as intuitive manner.

1.1 COMMUNICATION SYSTEMS

Although diverse in their nature, wired and wireless communication systems work together to provide end-user services. Figure 1.1 illustrates this aspect. Let us consider the following situation: A user located in a cell in geographical area A needs to

Advanced Integrated Communication Microsystems, edited by Joy Laskar, Sudipto Chakraborty, Manos Tentzeris, Franklin Bien, and Anh-Vu Pham

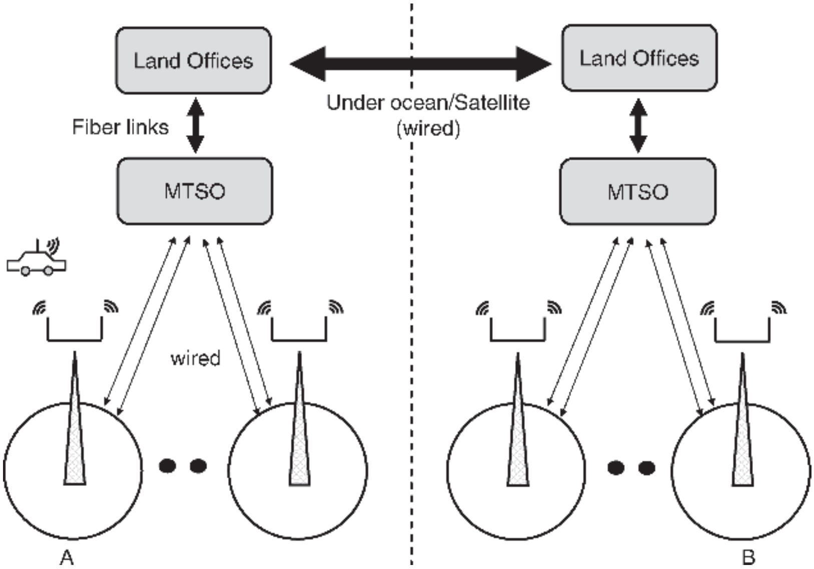


Figure 1.1. Coexistence of wired and wireless communication systems.

communicate to another user in a geographical area B while moving on the highway, at the end of work. Call from the mobile phone is accurately received by the base station in area A and communicated to the Mobile Terminal Switching Office (MTSO). Various MTSOs are connected to the central switching office by optical fiber backbone; they communicate with the central office, which communicates with its counterpart in B through optical fiber links laid underneath oceans or through geostationary satellites (with some communication delay). Modern communication systems mostly use optical fibers. As the message is received by the central office in B, it then diverts the traffic to a specific city, and the specific user gets the call from a telephone exchange. In case the end user is also mobile, the central office then communicates with another MTSO, which is responsible for delivering the message to the appropriate mobile user.

The entire process is complicated, in terms of its switching, traffic handling, and other network management issues. The above example has been used to illustrate the basic mechanism of a voice communication. Other types of high-data-rate communications are also feasible. For example, transferring large files, or multimedia movies, from one wireless device to another falls in the same category of high-speed wireless communications. Many times it is difficult to lay fiber optic cables because of geographical problems (rough terrains, mountains, etc.), and a direct line-of-sight wireless communication may be preferred. Our focus in this book is to provide an understanding of how to develop the physical-level hardware solution to enable such communication systems.

Our focus in this chapter is on the physical layer of these communication systems, in order to develop insight toward developing miniaturized hardware. A single chip, which can perform the functionalities of wireless communications at a desired data rate and frequency within a required power and area is the subject of this book. As the two communication systems are essentially diverse in nature, we focus on the various considerations toward wireless and wired communication systems. First we illustrate the nature of each of these communication systems and their fundamental aspects. Then we cover the key background needed for appreciation and design of such systems. This background is essential for developing any type of systems, wireless or wired.

1.2 HISTORY AND OVERVIEW OF WIRELESS COMMUNICATION SYSTEMS

The basic developments in the area of wireless communication date back to the early twentieth century. Since those early years, wireless engineering has come a long way. Most of the basic principles of the sophisticated radio architecture, as we see it today, were developed using vacuum tubes around 1930. Starting with the basic foundation provided by Maxwell (1883), and with subsequent inventions in wave propagation and wireless telegraphy by Hertz, Bose, Marconi, and others, wireless technology was born around 1900 in a very primitive form. Demonstration of a superheterodyne receiver by Armstrong dates back to as early as 1924. Various illustrations of Armstrong's superheterodyne receiver were reported during the 1920s and 1930s. At this time, radio pioneers considered the use of homodyne (/direct conversion) architectures for single vacuum tube receivers. For over two decades, the standard low-end consumer AM-tunable radio used a system of five vacuum tubes. A major milestone was set by the invention of the transistor by Bardeen, Brattain, and Schockley in 1948, which changed the world of vacuum tubes. However, implementing radios was a farsighted vision at that time. As semiconductor technologies became more mature, more circuit integration took place. Starting with small-scale integration in the standard integrated circuits, the trend moved toward more integration and high-speed microprocessors. With the tremendous growth in digital signal processing, very large-scale integration (VLSI), demands for ubiquitous computing and wireless applications increased.

During the 1990s, the maturity of digital electronics and signal processing hardwares led to the perception that a single-chip implementation of the front end could be feasible. This belief led to various developments of integrated filters, radio architectures based on frequency planning [super heterodyne to low intermediate frequency (IF) to direct conversion], and modulation techniques (such as DC-free spectrum) to combat known problems associated with direct conversion and so on.

Two fundamental operations of a receiver/transmitter include down/upconversion and demod(/mod)ulation. However, this is different in the case of coherent versus noncoherent radios. In the downconversion function, the desired signal is filtered and separated from the interferers, and it is converted from the carrier frequency to a frequency suitable for the demodulator for low signal processing power. Demodulation is performed at a lower frequency, either by a simple in-phase and quadrature phase (I/Q) demodulator or digitally sampled and performed by a digital signal processor (DSP). The latter allows for the use of complicated modulation schemes and complex demodulation algorithms. The demod(/mod)ulator and the other signal processing functionalities are usually performed using a digital signal processor, and its power consumption can be reduced by using advanced process technology nodes (which reduces the supply voltage and area). However, the down/upconversion functionality is not easily scalable, and the power consumption is a function of operating frequency, bandwidth, as well as intermediate frequency (which is dependent on blockers). Thus, numerous radio architectures are considered. Modern communication devices provide more and more integration on chip. The use of lower IF or elimination of IF from the frequency plan has many implications on the receiver/transmitter architecture. Low IF receivers combine the advantages of zero IF and IF architectures. It can achieve the performance advantages of an IF receiver, reaching the high level of integration as in a zero IF receiver.

1.3 HISTORY AND OVERVIEW OF WIRED COMMUNICATION SYSTEMS

Advanced wired communication systems today require transfer of multi-Gb/s data rate across bandlimited channels. Even computer hardware requires clock speeds of more than 2 GHz to be sent over motherboards. Overall, 10 Gb/s serial data have been transferred over FR-4-based backplanes, which were originally designed for 1-Gbps Ethernet applications. Advances in optical links and supporting electronics have dramatically increased the speed and amount of data traffic handled by a network system. Bandlimited channels continue to be a critical bottleneck for delivery of multigigabit serial data traffic.

The primary physical impediments to high data rates in legacy backplane channels are the frequency-dependent loss characteristics of copper channels. Above rates of 2 Gbit/s, the skin effect and dielectric loss in backplane copper channels distort the signal to such a degree that signal integrity is severely impaired. This dispersive forward channel characteristic contributes to the Inter-Symbol Interference (ISI).

Meanwhile, a major limiting factor to increasing transmission speeds and distances in fiber-optic communication links is modal dispersion causing ISI. Modal dispersion is caused as the numerous guided modes are transmitted in different paths in the multimode fiber (MMF) resulting in different receiving times at the receiver side of the fiber communication system. Modal dispersion becomes a severe factor as the length of the MMF is extended or the data rates are increased.

A brief comparison/contrast between wireless and wired systems can be represented as follows:

Electrical Characteristics	Wireless	Wired
Impact of channel	Mostly attenuation, and fading caused by path loss	Mostly dispersion caused by group delay variation
Bandwidth	Inherently narrowband	Inherently broadband
Effect of interferences	More interferers	Less interferers

(*Continued*)

Electrical Characteristics	Wireless	Wired
Synchronization problems	Very significant issue	A major issue to be considered
Modulation	A variety of modulation techniques are present starting from BPSK, QAM etc.	Mostly OOK, and some multilevel signaling in the electrical domain
Noise	Device noise plays a major role, as the signal is quite weak	Device noise is not an issue, as the signal levels are quite high
Architectures	Differ with each other in terms of frequency shift, up or down	Differ with each other in terms of synchronization schemes, half-rate/full-rate clock-data recovery systems, etc.
External components	Usually filter, balun, switch, duplexer (all electric/electromagnetic in nature)	Usually photodiodes, VCSEL, other lasers, and electronic couplers

1.4 COMMUNICATION SYSTEM FUNDAMENTALS

In a wireless communication system, communication channel characteristics are defined by the environment in which we decide to operate, and this may vary among rural, urban, suburban, hilly area, and so on. In the case of wired communication, the choice of channel is dependent on the distance we want to communicate over, and the overall cost of the material (usually multimode or single-mode optical fiber). Once again, our target is on the channel capacity and the Signal-to-noise ratio (SNR) degradations associated with it.

1.4.1 Channel Capacity

The capacity of the channel is defined by Shannon–Hartley theorem, which is defined as

$$C = B\log_2\left(1+\frac{S}{N}\right) = B\log_2\left(1+\frac{E_b}{N_0}\right) \times \left(\frac{C}{W}\right)$$

where C is the channel capacity (bits/s), $\frac{S}{N}$ is the SNR obtained from average signal and noise powers, and $\frac{E_b}{N_0}$ is the energy ratio of the bit to noise energy, also known as "bit-energy per noise-density."

This theorem shows the achievable limit on the transmission bit rate, whereas the accuracy is given by whether the transmission bit rate, $R \leq C$. With this condition, the probability of error could be sufficiently small by using some channel coding, whereas in the region of $R > C$, no channel coding would lead to a sufficiently small error rate.

1.4.2 Bandwidth and Power Tradeoff

The above equation also leads to interesting consequences in terms of two aspects, a bandwidth-limited transmission scenario and a power-limited transmission scenario. In a bandwidth-limited situation, the transmission bandwidth is higher than the channel bandwidth, and we use symbols to represent several bits, along with some channel coding. This is possible, however, with a compromise in higher bit energy per noise density. It is certainly possible to consider a situation in which one may be interested in transmitting a lower bit rate through a higher capacity channel, while operating in the power-limited region of the capacity $\frac{E_b}{N_0}$ plane. This situation is illustrated in Figure 1.2, and it leads to fundamental considerations while determining radio architectures. For example, it a spectrally efficient modulation scheme, more bits would be packed in a symbol, which leads to the requirements of moderate to high accuracy for the signal processing necessary. To design such systems, a certain amount of power should be consumed to ensure the accuracy of the signal processing. In power-limited modulation techniques, low spectral efficiency modulation techniques are usually preferred. These techniques are used for low/moderate data rate systems, where battery longevity is the prime consideration.

One can also conclude from the above equation that, in the regime of low SNR communication systems, the logarithmic nature can be expressed as

$$W = C \times \left(\frac{S}{N} \right)$$

which implies that one can extend the bandwidth significantly while using a low SNR. This is particularly applicable to ultrawideband systems. Communication is performed by embedding information in the amplitude, frequency, and/or phase of

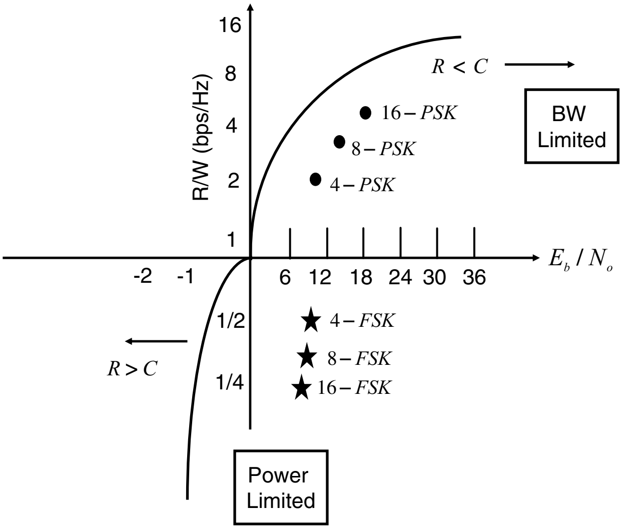


Figure 1.2. Bandwidth-efficiency plane.

the transmitted signal. Any communication system design is a tradeoff between the bandwidth and power usage. The task of the receiver is to recover the transmitted information successfully, which has discrete states w.r.t amplitude and time.

Radio architectures also evolve around these fundamentals. In a bandwidthlimited modulation scheme, spectral efficiency is a key factor, and the target is to "pack" a maximum number of bits into a symbol, in order to achieve high data rates. However, this process requires high signal processing accuracy. On the other hand, power efficient modulation schemes lead to low-power hardware, at the expense of low spectral efficiency. Depending on the application, each scheme should be chosen to fit the needs.

1.4.3 SNR as a Metric

Given a wireless environment, operating frequency, and distance, one can easily calculate the path loss, and the SNR degradations caused by multipath and shadowing effects. One can identify various geographical regions and map the associated SNR with them. Once these are determined, then, it is calculated how much SNR degradation is obtained from the RF/analog front end. After the signal processing at the RF/analog front end is performed, the demodulator obtains a specific SNR. We would then refer to the waterfall curve of the bit error rate in order to obtain a suitable modulation scheme. Thus, SNR is the major performance parameter that determines the choice of modulation scheme.

From an RF/analog perspective, a certain modulation scheme, bit error rate, and channel coding scheme defines the available bandwidth. The SNR degradation resulting from RF/analog blocks is contributed by the regular noise phenomena such as thermal noise, flicker noise, as well as intermodulation distortion product. These effects are further complicated in the case of wideband systems. The SNR improvements and, hence, the signal processing accuracies in the RF/analog front end are dependent on the power consumption.

The operating frequency is an important parameter in deciding the feasibility and cost of a communication system to be deployed. The propagation characteristics in a free space (path loss) is a function of frequency and the distance, and they are given by

$$L = \left(\frac{4\pi d}{\lambda}\right)^2$$

Thus, the path loss is lower at low frequencies, which leads to better signal propagation. However, the antenna size is inversely proportional to frequency of operation. However, it should be kept in mind that the above equation is a free space loss only. In an office environment or a home environment, the path loss assumes a much different profile, and the losses are usually much higher. In the case of mobile devices, the Doppler effect occurs between mobile devices, which needs estimation and compensation algorithms.

1.4.4 Operating Frequency

Although the above arguments hold good, choice of operating frequency is strongly motivated by licensed free frequency spectrum. Several frequency bands are dedicated for industrial, scientific, and medical applications under the FCC regulations. Medical applications usually operate in dedicated frequency bands because of the high reliability considerations of these devices. Such ISM bands are located in the 315M/433M/868M/915M/2400M bands. Associated with these center frequencies, there are various interference patterns from adjacent frequency bands. All ISM band devices have restrictions on maximum transmitter power as well. These restrictions, almost always determine the usable frequency band and the maximum distance achievable. The frequency allocation of various bands is shown in Figure 1.3 along with their applications.

Choice of frequency is a major decision point in the implementation of an integrated system. At lower frequencies, the data rate is lower, medium propagation is better, and a large antenna would be required. At higher frequencies, data rates are higher, medium propagation is worse, and a smaller antenna would be required for a low form factor solution. To trade off these constraints, most of the commercially available

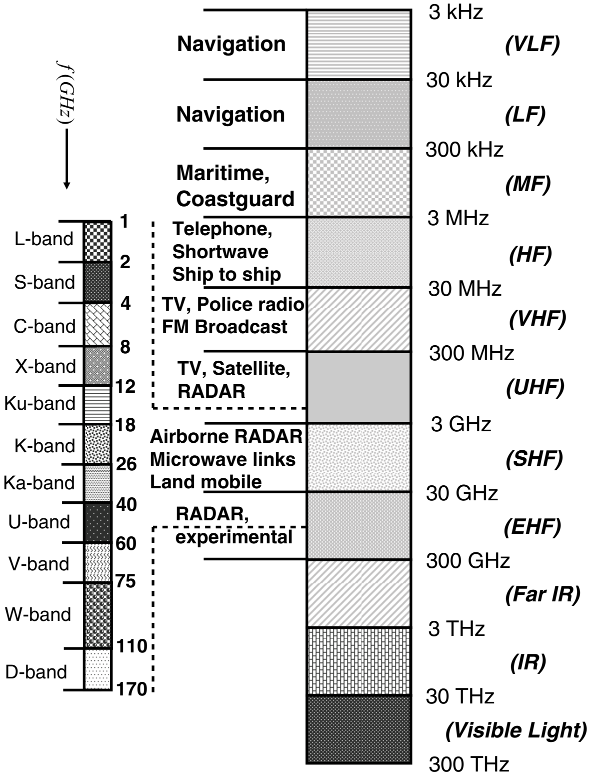


Figure 1.3. Frequency bands and allocations.

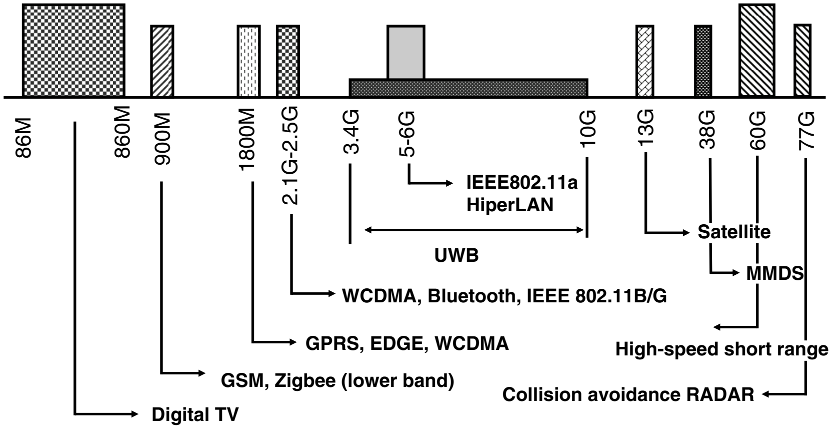


Figure 1.4. Commercial applications over frequency bands.

frequency bands range from 1 Ghz to 10 Ghz in present state-of-the-art cellular and wireless local area network (LAN) systems. Several emerging applications tend to operate in higher frequency bands as shown in Figure 1.4.

1.4.5 The Cellular Concept

Within the allocated frequency band of interest, multiple users can be accommodated by providing various frequency channels. It is the basis of the cellular radio system, as illustrated in Figure 1.5. A specific geographical area can be divided into multiple such cells, with frequency reuse planning. In practice, the individual cells are not

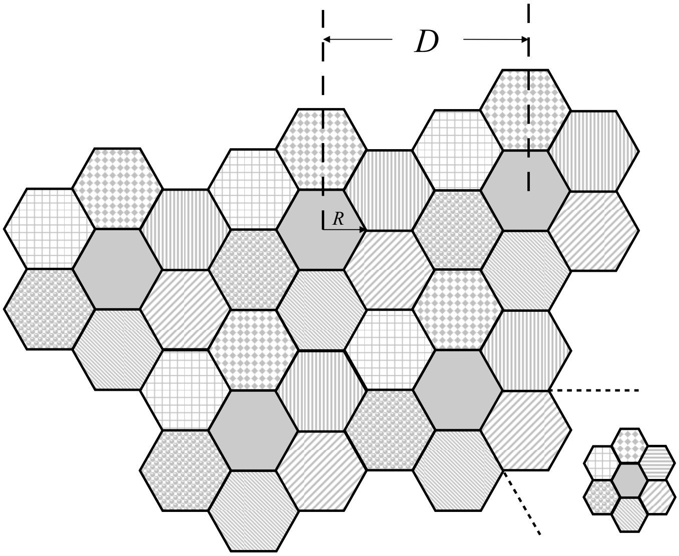


Figure 1.5. Cellular communication: frequency reuse, cell splitting.

hexagonal. In case of heavy traffic in an individual cell, it can be further divided into multiple smaller cells, and each one of the smaller cells operates at lower powers. Such allocations are dynamic in nature, which leads to increased flexibility of the cell-allocation scheme. This flexibility is illustrated in Figure 1.5, where different shades represent individual frequencies.

In analog communication systems, both the message signals as well as the time at which they are sent can assume continuous values. However, in a digital communication system, the information used and processed is discrete in time and amplitude. The fundamental aspects of any communication system design include (1) bandwidth, (2) power, and (3) error correction capability.

Obvious as it may seem, the bandwidth usage is specific to the FCC restrictions in specific countries under consideration. The task of the communication system designer is to select a specific frequency band and determine a modulation scheme such that the information transfer can be maximized at a given time. Depending on the situation, one can operate in the license-free ISM bands or the licensed bands with proper permissions from the governing agencies. Once the frequency is chosen, one should consider how much bandwidth is to be used, and what modulation technique is to be used.

1.4.6 Digital Communications

Our emphasis is on digital communication. In a digital communication system, information is arranged in a set of discrete amplitude as well as at discrete time instants. Such a signal, even being simply upsampled by a clock waveform, would lead to a digital waveform, which would lead to spectral spillover at the front end. At the same time, the antenna would need to be infinitely broadband in nature in order to accommodate all the useful information that is obtained. This would be highly inefficient, and we simply cannot transmit a digital waveform, however delicately it has been processed using careful techniques.

This issue is solved by using the concept of symbols. Symbols are formed from the sets of raw bits in the system, and as the information is discrete in amplitude and time, the symbols assume discrete states, and a diagram illustrating this is shown in Figure 1.6. Hence, one symbol may represent multiple bits at a time, and depending on the number of bits, the digital modulation is named. In binary notation, for an M ary communication, we obtain $\log_2 M$ symbols. Even symbols are digital in nature, so we have not really solved the spectral spillover problem.

The answer comes in constructing specific analog waveforms, which are bandlimited in nature. These waveforms can be obtained as a combination of three fundamental factors in information communication through waveforms by changing its (1) amplitude, (2) frequency, (3) phase, or a combination of them. Hence, we associate a specific analog waveform governed by our prespecific rule with each symbol. This process provides spectral containment. As the number M increases, we can associate more bits per symbol, and per analog waveform, or in other words, one waveform would contain the same information as so many raw bits. The nature of analog waveforms essentially determines the bandwidth, so the bandwidth in a digital communication is always associated with symbol duration.

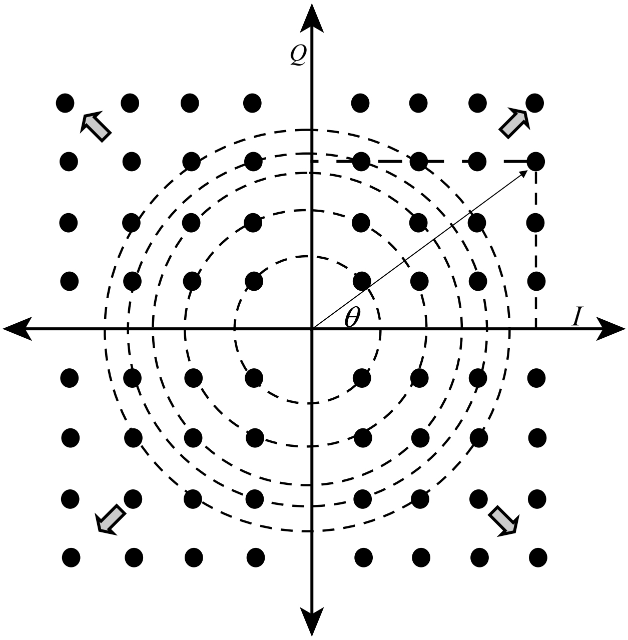


Figure 1.6. Signal constellation: illustration in polar/rectangular format.

1.4.7 Power Constraint

A communication system may be viewed as a combination of signal processing operations, which consume a certain amount of power. The signal processing can be continuous or discrete time in nature. If power consumption restriction is not present, the signal can be made arbitrarily large, and the individual symbol amplitudes can be made much differently from one another, which leads to their easy detection in case there is an error. However, we want to obtain the highest amount of information transfer in a given power budget. Hence, we need to obtain a proper choice of the modulation scheme. If the modulation order (*M*) is higher, the symbol constellation would become denser, and the distance between two symbols would reduce. In the case of a low order modulation scheme, the reverse is true. In the two cases, we have assumed the overall symbol power to be the same. Thus, for the same power, the higher *M* constellation symbols appear closer to one another, which leads to tolerance of the lower amount of impairment from noise, or require "higher SNR." Noise can appear from (1) quantization, (2) analog impairments, and (3) channel.

After the transmission is performed, the signal goes through several impairments in terms of channel characteristics and RF nonlinearity in the receiver. The digital receiver would then retrieve the correct state from the received impaired signal. There are several ways in which the transmitted signal can be distorted, including (1) intersymbol interference and (2) carrier offset.

Let us assume that the transmitted symbols are $x_1, x_2, x_3, \ldots x_n$. Assuming a linear superposition behavior, the received signal may obtain a value of $0.1 \times x_1 + 0.9 \times x_2$, instead of the second symbol x_2. This change may be caused by intersymbol interference, and in reality, it is a complex function of channel impulse response. At the same time, the carrier frequency generated from the VCO may provide a constant offset frequency, and the received symbols may appear to be as $x_1, x_2e^{j\phi}, x_3e^{2j\phi}, x_4e^{3j\phi}\ldots$. If the rotation of the symbols is larger, then symbols can get wound up in a manner in which the rotated symbols can significantly impact the detection of the received signal. A system designer always wants these degradations to be lowest, and comes up with the right selection of architectures and algorithms to mitigate these impairments.

1.4.8 Symbol Constellation

All of the impairments can be characterized in terms of symbol constellation, which is a graphical representation of symbols in a communication system with an orthogonal set of vectors (usually termed as in-phase and quadrature). Constellation represents the finite set of symbols in a digital communication system using a set of predefined states, using orthogonal axes. An example of orthogonal representation would be to represent symbols using amplitude and phase (I and Q), and it can be very well used in the case of BPSK, QPSK, QAM, etc. As the constellations become denser, which is the case of bandwidth-limited modulation, the SNR requirements become increasingly higher from RF/analog front ends. In the receiver, the digital demodulator delivers a reliable set of bits given distorted, quantized received signals at an oversampling factor N.

From the viewpoint of a communication system engineer, we would like to transmit the maximum possible bit rate while achieving the minimum probability of error with minimum available bandwidth and minimum SNR. We would also like to design the system for minimum complexity and maximize the number of users with a good quality of service in terms of delay and interference immunity. As these demands are contradictory to each other, a compromise needs to be obtained. Unlike the analog system, which works on reproducing original waveforms, digital communication sends waveforms to represent digits obtained by sampling the original waveform. Analog systems contain infinite energy, but finite power, whereas digital communication waveforms are of zero average power, with a finite energy. For this reason, the digital communication systems are better represented in terms of bit energy, with E_b/N_0 leading to bit error rate performance. From the above argument, it is clear why a communication system designer is always concerned with signal-to-noise degradations (SNR) in various signal processing blocks.

Currently, there have been various reports of communication system standards: architecture proposal. Key aspects of these standards include (1) communication channel under consideration; (2) center frequency, data rate, and distance; (3) modulation scheme; (4) connection protocol; and (5) targeted application.

1.4.9 Quadrature Basis and Sideband Combination

A specific way to understand transmitter and receiver architectures is by means of frequency translation and phase rotation. Let us consider direct conversion architecture as an example. In the transmitter, baseband signals are processed using a

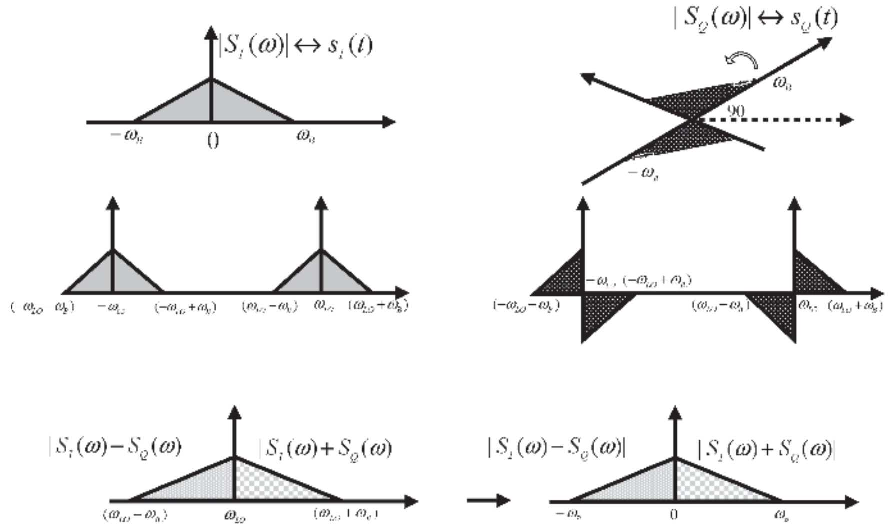


Figure 1.7. Frequency translation and negative frequency concept.

high-speed digital signal processor (depending on the data rate) to generate two streams of signals, in phase and quadrature. These streams are implemented by interleaving the original message sequence in in-phase and quadrature components and adjusting the delay between the two streams. This signal is then processed using a digital-to-analog converter (DAC), and finally up-converted by the LO frequency using quadrature phases, and combined at the output to obtain a single sideband. The mathematical synthesis can be represented as

$$S_{TX}(t) = A_{BB}\cos\omega_{BB}t \times \cos\omega_{LO}t - A_{BB}\sin\omega_{BB}t \times \sin\omega_{LO}t = A\cos(\omega_{LO} + \omega_{BB})t,$$

variations of this trigonometric formulation are also shown in Appendix A(1).

The frequency translation is illustrated in Figure 1.7.

1.4.10 Negative Frequency

In the frequency domain representation, the spectrum is symmetrically arranged around the LO frequency. The downconversion can be represented as a frequency translation in order to obtain the original signal centered around DC. Both sidebands contain a different amount of information. Both situations are consistent only if there is a "negative" frequency at the baseband. However, all along we are using real signals for illustration, and there is not a concept of negative frequency, we cannot generate it, and we cannot perceive it.

To understand this aspect, we represent signals as $S(t) = I(t) + jQ(t)$, where $I(t)$ and $Q(t)$ are real valued functions and "j" simply represents a "rotation" (or it could be thought about a transformation to construct a new variable). Thus, the frequency domain representation of this signal in frequency domain would contain $I(\omega) + Q(\omega)$

for positive ω and $I(\omega) - Q(\omega)$ for negative ω values. This is now upconverted at the transmitter and downconverted at the receiver. From $I(\omega) + Q(\omega)$ and $I(\omega) - Q(\omega)$, we can easily reconstruct $I(\omega)$, and $Q(\omega)$, and their time domain waveforms.

1.5 ELECTROMAGNETICS

Almost all developments in the area of communication systems, devices, and circuits can be correlated to some aspects of electromagnetics. Electromagnetic principles can well explain the propagation of waves, basis of wireless communication, skin effects of integrated inductors, Kirchoff's laws governing all areas of circuit design, standing waves formation, and the nature of electric field in scaled semiconductor devices.

1.5.1 Maxwell's Equations

Time-varying electrical and magnetic fields and their relationship with one another can be governed by Maxwell's equations, which is a generalized form of experimental results obtained by many researchers. In terms of generalized spatial coordinates, they can be stated as $E(x, y, z, t)$ for electric field and $B(x, y, z, t)$ for magnetic field. In free space, they are governed by the following four fundamental equations:

$$\begin{aligned} \nabla \cdot B &= 0 \\ \nabla \cdot D &= 0 \\ \nabla X H &= J + \frac{\partial D}{\partial t} \\ \nabla X E &= -\frac{\partial B}{\partial t} \end{aligned}$$

where $B = \mu_0 H$ and $D = \varepsilon_0 E$ in free space. These equations uniquely determine the nature of magnetic and electric fields at any spatial point as a function of time. The last two equations imply the inherent coupling between electric and magnetic fields (change in electric field produces change in magnetic field and vice versa). Hence, two coupled first-order differential equations lead to formation of a second-order differential equation in individual variables, and they form the basis of a standing wave, which is used in the context of almost all electromagnetic phenomena. The first equation implies the absence of magnetic monopoles, and the second equation implies that the divergence of electric field is dependent on the net electric charge.

1.5.2 Application to Circuit Design

As a first application of the above equations, we consider the circuit design principles. The assumption is that there is no coupling between electric and magnetic fields, which is obtained by setting $\mu_0 = 0$, and $\varepsilon_0 = 0$, leading to

$$\begin{aligned} \nabla X H &= J \Rightarrow \nabla \cdot (\nabla X H) = \nabla \cdot J = 0 \\ \nabla X E &= 0 \Rightarrow \oint E \cdot dl = 0 \Rightarrow \oint -(\nabla \cdot V).dl = 0 \end{aligned}$$

Divergence of curl of a vector is zero, and we use Stoke's theorem of line integral to obtain the second formulation. The first one implies that there is no divergence of current, implying that the sum of all the currents flowing to a node would be zero. The second one implies that in a loop, the sum of all the voltages along the loop would be zero. These form the basic theory of any circuit operation (Kirchoff's laws). The assumption of decoupled electric and magnetic fields is valid under the assumption that the lengths of the loop under consideration are much smaller than the wavelength under consideration. This is related to the physical dimension of the circuit under consideration, and in a semiconductor substrate, the wavelength is reduced by the relative dielectric constant.

1.5.3 Signal Propagation in Wireless Medium

In the above discussion, we have obtained the fundamentals of circuit theory under the assumption that electric and magnetic fields are uncoupled. In a coupled relationship, the wave propagation can be illustrated. A loop of wire carrying time-varying electric current causes a time-varying magnetic field around it. This changing magnetic field causes a continued time-varying electric field, and this happens in a three-dimensional fashion and with speed of light in the medium under consideration. Figure 1.8 illustrates this concept, which is the fundamental basis of radio propagation through air.

Maxwell's equations related to divergence the of electric field ($\nabla \cdot D = \rho$) provides fundamental equations related to device physics in the case of semiconductor junctions, maximum electric field, and so on, and they lead to a discipline well known as "electrostatics." Similarly the first equation related to magnetic fields (magnetostatics, also known as Biot–Savart's law) leads to understanding the nature of magnetic field lines resulting from inductors and so on. Hence, the generalized Maxwell's equations explain almost all the aspects, including circuit design, wave propagation, electromagnetic field lines, and so on. In modern integrated systems, these can be used to solve various problems, and in many cases, they are solved numerically in the case of practical problems in order to maintain computation speed and accuracy.

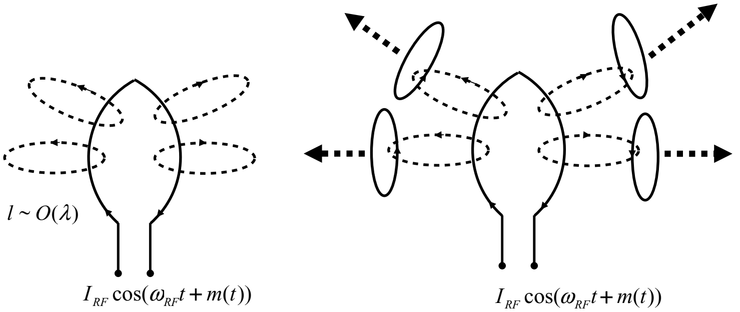

$l \sim O(\lambda)$
$I_{RF}\cos(\omega_{RF}t + m(t))$
$I_{RF}\cos(\omega_{RF}t + m(t))$

Figure 1.8. Propagation of waves from an antenna.

1.6 ANALYSIS OF CIRCUITS AND SYSTEMS

Several methods can be used to represent communication circuits and systems. We will illustrate a few of the analysis tools, which are used to obtain numerical efficiency, spectral information, as well as physical insights.

The first in this category of tools are a few transformations and signal processing components. The mathematical nature of signals in any communication systems can be represented in time or frequency domain. Transformations help with the conversion of complicated differential equations to linear equations, subject to the initial conditions, for simplified mathematical analysis. These processes were originally developed to analyze partial differential equations with boundary value problems, and later they were adopted in a variety of engineering disciplines.

1.6.1 Laplace Transformation

Laplace transformation is defined as $F(s) = \int_0^\infty e^{-st} f(t)dt$. This integral exists when *f* (*t*) grows slower than e^{bt}, such that convergence is obtained. *f*(*t*) need not be a continuous function, and it may be simply a piecewise linear function. If the transformation exists, it is uniquely determined. This transformation can be applied to convolution of two functions, which provides the multiplication of individual Laplace transforms. Laplace transformation can be used to provide impedances of inductance and capacitance, which are obtained under the conditions that the current through an inductor remains the same before and after an event occurred at time instant τ, while the voltage across a capacitor remains the same before and after a time instant τ. In the case of a simple example, we assume the initial current and charge values are 0, respectively. Since inductor and capacitors are represented by differential equations in the time domain, their voltage and current waveforms are represented as follows:

$$V_L(t) = L\frac{di}{dt}$$

$$V_C(t) = \frac{1}{C}\int \frac{di}{dt}$$

The Laplace transformation implies that $V_L(s) = (sL)I(s)$, and $V_C(s) = \frac{1}{sC}I(s)$; hence, inductor and capacitors are represented by frequency-dependent impedance values of sL, and $1/sC$, respectively.

1.6.2 Fourier Series

Analysis of periodic waveforms can be represented by Fourier series expansions. We first start with the analysis of periodic signals in the time domain. Periodic functions occur in numerous places, such as the LO drive of the mixer, which can assume various waveform types (sinusoidal, square, etc.). On the other hand, the input RF signal has a sinusoidal waveform shape. For a time domain waveform with period $p = 2T$, the

Fourier series can be represented as follows:

$$s(t) = a_0 + \sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi}{T}t + b_n\sin\frac{n\pi}{T}t\right)$$

and the Fourier coefficients are determined by

$$a_0 = \frac{1}{2T}\int_{-T}^{T} f(t)dt$$

$$a_n = \frac{1}{T}\int_{-T}^{T} f(t)\cos\frac{n\pi t}{T}dt, \quad n = 1, 2, 3\ldots$$

$$b_n = \frac{1}{T}\int_{-T}^{T} f(t)\sin\frac{n\pi t}{T}dt, \quad n = 1, 2, 3\ldots$$

For a different period, T can be replaced accordingly in order to obtain the desired Fourier series representation. As an example, a periodic square waveform, and a half-wave rectifier, can be considered, which are classical waveform shapes in electrical systems. The square waveform contains odd harmonics of the period, whereas the half-wave rectifier contains even harmonics of the period. The Fourier transform of the square waveform illustrated in Figure 1.9 is given by

$$s(t) = \frac{A}{2} + \frac{2A}{\pi T}\left[\cos\frac{\pi t}{2T} - \frac{1}{3}\cos\frac{3\pi t}{2T} + \frac{1}{5}\cos\frac{5\pi t}{2T} - \ldots\right]$$

which contains only the odd harmonics of the waveform frequency under consideration.

A phase shift (or delay) in the original signal $s(t)$ would lead to integral multiples of phase shift, depending on the harmonic tone under consideration, as phase is

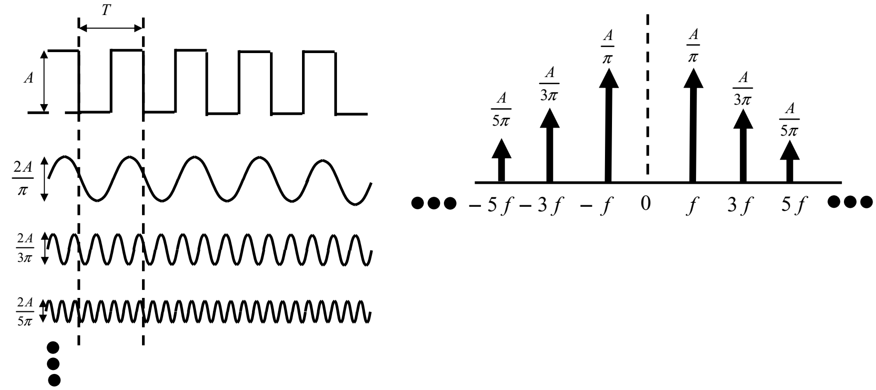


Figure 1.9. Spectral contents of a square waveform.

multiplied along with frequency. In the case of the half-wave rectifier waveform, the expansion is provided as follows:

$$s(t) = \frac{A}{\pi} + \frac{A}{2}\sin(\omega t) - \frac{2A}{\pi}\left[\frac{1}{3}\cos(2\omega t) + \frac{1}{15}\cos(4\omega t) + \ldots\right]$$

1.6.3 Fourier Transform

A more generalized case can be formulated by analyzing aperiodic signals, which can be formulated as signals limited in the time domain with a period of ∞. For a time-limited signal, we are interested in the equivalent frequency domain characteristics of the same signal, and the frequency spread is attributed to the bandwidth of the signal. When the signal is limited in time, the frequency spread increases and vice versa. For an aperiodic signal, the Fourier transform is given as follows:

$$S(f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}dt$$

and the existence of Fourier transform requires: (1) $s(t)$ to have a minimum number of maximum and minimum and a single value, (2) an finite number of discontinuities, and (3) absolute integrability $\int_{-\infty}^{\infty} |s(t)dt| < \infty$.

Some frequently encountered signals can be evaluated in terms of the Fourier transform at this stage. A rectangular pulse, strictly limited in time, would lead to a "sinc" shape in the frequency domain;, similarly, a "sinc" shape in the time domain represents a rectangular waveform in the frequency domain. Various properties of a Fourier transform are illustrated in Appendix A(2), Figure 1.9. It can also be shown that for pulse signal families, the product of signals duration and the bandwidth is a constant. The energy contained in the signal can be evaluated from either frequency or in time domain representations, $E_s = \int_{-\infty}^{\infty} |S(f)|^2 df = \int_{-\infty}^{\infty} |s(t)|^2 dt$ (also known as Rayleigh's energy theorem). The concept of bandwidth can be also explained from the frequency spread of the signals, and it is determined as a frequency range within which maximum signal energy is contained. Common methods of indicating bandwidth include when the signal power is 3 dB below its peak value and can be used as a performance metric for low-pass, band-pass systems. In the case of "sinc" type pulses, "null-to-null" spacing in the frequency domain contains maximum energy (almost 92%), and it can be used as a measure of bandwidth.

Fourier transformation of standard functions are shown in Appendix A(2).

1.6.4 Time and Frequency Domain Duality

Analysis of convolution occurs in the same manner as illustrated before, in the case of Laplace transformations. Fourier transforms can be very powerful in analyzing complicated functionalities in the time domain. In nonlinear circuits and systems, often a square and cubic law characteristics are common, and they can be obtained as a result of time domain multiplication. A time domain multiplication leads to a convolution in frequency domain. The relationship is shown as follows:

$$s_1(t)s_2(t) \Longleftrightarrow \int_{-\infty}^{\infty} S_1(u)S_2(f-u)du$$

where $S_1(f)$ and $S_2(f)$ denote the Fourier transformation of $s_1(t)$ and $s_2(t)$, respectively. To evaluate the convolution of two signals in the frequency domain (the same procedure is true for the time domain as well), we first flip the frequency axis of one signal while keeping the other intact. Then the flipped axis variable is moved toward the right, and as it moves, integration of the product is performed.

In a practical case, consider the blocker scenario in a WCDMA standard, where we are to evaluate the impact of second-order nonlinearity upon the SNR requirement at the demodulator. The frequency domain representation of the blockers, as well as the desired tone, is shown in Figure 1.10. Instead of performing this complicated multiplication, we simply obtain the frequency domain representation of the input signals to the amplifier (a modulated signal, a continuous wave blocker, an amplitude modulated blocker). Then we perform a flip in the frequency axis and slide the "flipped" terms to the right of the frequency axis (starting from an infinite offset from the center),

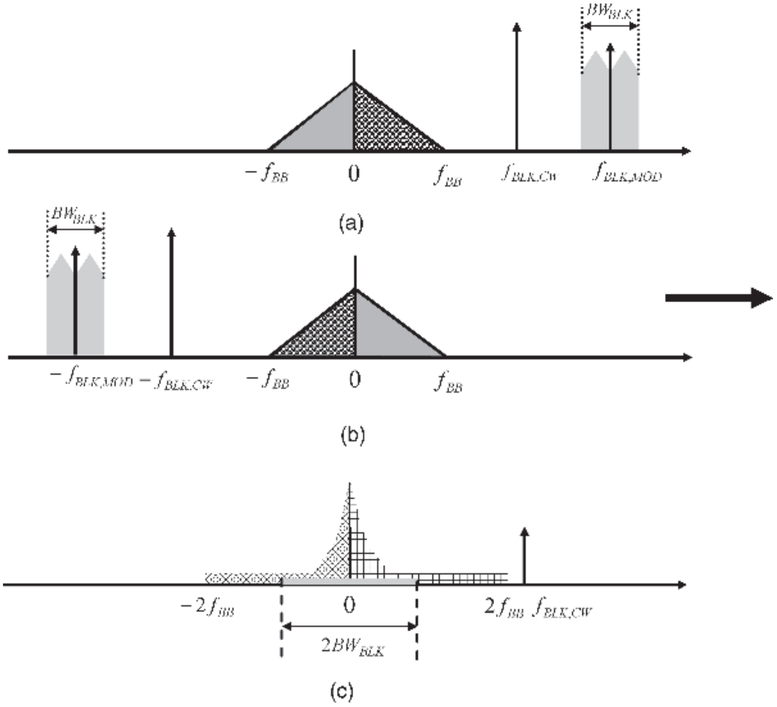


Figure 1.10. Analysis of modulated signal in the presence of blockers in nonlinear systems.

and we obtain the frequency domain representation of the nonlinear signal. As can be observed from the graphical illustration, the degradation is more when a modulated blocker with certain bandwidth is present, as opposed to a single tone. The convolution clearly illustrates the “spectral spreading” of the nonlinear terms in the desired bandwidth, and SNR degradation caused by the same. It can be easily observed that the evaluation in the time domain would be difficult, as the modulated blocker needs to be represented with time domain nonlinearity, leading to many terms, which are, hence, difficult to handle. At the same time, the graphical representation provides easily interpretable insights. For a cubic nonlinearity, convolution can be performed once more by flipping the frequency axis. Convolution-based evaluation becomes very effective in the case of multicarrier signals with uniform power distribution [a case for multicarrier orthogonal frequency devision multiplexing (OFDM) signals]. In this case, the frequency domain profile is rectangular and the time domain waveform is Gaussian in nature (because of presence of many carriers) with a high crest factor. Second-order nonlinearity leads to a cubic profile in the frequency domain, and cubic order nonlinearity leads to the parabolic shape of the frequency domain profile. When two sequences of bandwidth ω_1 and ω_2 are convolved with each other, the result would provide a frequency component up to $\omega_1 + \omega_2$. Hence, when a signal is convolved with itself, it “spreads” in frequency, which leads to SNR degradation throughout the bandwidth under consideration. In the case of unmodulated tone, no “spreading” is observed, providing a lower amount of SNR degradation.

1.6.5 Z Transform

In mixed signal systems, often Z transforms are useful in order to represent the operation of sampled signals. This is especially the case when the signal processing occurs at different time instants. It is true for a switched capacitor circuit, which stores charge at time τ_1 and transfers at time instant τ_2. These discrete time systems are well represented using Z transforms. Z transform is convenient in analyzing discrete time systems; e.g., digital filters and switched capacitor circuits. Z transforms are especially helpful in analyzing systems which are discrete in value (amplitude) and time (sampled). Analog/digital converters, especially sigma-delta type ones, are extensively analyzed using Z transforms. Analogous to the continuous time case, a–Z transform represents the frequency content and shaping function in the case of sampled data systems. They can be correlated to the continuous time counterparts with appropriate analog sampling frequencies. The transform can be represented by

$$F(Z) = \sum_{-\infty}^{\infty} f(n) z^{-n}$$

where $f(n)$ is a discrete sequence, which can assume any amplitude values (usually determined by the quantization of the system). Z is usually represented as $Z = e^{-j\Omega_n}$. Most of the properties of Z transforms are similar to the Laplace and Fourier transforms discussed before.

1.6.6 Circuit Dynamics

Although the transformations illustrated above provide computation flexibility and speed, we need to be careful not to forget the true nature of the circuit dynamics. For example, a large signal charging/discharging of capacitor and time domain dependence of current/voltage waveforms can be easily forgotten by representing the capacitor by an impedance $1/\omega C$. We can use the transformation of impedances in frequency domain in order to relate voltage and currents through them in the steady state. We must use these transforms to solve complicated networks, but under a given situation, they must clarify them from circuit dynamics, which are captured well in the charging/discharging behavior of components.

1.6.7 Frequency Domain and Time Domain Simulators

Circuits and systems can be analyzed in the time or frequency domain. Both approaches are common, and for a given circuit complexity (number of components, feedback loops, etc.), the time domain proves efficient when many harmonics are involved in the waveform (e.g., a square waveform). This may be the situation in digital circuits, where mixers are driven with a large signal square waveform shape. However, a difficulty, which is often faced with time domain simulators, is the presence of multiple time constants that are varying by orders of magnitude from one another. This is often the case for integrated transceivers where a low-frequency signal is upconverted to an RF signal and an RF signal is downconverted to a low frequency signal. In both cases, the system would be allowed to settle within the limits of the time constant of the circuit. Frequency domain simulators prove efficient in these cases, as knowing the nature of harmonic tones and their placement along the frequency axis would require few iterations for simulation convergence. Such techniques are well known as "harmonic balance," and the signals are treated as a combination of DC and a finite number of harmonics of the signal. It solves for magnitudes and phases of all spectral lines in the frequency domain simultaneously. The frequency domain current and voltages are adjusted w.r.t. their amplitude and phase characteristics until their sum equals the input current and voltages according to Kirchoff's laws. Currently, frequency domain analyses are becoming computation efficient by incorporating "envelope simulation" techniques, whereasthe time domain simulators tend to analyze various parts of the circuits w.r.t different timesteps, and by correlating the results w.r.t. sampling techniques.

1.6.8 Matrix Representation of Circuits

In this section, we will discuss the various forms of circuit representations using matrix form. Usually such representations are generic in nature, and they are applicable to transistors also. Modern technologies use many parameters to represent a transistor model, and they may appear as a complicated circuit themselves. Most commonly used under this category are the Z (impedance), Y (admittance), S (small signal), and H(hybrid) Matrices. These representations assume a "black-box" representation of the circuit element, assuming a two terminal model, in which one is input (terminal 1) and another is output (terminal 2). The matrix relationships are represented as follows.

$$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix} \times \begin{pmatrix} I_1 \\ I_2 \end{pmatrix}$$

$$\begin{pmatrix} I_1 \\ I_2 \end{pmatrix} = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix} \times \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$$

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \times \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

$$\begin{pmatrix} V_1 \\ I_2 \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \times \begin{pmatrix} I_1 \\ V_2 \end{pmatrix}$$

Each one of these representations is capable of describing the performance of two-port networks completely. Z and Y parameters are homogeneous, whereas H parameters combine voltage and currents. The fundamental difference in the case of S parameters is that they tend to use wave reflection methodology to represent a network. Each of these parameters can be interconverted as illustrated in Appendix A(3), and they can be used in appropriate scenarios. Although the other matrices do not require a standard reference impedance, any conversion from/to S parameters requires a reference impedance (50 Ω can be used assuming that characteristic impedance is not frequency dependent). All of these parameters capture the linear behavior of a network.

1.6.8.1 S Parameters. S parameters are widely used in microwave frequencies because of easy measurements (measurements are based on signal reflection and transmission), and convenience in using them for modeling purposes. In the power domain, they can be represented as follows:

$$\begin{pmatrix} |B_1|^2 \\ |B_2|^2 \end{pmatrix} = \begin{pmatrix} |S_{11}|^2 & |S_{12}|^2 \\ |S_{21}|^2 & |S_{22}|^2 \end{pmatrix} \times \begin{pmatrix} |A_1|^2 \\ |A_2|^2 \end{pmatrix}$$

where $|B_i|^2$ denotes the reflected power, $|A_i|^2$ denotes the incident power, $|S_{11}|^2$ denotes the reflected power at port 1, $|S_{21}|^2$ denotes the transmitted power from port 1 to 2, $|S_{12}|^2$ denotes reverse isolation, and $|S_{22}|^2$ denotes the output reflectance. To obtain the individual S parameters, the other terminal is terminated using characteristic impedance Z_0. To obtain S_{11}, port 2 is terminated. The accuracy of these parameters depends on the termination quality (how close Z_0 is to 50 Ω). The magnitudes of S_{11} and S_{22} are always less than 1, whereas S_{21} can have a magnitude greater than 1 (gain) and S_{12} is usually less than 1 (reverse isolation). S parameters also provide the phase shift information through the network, as they are essentially complex numbers. In the case of passive devices, reciprocity holds good, which leads to $S_{21} = S_{12}$. The magnitudes of the following cases can also be observed:

$S_{ii} = -1$: Amplitudes are inverted and reflected (0 Ω)

$S_{ii} = 0$: No reflections terminated at (50 Ω)

$S_{ii} = 1$: Voltage reflections without inversion

In the case of passive circuits, all values of S_{mn} are between -1 and 1, and in general, this implies that m is the output port and that n is the input port.

Impedance matching is an important consideration in designing integrated systems, and two critical examples occur in the termination of clock distribution networks in high-speed digital system and in the input of the low noise amplifier (LNA) in the case of wireless systems. Unterminated lines in digital systems result in reflections, which lead to significant distortions of the square waveform. A typical input matching of 15 dB can be adopted for narrowband wireless standards.

S parameters are important at a high frequency. At low frequencies, the voltage and current waveforms are the same at all points along the line. As frequency increases, line lengths are comparable with the wavelengths, and when the line is not terminated in Z_0, an entire signal is not absorbed by the load and reflected back to the source. In the case of a short-circuited termination, reflected and incident voltage waveforms would be equal in magnitude but oppositely phased. In the case of an open-circuit configuration, reflected and incident voltage waveforms are in phase, whereas the current waveforms are oppositely phased. In the case of a perfect termination, no standing wave is formed, and the energy flows from the source to the load in one direction. In the case of reflections, the ratio of maximum to minimum values of the RF envelope is termed the voltage standing wave ratio (VSWL). In the case of a perfect termination, VSWR = 1 and infinity for full reflection.

1.6.8.2 Smith Chart. A graphical representation of impedances can be made via a Smith chart, which can be effectively used in the case of designing matching networks. It starts with computing the reflection coefficient Γ, which is defined as $\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0}$, and for load impedance variation in the range of $0 < Z_L < \infty, -1 < \Gamma < +1$. Hence, in the Smith chart, we can plot a set of impedances conforming to certain constraints, and the Smith chart essentially would contain the impedance states, which determines the reflection coefficient. This transformation is graphically illustrated in Figure 1.11. A polar plot can be represented as well.

From the impedance transformation, the rightmost point in the Smith chart denotes infinite impedance, and the leftmost point (diametrically opposite) denotes zero impedance. The center point of the Smith chart denotes the characteristic impedance Z_0, and for a perfect match, the impedance would coincide with the center. In practical cases, however, the matching levels are determined by the distance of the impedance from the center. The upper half of the Smith chart contains inductive impedance states, and the lower half contains capacitive impedance states. With increasing frequency, the impedances always traces clockwise.

Impedance states provide important graphical information for circuit designers, and they are traditionally used in designing high-performance stand-alone RF circuits such as LNA, power amplifier (PA), as well as matching networks. For example, a set of impedances can be plotted in a Smith chart, which optimizes the noise figure of the

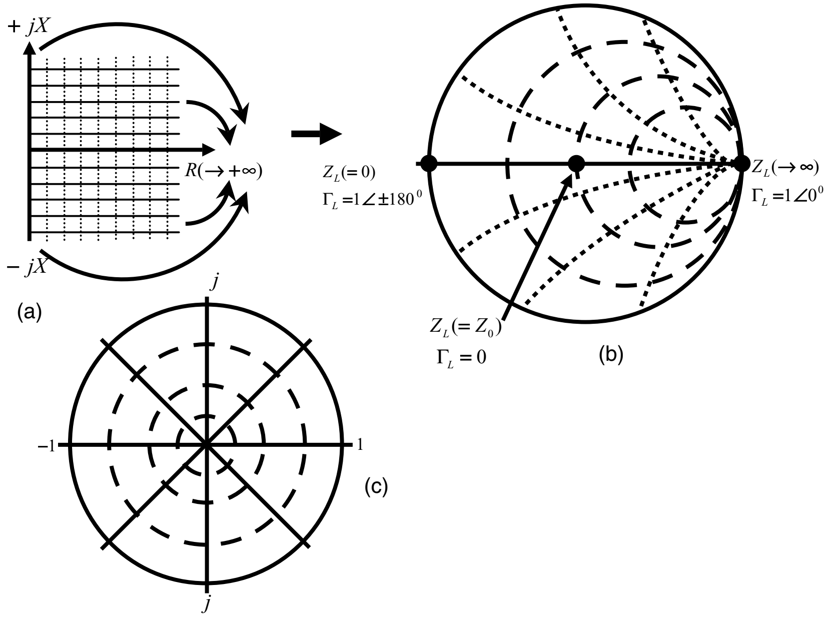


Figure 1.11. Rectangular, polar, and Smith chart representation of impedances.

amplifier, and simultaneously, the impedance states for maximum available gain can also be plotted. The locus of such impedances usually results in circles in the Smith chart (constant noise, gain, and stability circles can be plotted in the Smith chart as well) and the intersection of these circles would determine the optimum impedance states for circuit performance. In the case of power amplifiers, a load pull technique is commonly used to provide the designers realizable impedance states to maximize power transfer.

Impedance matching can also be graphically realized by Smith charts. The output impedance of a transistor is usually capacitive, and for maximum power transfer, we require an inductive impedance match. We first plot the device output impedance and its conjugate in the same Smith chart. Then we consider the 50 Ω load impedance and work backward. A series capacitor with the 50 Ω impedance provides an impedance state that is capacitive and is represented in the lower half of the Smith chart. This capacitive impedance is then considered in its admittance domain by flipping the Smith chart along its real axis. Finally, an inductive admittance takes it to the desired conjugate impedance. This simple graphical illustration suggests the use of an L-type matching network, which is commonly used at the output of the circuit. In practice, however, these components have finite Q, and the quality of matching is affected by the achievable Q from the components.

1.6.8.3 Practical Applications of S Parameters. S parameters are useful in device modeling. In the modeling step, the first part consists of an S parameter measurement, and then it converts to the appropriate parameters for better interpretation.

Critical performance parameters such as f_T and f_{MAX} can be easily interpreted from S parameters.

1.6.8.3.1 Amplifier Design. S parameters are commonly used in classic amplifier designs. A few illustrations include

$S'_{11} = S_{11} + \frac{S_{12}\times S_{21}\times\Gamma_L}{1-S_{22}\times\Gamma_L}$, input reflection coefficient with arbitrary load impedance Z_L

$S'_{22} = S_{22} + \frac{S_{12}\times S_{21}\times\Gamma_S}{1-S_{22}\times\Gamma_S}$, output reflection coefficient with arbitrary source impedance, Z_S

$A_v = \frac{S_{21}\times(1+\Gamma_S)}{(1-S_{22}\times\Gamma_L)\times(1+S'_{11})}$, voltage gain with arbitrary Z_S and Z_L

$K = \frac{1+|D|^2-|S_{11}|^2-|S_{22}|^2}{2\times|S_{12}\times S_{21}|}$, stability factor, where $D = S_{11}\times S_{22} - S_{12}\times S_{21}$

In the case of obtaining the unity current gain cutoff frequency f_T, we first obtain the S parameter measurement data and convert it to H parameters using the following relationship.

$$H_{21} = \frac{-2S_{21}}{(1-S_{11})\times(1+S_{22})+S_{12}\times S_{21}}$$

Usually transistors behave as a single-pole, low-pass filter, and f_T is determined by the frequency where $|H_{21}| = 1$.

1.6.8.3.2 Modeling of Passive Circuits. S parameters can be useful in constructing models of passive circuits. Since our target is to obtain the lumped element representation of such networks and to obtain the values of each of the lumped element components, a transformation to either Y or Z parameters should be applied. From these parameters, the individual lumped element components can be extracted. The networks are usually represented by a series and a parallel combination of lumped elements, which can be accurately extracted from the Y or Z parameters. This procedure is applicable in the case of package models and spiral inductor models.

Inductor models at high frequency can be obtained using S parameter measurement as well. In this case, the S parameters are first obtained using a two-port or one-port measurements. In RF circuits, often a differential inductor is employed for area efficiency, and a two port S parameter is most appropriate. From two-port S parameter measurement data, one-port S parameter data can be obtained (as shown in Appendix A) and subsequently converted to Z parameter data using

$$Z_{11,1p} = \frac{1+S_{11,1p}}{1-S_{11,1p}}\times Z_0$$

and the Q factor given as

$$Q = \frac{\mathrm{imag}(Z_{11,1p})}{\mathrm{real}(Z_{11,1p})}$$

w.r.t. Y parameters, it is given as

$$Q = \frac{-\text{imag}\left(Y_{11,1p}\right)}{\text{real}\left(Y_{11,1p}\right)}$$

A few commonly used networks, and their S, Y, and Z parameters, are provided in Appendix A(5). Many of these networks can be interpreted as a combination of T or π configurations.

1.7 BROADBAND, WIDEBAND, AND NARROWBAND SYSTEMS

The small signal response of any transistor-based circuit is usually broadband (determined by the f_T of the circuit). It is determined by the parasitic capacitance of the device as well as by the load capacitance. It forms a single-order pole at the output. Any signal processing operation in analog requires power, and the power consumption is proportional to the frequency of operation and square of the bandwidth under consideration. A broadband system is capable of operating from DC to its bandwidth. Wideband systems operate over bandwidths greater than the center frequency BW $\sim 1.5f_c$. Narrowband systems usually operate with BW $\sim 0.2f_c$, a commonly used scenario in wireless systems. Operating over narrow bandwidths reduces the power consumption in the front end. This result is quite intuitive, as we consume lower power and process smaller bandwidth signals. The characteristics of a semiconductor active device are usually broadband, limited by the parasitic device capacitance. To use such circuits, a narrowband signal processing element would need to use frequency-selective load impedances. Such components can be very easily implemented by a parallel combination of two frequency-dependent reactances, one of which grows with frequency (inductor) and another decreases with frequency (capacitor). Thus, we obtain a high impedance at a certain frequency (resonance) and a low impedance away from it. Broadband and wideband systems pose a severe group delay restriction on the circuits that perform signal processing, and they require a wideband antenna, which is challenging to implement. Narrowband systems can be implemented by selective peaking of wideband device characteristics using frequency-selective impedances.

1.7.1 LC Tank as a Narrowband Element

Each component in an LC resonant circuit can be represented in a series/parallel combination of the reactance and Q factor as illustrated in Figure 1.12. In a series representation, the Q factor is given by $Q_s = \omega L/r$ (reactance divided by resistance), where r is the series resistance. The parallel equivalent of this circuit consists of the same inductance, and a parallel resistance denoted by $R = Q_s^2 r$, and the Q factor is given by $Q_P = R/\omega L$ (resistance divided by reactance). Of course, we cannot obtain a different Q factor just by merely changing a series or a parallel combination of a component representation. The same is true for capacitors, where $Q_s = 1/\omega Cr$, and a parallel combination would provide $Q_P = \omega CR$, where $R = Q_s^2 r$.

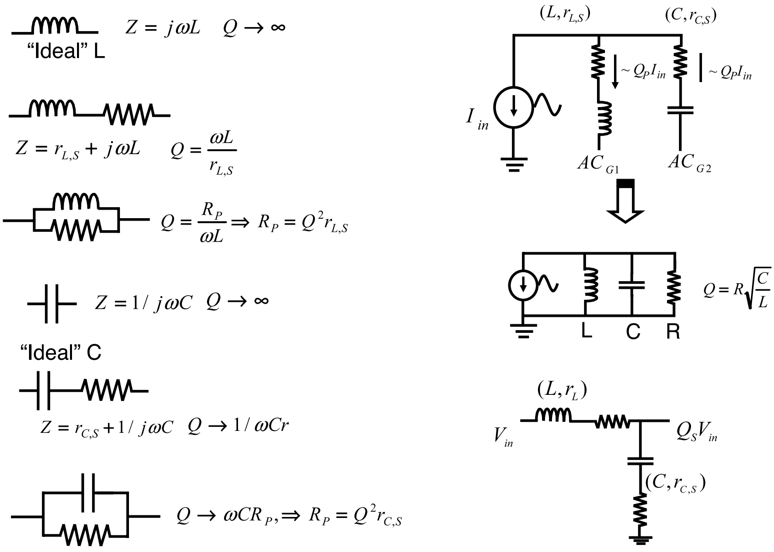

$Z = j\omega L \quad Q \to \infty$
"Ideal" L
$Z = r_{L,S} + j\omega L \quad Q = \frac{\omega L}{r_{L,S}}$
$Q = \frac{R_P}{\omega L} \Rightarrow R_P = Q^2 r_{L,S}$
$Z = 1/j\omega C \quad Q \to \infty$
"Ideal" C
$Z = r_{C,S} + 1/j\omega C \quad Q \to 1/\omega Cr$
$Q \to \omega CR_P, \Rightarrow R_P = Q^2 r_{C,S}$
$(L, r_{L,S})$
$(C, r_{C,S})$
$\sim Q_P I_{in}$
$\sim Q_P I_{in}$
I_{in}
AC_{G1}
AC_{G2}
L
C
R
$Q = R\sqrt{\frac{C}{L}}$
(L, r_L)
V_{in}
$Q_S V_{in}$
$(C, r_{C,S})$

Figure 1.12. Series and parallel representation of LC circuits.

The above analysis can then be extended to *LC* resonant tanks. Each component can be transformed to its parallel equivalent, and at resonance, the inductive impedance will cancel the capacitive impedance. The resistive part from each of the components is connected in parallel, and the *Q* factor at resonance is determined by the ratio of the parallel resistance to the individual reactance.This implies that the *Q* factor of an *LC* resonant tank is obtained by the parallel combination of the individual *Q* factors. Usually on-chip capacitors provide $Q > 25$, and the overall *Q* is dominated by the inductor *Q*. Hence, we emphasize the unloaded *Q* factor of inductors in conjunction with *LC* resonant circuits. The circuit performance always depends on the "loaded *Q*" of the *LC* tank, which is governed by the parallel combination of the individual component *Q* factors.

1.7.2 LC Tank at Resonance

At resonance, the individual currents through the resonating components are increased by *Q* times the input current. However, this current is circulated through the inductor and the capacitors, and it does not flow anywhere else. The sum of the branch currents has to be the same as the input current to validate Kirchoff's laws.

Let us consider a transconductor stage loaded with an *LC* tank in the limits of a small signal operation. As shown in Figure 1.13, it can be observed that the voltage swing across the tank at resonance is given by (assuming high output impedance of the cascode pair)

$$V_{\text{TANK}} = g_m V_{in}(\omega L)Q$$

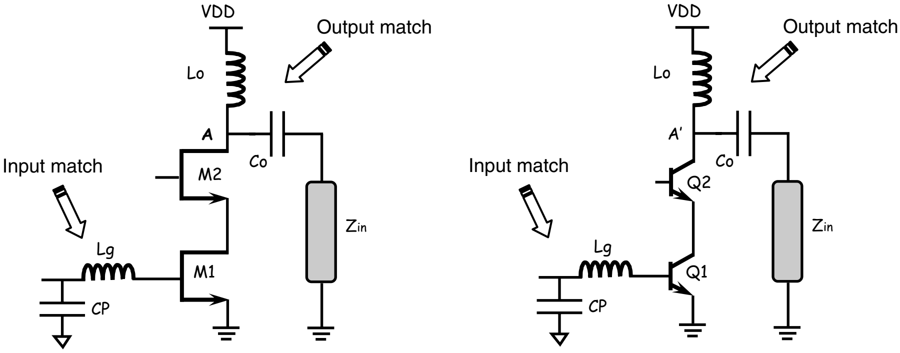


Figure 1.13. Inductively loaded MOS and bipolar stages.

1.7.3 Q Factor, Power, and Area Metrics

This result, is very important, as it implies that a Q factor of 3 can reduce the power dissipation by a factor of 3. We also assume here that the amplifier is operating at open loop. Q is the loaded quality factor associated with the tank ($= Q_L || Q_C || Q_{Cas}$). Assuming that we use the highest quality factor capacitors in tank load, and a high Q of the output parasitic capacitance from the cascode device, Q_L would dominate in the determination of loaded tank Q. At the same time, it can be observed that $g_m = \frac{I_C}{V_t}$ in the case of bipolar and $g_m = \sqrt{2\beta I_d}$ in the case of long-channel MOS devices. Thus, at a specific operating frequency (ω) and an operating input signal condition (V_{in}), the transconductance, hence, the DC current, and the power can be reduced proportionally with an increase in the LQ product. Hence, one can increase L or increase Q, or obtain an intermediate optimization stage. A higher value of L would lead to an increase in inductor area (given by the number of turns and the outer diameter), and to a reduction in capacitor area for operating at the same resonating frequency. The opposite would happen for a reduced L, and the capacitor area consumed would be more. If the Q is increased, then the circuit would provide a narrowband frequency response. Hence, two types of optimization domains would exist:

1. High L, low Q, low C, higher bandwidth, lower power
2. Low L, High Q, high C, lower bandwidth, lower power

1.7.4 Silicon-Specific Considerations

It would also be observed that as the Q is increased for the amplifier stages, it leads to increasingly higher signal swing and narrower bandwidths, which then leads to saturation of the subsequent stages and more susceptibility to process variation. As mentioned, the passive components do not consume any voltage headroom, and it is suitable to obtain signal swings beyond supply rails (assuming that the reliability considerations of the devices are satisfied for large swings). Thus, it leads to high dynamic range systems in the analog/RF domain. It can also be observed from the above argument that high Q is not always desirable. In fact, one can design circuits for low-power applications by using a Q factor of 4–5, which is fairly reasonable for silicon-based process technologies. Hence, ideally one does not have to look for technologies where a Q of 1000 is to be achieved. This may be desired in the case of oscillators for obtaining a very good phase noise at lower current consumption. However, oscillators always operate with a PLL, but the amplifiers usually operate in open loop. Hence, very high Q is attractive for VCOs, but not desirable for amplifiers. At the same time, it can be observed seen that the impedance at resonance is governed by the ratio of $(Q_L\sqrt{L/C})$. Various types of passive components are used in integrated systems, and we would explain the passive components a little later. While being useful for the power consumption perspective, they tend to consume large area on chip, and they generate electromagnetic cross-talk with other components present in the same substrate. It must be noted that passive components do not provide power gain (it would be wonderful if they did!). They can provide voltage gain or current gain.

1.7.5 Time Domain Behavior

The above illustration provided frequency domain representation of L-C resonators. We have observed the frequency-dependent nature of individual impedance components and have obtained a frequency where they cancel each other to provide high impedance. The transient domain analogy is also possible, and one can apply a current step at the input of the LC tank, where the voltage waveform would “ring” for Q cycles before reaching its steady state value. The frequency of oscillation determines the resonating frequency. It is caused by the fact that voltage across a capacitor is obtained by “integrating” the current over time, whereas the same for the inductor is obtained by “differentiating,” thus leading to the formulation of a second-order differential equation.

1.7.6 Series/Parallel Resonance

The above illustrations are true for circuits at resonance, which can be a parallel resonance, or a series resonance, based on the circuit configurations and component arrangements, as illustrated in Figure 1.14. Series resonance can be used at the input of the LNA circuit to provide voltage gain, thereby improving the noise figure. Parallel resonance provides current gain at resonance, and this is true for an LC resonant tank or a matching network. At resonance, multiple reactance elements come in parallel to one another; the current through each element is multiplied Q times at resonance. In an L matched section, this current flows through the branch consisting of 50 Ω of impedance and responsible for the voltage swing at the output. As illustrated in Figure 1.14, both common-mode and differential-mode impedances must be taken into consideration.

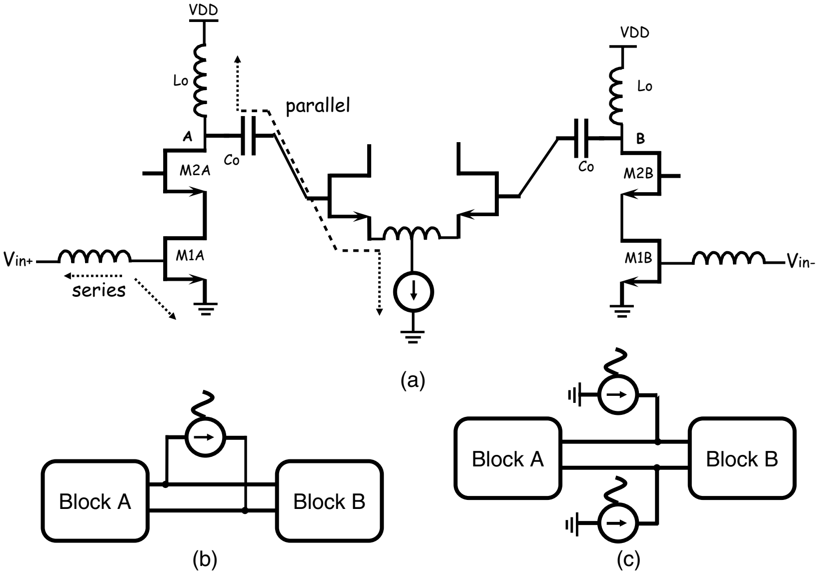


Figure 1.14. Two building blocks w.r.t. interfacing impedance and Q.

1.8 SEMICONDUCTOR TECHNOLOGY AND DEVICES

Because of the high mobility of III–V transistors, they were preferred in the early years for developing integrated radios. Most III–V transistors also provide a direct alignment of valence and conduction bands, leading to light-emitting behavior from these materials. Hence, they gained significant popularity: (1) high cutoff frequency to perform frequency translation and other high-speed operations, and (2) light emission capability, so that they can be integrated as a part of high-speed digital signal processing systems. However, cost was a major factor for these devices. Large wafers had manufacturing issues, whereas much complex functionality could not be easily integrated. At the same time, in many cases, the differences between the electron and the hole mobility were significant, such that feasibility of complementary circuit topologies was difficult. However, at the time of their popularity, radios used to be a combination of separate functional blocks, and were not viewed as complicated integrated systems. Many of these devices are still used in the high-speed industry, such as a defense electronics and RADARs, for their superior performance. However, in many of these cases, the quality of substrate was also a winning factor, as high *Q* and low parasitic passives, such as inductors, capacitors, and resistors could be easily realized and were accurate. Typical examples include GaAs MESFET, GaAs bipolar, InP bipolar, GaAs pHEMT, and so on. Each of these devices is different from another in terms of its physical construction and operation according to band-gap theory.

1.8.1 Silicon-Based Processes

Developments on silicon-based platforms started with the invention of bipolar transistors (1947). During the 1980's, various fabrication difficulties related to CMOS devices were solved and CMOS based digital circuit techniques became more and more popular. During the late 80's Germanium doped graded base profile bipolar transistors could achieve both the speed (f_T) and the RF performance (f_{MAX}) comparable with the III–V semiconductors. As the CMOS technology nodes were scaled successively, the high-frequency handling capability increased, and the minimum feature size was reduced. This improvement led to a lower voltage, lower geometry device. Although it became popular in the digital domain, there were several factors to consider:

1. Fineline CMOS is costly, because of the maskset and lithographic precision.
2. Silicon substrate provides moderate to low Q passives.
3. Process controllability is poor in fineline CMOS, which leads to worse component matching.
4. There may be incremental area advantages for high-density I/O circuits (circuits may be pad limited).

Although digital circuits enjoy scaling in terms of power and area (although leakage is a significant issue for the advanced nodes), RF/analog circuits may suffer because of lower breakdown voltages, poor matching, and lower drive strengths.

The proper choice of semiconductor platform is a critical decision in the implementation of integrated communication microsystems. The basic elements of any high-frequency communication circuit are transistors and high Q, high-density, low-parasitic passives. For the analog/RF building blocks, the transistor cut-off frequency f_T and the maximum oscillation frequency f_{MAX} are the key performance metrics, which usually increase with each technology node. This implies that one can realize progressively higher frequency circuits and systems using advanced silicon-based technologies. At the same time, the power consumption at a specific frequency of operation reduces with increased cutoff and oscillation frequencies. Advanced MOS transistors can be operated in a subthreshold region, which has been proven to be effective in terms of their low-power consumption. Advanced CMOS technologies tend to demonstrate subthreshold cutoff frequencies in the GHz range. To realize higher transconductance, the device sizes need to be significantly large, which leads to parasitic loading. The weak inversion region also has a worse noise performance when compared with the strong inversion region. As a MOSFET is driven into a subthreshold region from strong inversion (by reducing v_{GS}), NF_{min} increases sharply and then becomes saturated at a higher value [23].

1.8.2 Unity Current and Power Gain

The maximum device cutoff frequency (f_T), maximum oscillation frequency (f_{MAX}), broadband noise factor (NF_{min}), and flicker noise ($1/f$) profile become the determining

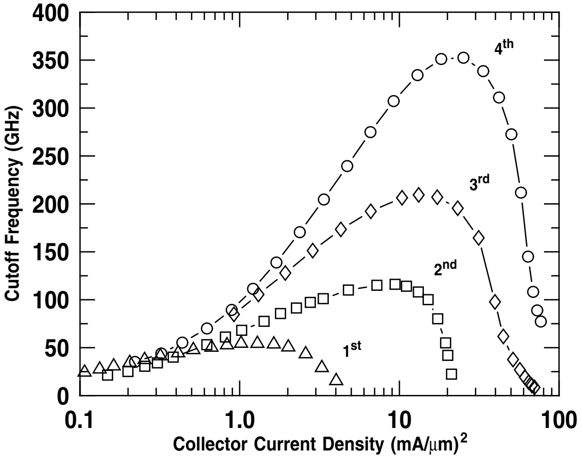


Figure 1.15. Cutoff frequency versus current density in bipolar.

factors for RF designs at nanometer geometries. These parameters play an important role, irrespective of their device technology (bipolar or MOS).

Figures 1.15 [24] illustrates the scaling impact on $f_T$across various generations of SiGe HBT transistors. As f_T increases, it is observed that a transconductor would require progressively lower power consumption at a certain frequency of operation (f_T GHz). In other words, increasing f_T enables the feasibility of RF designs at progressively higher frequency regime.

Figure 1.16 illustrates the scalability impacts on the cutoff frequency of MOS, across two technology nodes, 130 nm and 90 nm [25,26]. The cutoff frequency is given by

$$f_t = \frac{g_m}{2\pi(C_{gs} + C_{gd} + C_{gb})} \tag{1.1}$$

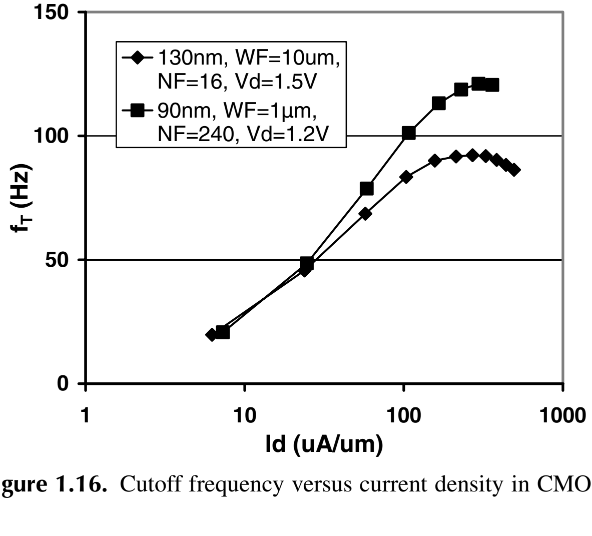


Figure 1.16. Cutoff frequency versus current density in CMOS.

where g_m is the transconductance and C_{gs}, C_{gd}, C_{gb} indicate the gate-source, gate-drain, and gate-bulk capacitances associated with the respective terminals. Using short channel approximations for g_m, and expressing the capacitances in terms of the area parameter,

$$\begin{aligned} f_t &= \frac{\mu_n C_{ox} W E_{sat}}{4\pi(C_{gs}W + C_{gd}W + C_{gb}WL)} \\ &= \frac{\mu_n C_{ox} E_{sat}}{4\pi(C_{gs} + C_{gd} + C_{gb}L)} \end{aligned} \tag{1.2}$$

In conjunction with f_T, f_{MAX} determines the frequency of unity power gain and is given by

$$f_{\max} = \frac{f_t}{\sqrt{8\pi C_{gd} R_G f_t + 4g_{ds}(R_G + R_s)}} \tag{1.3}$$

It can be observed that, although f_T is a relatively straightforward expression in terms of forward current gain, f_{MAX} is a complicated function of device geometry and layout. The gate resistance plays a significant role in determining f_{MAX}, along with the substrate and gate resistances associated with the device.

In addition to their high-frequency behavior, two device noise mechanisms such as broadband noise (thermal noise) and flicker noise become important for nanometer MOSFETs. The broadband noise factor of MOSFET is given by

$$NF_{\min} = 1 + B\frac{f}{f_t}\sqrt{g_m(R_s + R_G)} \tag{1.4}$$

which implies that the minimum noise that can be obtained from these devices is dependent on the transconductance g_m, the gate resistance R_G, and the substrate resistance R_s. It can be observed that the improvement of broadband noise can be achieved by optimized layout, which also helps improve f_{MAX}. The substrate resistance directly impacts the noise factor, and hence, a reduction in the substrate resistivity would lead to minimization of the minimum noise factor.

1.8.3 Noise

The flicker noise contributed by the MOS transistors is given by

$$\overline{i_n^2} = \frac{K}{f} \cdot \frac{g_m^2}{WLC_{ox}^2} \cdot \Delta f \approx \frac{K}{f} \cdot \omega_T^2 \cdot A \cdot \Delta f \tag{1.5}$$

The flicker noise in MOSFET is usually much worse compared with bipolar transistors, as the fluctuation of the carriers in the channel occurs in the presence of traps in the oxide. For a given transconductance, a larger gate area and thicker oxide reduces the contribution of flicker noise. Since an increased area implies the loading for RF/analog circuits, careful design optimization needs to be performed.

Broadband noise is important for circuits such as low noise amplifiers, whereas flicker noise impacts the low-frequency circuits such as differential amplifiers and OP-Amps, in addition to the flicker noise upconversion in the cases of VCOs, frequency dividers, mixers, and other nonlinear frequency conversion circuits. The design in the RF/analog regime needs to be optimized in terms of power and area; hence, these formulations need to be used in specific cases, depending on the circuit under applications.

1.8.4 Bipolar vs. MOS

It is important to understand the differences between two devices in order to use them optimally in a transceiver architecture. Bipolars are vertical devices with regard to the current flow, whereas MOS current flow is lateral in nature and most of the action (electron transport) happens at the surface. From a circuit perspective, transconductance of a bipolar transistor is dependent only on the current it is biased at, and is not scaleable, whereas in the MOS device, the transconductance is a function of both the bias current and the device size. The transconductance is proportional to the bias current for a bipolar device (whereas the square root in MOS), which leads to the fact that bipolars provide superior transconductance performance. However, the collector and emitter terminals are asymmetric in nature in terms of electrical performances, whereas drain and source symmetrical in nature in a MOS device. Hence, MOS transistors can be used as excellent switches and are popular in digital circuits, sampling switches, and passive mixers. At deep submicron technology, bipolars tend to have superior output impedance performance compared with MOS. Area wise, construction of MOS device usually takes 15% more area compared to the bipolar device of the same transconductance performance. Being a surface device, MOS is susceptible to electron interaction with the trap states in the gate oxide, which contributes more flicker noise than a bipolar device.

A BiCMOS technology is optimum for integrated radios. Superior bipolar devices can be used to develop better low noise amplifiers, low $1/f$ noise VCO cores, and baseband amplifier stages, whereas MOS can be used for superior switching performance. In terms of fundamental device operations, bipolar is a minority carrier device, whereas MOS is a majority carrier device. Different parts of bipolar transistor characteristics are well modeled using exponential characteristics, whereas MOS is mostly a square law device and empirical modeling is used. However, at deep submicron MOS, this differs significantly. Bipolar modeling is complicated by the fact that the collector and emitter terminals are asymmetrical in nature, and the current distribution in the collector is difficult to model in advanced geometries. The difficulty in MOS modeling originates in order to construct a continuous model across all regions of operation. In their inherent nature, the operating regions of MOS devices are sometimes difficult to represent using a single mathematical equation, and usually, numerical fitting techniques are used. While modeling a device, accuracy over a wide operating range is often obtained using multiple parameters, which may imply intensive analysis time for the circuit simulators and so on. Although an accurate model can be obtained using fewer parameters for a fixed device geometry (and a large device can be an array of small devices) to obtain faster simulation of circuits, scalability is often, preferred by the designers. Moreover, in advanced CMOS technologies, all functional blocks can be integrated into one substrate, which leads to the popularity of CMOS technology nodes.

1.8.5 Device Characteristics

Without moving to much detail in modeling aspects, we intend to provide readers with basic intuition of the fundamental device parameters to develop circuits and systems. At the same time, a hand analysis is almost impossible using multiple model parameters. Key performance metrics are discussed in the following subsections.

1.8.5.1 DC Characteristics. I_D versus V_{GS} for MOS, I_C versus V_{BE} for bipolar MOS transistors operating in linear region $V_{DS} < V_{GS} - V_t$, for the triode region with a voltage variable resistor of $r_{ON} = \left[\mu C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_t)\right]^{-1}$

$$I_D = \frac{\mu C_{ox}}{2}\left(\frac{W}{L}\right)\left[2(V_{GS} - V_t)V_{DS} - V_{DS}{}^2\right]$$

MOS transistor in saturation region.
$V_{DS} > V_{GS} - V_t$ for saturation region, which leads to the square law characteristics,

$$I_D = \frac{\mu C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_t)^2$$

The threshold voltage V_t is given by, $V_t = V_{t0} + \gamma\left[\sqrt{2|\phi_F| + V_{SB}} - \sqrt{2|\phi_F|}\right]$.

In the case of bipolars, the corresponding relationships are given as follows:

$$I_C = I_S e^{(V_{BE}/V_T)}$$

1.8.5.2 Output Impedance. The real part of the output impedance is governed by the output DC characteristics, and the imaginary part is governed by a combination of one or more capacitances associated with the device. At RF frequencies, the capacitance usually plays a dominating role in determining the output impedance. At low frequencies and DC, the real part becomes important for construction of current sources and so on. Both the magnitude and the Q factor of this impedance are important for circuit design.

In the case of bipolar transistors, $r_O = \frac{V_A}{I_C}$, where V_A is the "Early" voltage and I_C determines the bias current. In the case of MOS, this is given as $r_O = \frac{1}{\lambda I_D}$, where λ denotes the channel length modulation parameter.

1.8.5.3 Capacitive Elements. Two types of capacitors are associated with transistors: (1) The geometry-dependent capacitor and (2) the bias-dependent capacitor. These capacitors are associated with MOS and bipolar.

In the case of MOS transistors, the source-drain regions form a diode structure with the substrate, and the capacitance is voltage dependent, which is given by

$$C_{SB} = \frac{C_{SB0}}{\sqrt{1 + V_{SB}/\phi_0}}, \quad \text{and} \quad C_{DB} = \frac{C_{DB0}}{\sqrt{1 + V_{DB}/\phi_0}}$$

The input capacitance referred to as the gate terminal is geometry dependent, and it is given by $C_G = C_{OX}(W.L)$. This capacitance is essentially divided between the drain and source terminals, depending on the geometrical shape of the formed channel, which differs from a triode region to a saturation region.

Triode/Linear region:

$$C_{GS} = \frac{1}{2}C_{OX}(W.L), \; C_{GD} = \frac{1}{2}C_{OX}(W.L)$$

Saturation region:

$$C_{GS} = \frac{2}{3}C_{OX}(W.L), \; C_{GD} = \frac{1}{3}C_{OX}(W.L)$$

In the case of bipolar transistors, the $C - B$, and $C - S$ junction capacitances are given by

$$C_{EB} = \frac{C_{EB0}}{(1 + V_{EB}/\phi_0)^n}, \; C_{CB} = \frac{C_{CB0}}{(1 + V_{CB}/\phi_0)^n}$$

The input capacitance C_B is bias dependent and given as $C_B = \tau g_m$, with $C_{BE} = C_B + C_{jE}$.

Voltage variable capacitors provide distortions to the signal waveforms. They also provide AM–PM conversion in large signal swings.

1.8.5.4 Device Noise.

Device noise determines the fundamental limits to available signal-to-noise ratio in any circuit. Mainly three types of noise occur in devices:

Thermal noise: This is associated with random flow of electrons, and not associated with any bias current; it is present in all devices. The spectral density id given by

$$\langle i_n^2 \rangle = \frac{4kT}{R}\Delta f$$

Shot noise: This is associated with DC current flow, and it is always associated with a junction and independent of frequency. The spectral density is given by

$$\langle i_n^2 \rangle = 2qI\Delta f$$

Flicker noise: This is associated with DC current and present in all active devices; the spectral density is given by

$$\langle i_n^2 \rangle = K\frac{I^a}{f^b}\Delta f$$

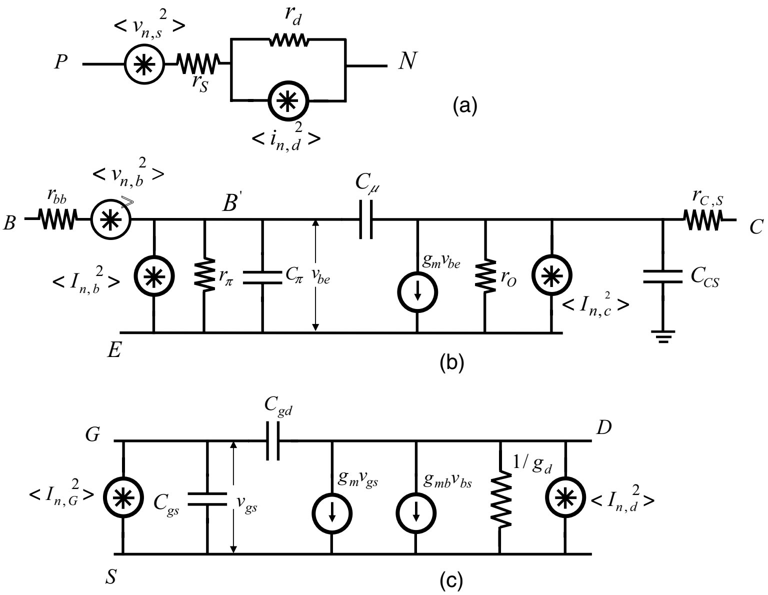


Figure 1.17. Noise models for diode, bipolar, and MOS transistors.

These equations can be applied to diode, MOS, and bipolar devices, respectively, and the noise model can be obtained as illustrated in Figure 1.17. Noise can be referred to the output as well as to the input, and both methods can be used in the circuit design process. Both are related to each other by the ratio of transconductances. Flicker noise is an important consideration in MOS transistors, and it is physically related to the number of surface states and to the change of threshold voltage from the gate oxide capacitance C_{OX}. The input-referred flicker noise spectral density is represented by

$$\langle v_{f,G}^2 \rangle = \frac{K_f}{WLC_{OX}f}\Delta f$$

which is independent of bias current. However, when referred by drain, this is given by

$$\langle i_{f,D}^2 \rangle = g_m^2 \langle v_{f,G}^2 \rangle = \frac{K_{f,D} I_D}{L^2 f}\Delta f$$

which is dependent on the bias current and the channel length. This result bears very important conclusions in the case of circuit designs: (1) Circuits, which do not consume any current (such as passive mixers, etc.), are inherently quiet in terms of flicker noise performance. (2) The output noise current is inversely proportional to the square of channel length, and migration to smaller channel lengths would contribute more flicker noise.

On the other hand, the spectral density for drain-referred thermal noise is given by $\langle i_{n,D}^2 \rangle = 4kT\beta g_m \Delta f$, where β varies between 2/3 and 2 from the saturation region to the

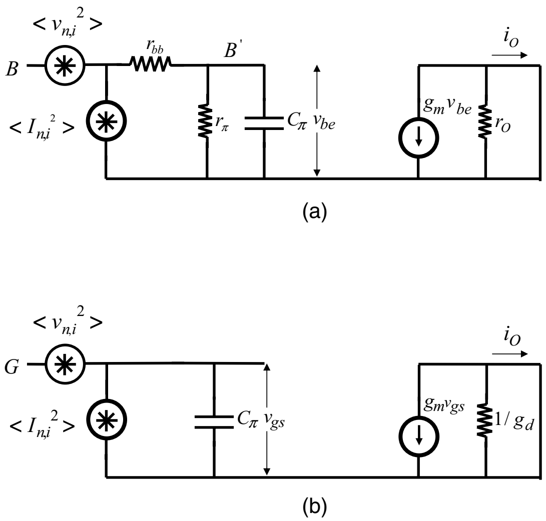


Figure 1.18. Equivalent noise models for circuit analysis.

subthreshold. Drain current caused by the input signal is given by $i_D^2 = g_m^2 V_{gs}^2$; thus, increasing g_m leads to a better SNR w.r.t thermal noise.

Gate leakage in MOS leads to shot noise, which can be represented as $\langle i_{n,G}^2 \rangle = 2qI_G\Delta f$. This is uncorrelated from the drain noise terms described above. Bipolar transistors can be analyzed in the same manner, starting with fundamental noise equations.

Using these noise models, important insight can be developed to construct optimum impedance for low noise circuits. Representation of bipolar and MOS circuits are illustrated in Figure 1.18. Noise models can be viewed as various noise sources in conjunction with the small signal models. In the case of bipolars, both the input referred voltage and the current noise terms are present. The voltage noise becomes dominant when being driven from a low impedance source, and the current noise becomes dominant when being driven from a high impedance source. Thus, an optimum noise figure is obtained in between the two impedances. In the case of MOS stages, the input noise current source is dominant, which leads to a high driving impedance for optimum noise performance.

Noise has an important implication in designing circuits; therefore, we will discuss a few typical cases. Often in communication systems, signals are sampled w.r.t. a clock waveform. To prevent aliasing, the clock frequency is chosen to be an integral multiple of the message signal. Although this process ensures signal reconstruction, it also downconverts noise from various clock harmonics, and places them in a band of interest. This action is typical of a sampling switch. In a simple $R - C$ stage, when resistor increases, the magnitude of the noise associated with it increases, but the bandwidth reduces. Similarly, when resistance is reduced, thermal noise reduces, but bandwidth is increased. The integrated noise is the same in both cases

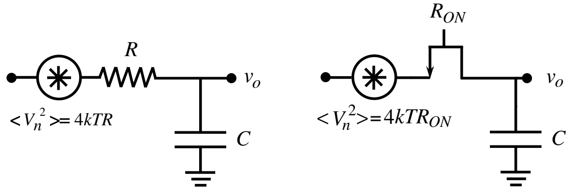


Figure 1.19. Noise in R–C stages.

and is given by

$$\langle V_{o,n}^2 \rangle = \int_0^\infty \frac{4kTR}{[1+(2\pi fRC)^2]} df = \frac{kT}{C}$$

However, it is also true for a bandlimited system. In a circuit, where multiple poles and zeros are present and the frequency response extends to infinity (a pole followed by a zero), the integration bandwidth does matter in the integrated noise consideration. Noise is contributed by the resistance only, but the capacitance comes into picture because of the band limitation of the white noise. This process is illustrated in Figure 1.19.

As this is dependent on the capacitor size, the output noise is invariant even when a MOS switch is present instead of a resistor. Thus, to suppress this noise, a large capacitor is to be used. To account for integrated noise and thermal noise spectral density, periodic steady-state noise and small-signal noise analysis can be used using circuit simulators.

1.8.5.5 Breakdown Voltage. From a device design perspective, device speed and breakdown voltages are important. They are related by the fundamental relationship, known as Johnson's limit $BV \times f_T = K$. Hence, the speed of a device cannot be increased arbitrarily without compromising the breakdown voltage. Hence, breakdown plays a critical role in submicron geometries, when designed for a high RF frequency.

Breakdown can fundamentally occur because of (1) application of an electric field across a semiconductor junction, which is more than the maximum electric field to be sustained at that junction, and (2) damage to the oxide caused by electrons moving in high velocity (caused by a high-input electric field). In the first case, the device can still be recoverable, but in the second case, it is permanent damage to the device.

In a bipolar device, two critical voltages associated with breakdown are V_{CBO} and V_{CEO}, which are called the "collector to base breakdown with emitter open" and "collector to emitter breakdown with base open," respectively. They are related to each other by

$$BV_{CEO} = \frac{BV_{CBO}}{\sqrt[n]{\beta_F}}$$

In a MOS transistor, similar quantities are referred to as "gate to drain breakdown" and "drain to source breakdown." The gate breakdown voltage is significantly lower compared with the drain-to-source breakdown voltage.

In large signal circuits and systems, breakdown is a critical consideration, and it depends on operating temperature as well. To obtain a higher breakdown voltage, thick gate oxide transistors can be used in MOS technologies with the compromise of lower transconductance. In a bipolar transistor, collector current has a positive temperature coefficient, and circuits are susceptible to thermal runaway. This is a major consideration in PA design, and resistors in the emitter are used to counter this effect, as RF power devices comprises arrays of small bipolar junction transistor (BJTs). On the other hand, the drain current for a MOS device has a negative temperature coefficient, which prevents thermal runaway, and multiple MOS transistors can be connected in parallel without ballasting requirements. Because of the inherent switching functionality of MOS devices, they are popular for switching the mode operation of PAs.

1.8.5.6 Technology Scaling. As semiconductor technologies are scaled, the physical device dimensions (W,L,t_{OX}) get reduced by the scaling factor α, supply voltage also drops by the same factor, delay of digital gates reduces by α (delay $=CV/I$), and so on. However, the wiring delay remains the same (the resistance of wires increases and the capacitors reduces). Hence, at scaled geometries, interconnecting delays play a critical role in gate delays. Power dissipation in digital gates reduces, and so is the power-delay product, which contributes positively to the performance of digital gates. In terms of RF/analog performance, $f_T$andf_{MAX} play a significant role in determining power consumption at a specific center frequency.

The impacts of scaling can be categorized in terms of analog and digital circuits as well. In a digital circuit, the number of gates would be an important parameter, and a technology scaling would significantly reduce the area of the digital part of the chip. Although there are several advantages to scaled CMOS geometries, they tend to provide significant leakage currents.

At deep submicron CMOS technology nodes, gate leakage contributes to a significant fraction of the power consumption of large digital chips. In RF circuits, it may lead to noise figure degradation. Leakage mechanisms can be categorized in the following major categories: (1) drain-induced barrier lowering (DIBL), (2) gate-induced drain leakage, and (3) hot carrier effect. Figure 1.20 illustrates the various leakage mechanisms in deep submicron MOS devices. In addition to these effects, process variation plays a critical role in analog circuits. As the lithographic geometries are extremely small, a little variation in geometry may lead a to a large variation of threshold voltage, transconductance, and so on. Broadband noise performance improves with scaling, but the flicker noise performance usually gets worse. Hence, the impact in terms of a continuous-time signal processing block (analog/RF circuit) includes (1) reduced supply voltage and dynamic range limitation; (2) increased component variation, which leads to the need of calibration circuits; (3) increased leakage; and (4) increased mask cost. However, sometimes the use of analog/RF blocks along with digital in the same substrate is encouraged to obtain a single die solution (system on a chip).

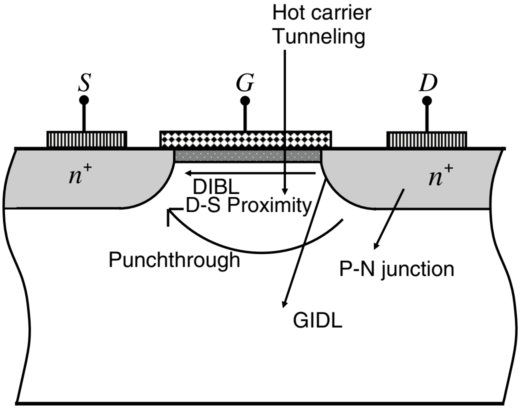


Figure 1.20. Leakage mechanisms in deep submicron CMOS technologies.

With this basic introduction, we will now consider a few testbenches in order to evaluate semiconductor technology platforms. These testbenches are generic in nature and routinely used by circuit designers. They can be used toward any technology platform at hand: Silicon CMOS, BiCMOS, GaAs, and so on.

1.8.6 Passive Components

Passive components are a key aspect to the development of analog/RF circuits. Commonly used passive components include:

1. Resistors
2. Capacitors
3. Inductors
4. Transformers

1.8.6.1 Resistors. Resistors are used mostly for the following functionalities:

1. Load impedance of the circuit (usually an amplifier/current mode logic etc.)
2. Biasing

Key considerations in using resistors include:

1. Parasitic capacitance
2. Component matching accuracy
3. Component variation with process and temperature
4. Voltage variation
5. Sheet resistance

Any resistively loaded circuit is essentially a broadband circuit. These circuits consume voltage headroom whenever they are used in the signal path, as part of amplifiers, mixers, and so on. The bandwidth is determined by the *RC* product, where *C* denotes the output capacitance (consists of resistor's parasitic capacitance and the output capacitance of transistors). Component matching is a critical performance in determining I/Q accuracy, asymmetry in balanced amplifiers, and so on. Component matching improves with a large component area. However, parasitic capacitance increases with area and nonlinear capacitance from reverse-biased diodes increases with an increase in area.

Resistors are categorized by their sheet resistance, which indicates their to area, and in fineline CMOS technologies, process and temperature variations of resistors may lead to as large as ±40% in terms of their values. Often, "dummy" resistors are placed alongside the main resistors to improve component matching performance. Although one can ignore the process variations of DC biasing resistors, variations in the load impedance may cause transistors to run out of voltage headroom.

1.8.6.2 Capacitors. Capacitors are the second-most area-consuming elements in ICs following inductors. The main usage of capacitors includes:

1. *LC* resonating tank
2. Coupling RF signals
3. Decoupling for power supply

Capacitors are categorized by the following performance aspects:

1. Capacitance density
2. Voltage variation
3. Bottom plate parasitics
4. *Q* of main capacitor as well as its bottom plate
5. Process variation
6. Leakage (in the case of MOSCAPs)
7. Breakdown voltage

Capacitance density directly implies the area consumption on chip. Capacitive impedance decreases with frequency. Since driving higher impedances provides improvement in power consumption, we prefer lower capacitance values with a high *Q* factor (to reduce tank loading). However, a lower value of the capacitor may lead to more fringing capacitance and poor component matching. Bottom plate capacitance and its *Q* factor are major considerations, as it would lead to signal shunting to ground. Bottom plate capacitance is proportional to the area. In large signal circuits such as the PA driver, the voltage variation of capacitance is important. It is determined from the $C - V$ characteristics of a capacitor. In the case of AC coupling capacitors, a large capacitance value with low bottom plate capacitor value is desired. It is also desired that the voltage variation of coupling capacitor be as small as possible. The Q of the bottom plate capacitor would directly impact the Q of the input impedance.

A large variety of capacitors are used in mixed signal systems. Three common types of capacitors are used frequently in communication circuits:

1. FLUXCAPs
2. MOSCAPs
3. Varactors

MOSCAPs are used in the case of supply bypassing capacitors, and they may be used as coupling capacitors. Whenever a MOSCAP is used, DC leakage current may alter the DC operating point, and one must be very careful to ensure that it is modeled. MOSCAPs require proper biasing, and the capacitance value depends on the AC voltage swing. FLUXCAP, on the other hand, is made in a comb-like structure using metal structures, and it provides the lowest voltage variation. Varactors (voltage variable capacitors), on the other hand, are accumulation-type capacitors, and they are intended for large voltage variation, as they directly impact the analog tuning characteristics of VCOs. One needs to be extremely careful when using voltage variation of capacitors, as any variation in the AC swing (amplitude noise) would lead to a frequency shift. In the case of varactors, the ratio of maximum-to-minimum capacitance provides an important design guideline, and it is optimized around $4 \sim 5$ for most cases.

Capacitive impedance is 160 ohm per pF at 1 GHz, and it can be calculated at commonly used wireless frequencies such as 0.9 GHz, 2.4 GHz, 5.2 GHz, and so on.

1.8.6.3 Inductors. Inductors are essential to the development of high-frequency circuits. They add a "zero" in the transfer function of the circuit when loaded at the output, thereby boosting the high-frequency response of the circuits. Narrowband tuning and filtering is essential for the RF front end, and this is manifested by "tuning out" the capacitive load of the transistors. The current gain is dependent on the Q factor of the inductor (a high Q factor provides more gain at resonance). A fundamental advantage of using inductors is the achievable swing beyond the supply voltage, which is essential for high dynamic range, low-voltage circuits. Inductors can well extend the bandwidth of the circuit, in the case of both high-speed digital or analog/RF blocks. In summary, usage of an inductor in circuit blocks can be illustrated as follows:

1. Resonating tank
2. RF chokes
3. Lossless feedback
4. Matching networks

The Q factor of on-chip inductors is of the order of 8–10, and it depends on the area of the components. A larger area usually provides more inductance and Q, but it leads to more electromagnetic cross-talk (as the number of flux lines is proportional to the area).

Assuming that one can obtain a high Q from the on-chip capacitors (depends on the frequency of operation and on layout of the capacitors), the Q factor of the LC tank would be limited by the Q factor of the inductor. This fundamental issue has motivated various developments of area-efficient high Q inductors in digital CMOS technologies. In rest of this section we will pay close attention to various types of inductor topology and to key performance issues.

The following parameters can be used to benchmark inductor performances:

1. Inductance value
2. Q factor
3. Parasitic capacitance and Q factor of parasitic capacitance
4. Self-resonating frequency
5. Area

1.8.6.3.1 Inductor Geometry. Inductors are categorized by the following key geometrical parameters:

1. Number of turns (n)
2. Turn width (W)
3. Turn spacing (S)
4. Outer diameter (OD) (determines the area)

Qualitatively, when the inductor turns are closer, mutual coupling increases, as well as the interwinding capacitance. Hence, the inductance increases (total inductance self-inductance of each turn + mutual inductance effects between two turns), and self resonating frequency decreases. As the width W increases, the series resistance of the turns drops, and the Q factor improves. The inductance also drops because of the current flowing through the edges of the conductor. The physical construction of inductors is illustrated in Figure 1.21. The inner and outer diameters are related to each other by $OD = ID + 2n \times (W + S)$.

The inductance of an inductor is given by $L \sim n^2 \times OD$, where n is number of turns and OD is the outer diameter of the inductor. The series resistance is given by $R_S \sim n^2$. An optimum Q factor is usually obtained in the case of $ID \sim 0.5 \times OD$. While using inductors from a standard technology library, various combinations of the geometry parameters can be used and the right dependence can be obtained.

1.8.6.3.2 Self-Resonating Frequency. Let us now consider the self-resonating frequency of inductors. The self-resonating frequency determines the usable limits of an inductor. A simplified electrical model of inductor is shown in Figure 1.22.

In a center-tapped differential inductor, the parasitic capacitance can be obtained by connecting all the terminals, and by observing the overall capacitance by injecting a current source at the common terminal. The combined capacitance (C_P) can be distributed by placing two capacitors of $C_P/4$ at each end and a capacitor of $C_P/2$ at the center terminal. The self-resonating frequency is determined as $1/2\pi\sqrt{L_h C_t}$, where L_h

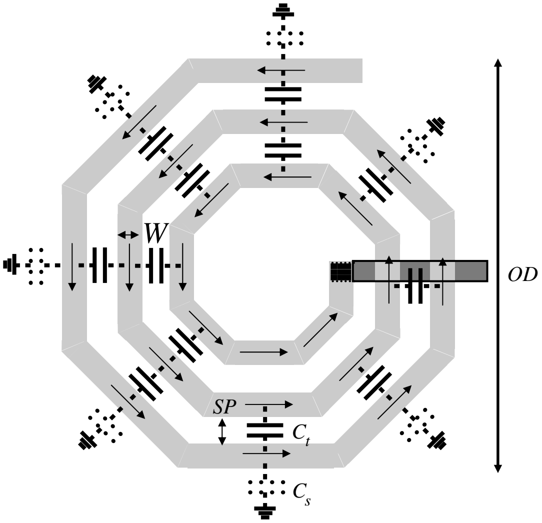


Figure 1.21. A single-ended inductor and its electrial components.

is the inductance of the half section of the inductor, and $C_t = C_P/4$. Physically, this capacitance is attributed by the interwinding capacitance and the capacitance to substrate. If the self-resonating frequency is too low, then one can reduce the outer diameter (smaller area reduces capacitance to substrate) or use a larger turn spacing S. Reducing the diameter by a factor of two would lead to a factor of 4 reduction in area, reducing the parasitic capacitance to substrate. Thus, larger inductors tend to exhibit lower self resonating frequencies. This phenomenon can also be illustrated due to the fact that, $L \sim ID$ and $C \sim ID^2$, hence, the rate of increase in capacitance is higher than the inductor, leading to lowered self resonating frequency.

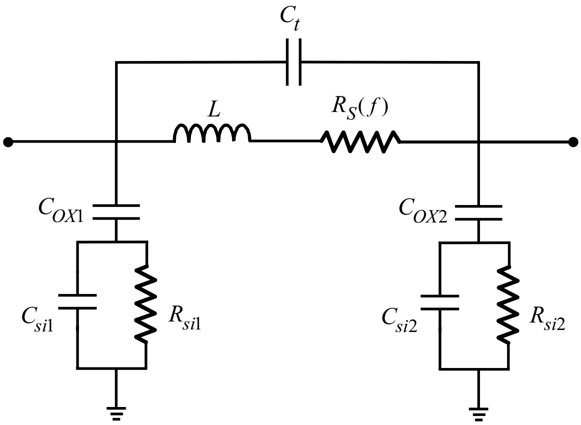


Figure 1.22. Inductor model.

1.8.6.3.3 Geometry and the Q Factor. The Q factor of inductors is dependent on the ratio of inductive impedance to its series resistance, and it is given by the ratio of area to the conductor length. For the same area of circular cross section provides the highest area-to-length ratio $(2/\sqrt{\pi})$, which leads to a high Q factor. However, it can be observed that an octagonal geometry is often a compromise, as it is easier to fabricate. Inductors generate electromagnetic fields, which propagate to adjacent circuit components, and create electromagnetic cross-talk. To prevent this action from occurring two approaches can be taken:

1. Isolate the inductor structure by providing a high resistivity exclusion zone around it.
2. Provide a low impedance substrate shield (patterned ground shield).

Both approaches are effective, depending on substrate resistivity. However, in the first approach, the electromagnetic field decays as $1/r$, where r is the distance from the excitation to the point of observation. However, the exclusion area uses much additional overhead in terms of area and processing step. In the second approach, however, the shield uses a patterned ground shield, which leads to an $1/r^2$ decay of the electromagnetic field and then to less cross-talk to the adjacent circuit components. This result can be explained with the help of image theory, which implies the conceptual formulation of image components in terms of current and charge in the substrate. For the sake of simplicity and understanding, we assume that the ground plane can be "perfect" and that the created image components would have the same strength as the original excitation element.

1.8.6.3.4 Single-Ended and Differential Inductors. We will now consider two types of inductor topologies: (1) the single-ended inductor and (2) the differential inductor. Figure 1.23 illustrates differential inductor configurations.

A differential inductor is compact in size compared with the single-ended inductor. In an inductor structure, two current flow paths exist: (1) a direct path

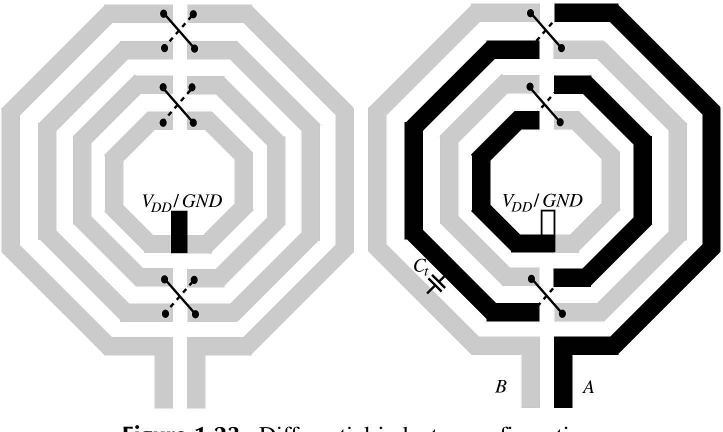


Figure 1.23. Differential inductor configuration.

through inductive component of the structure and (2) a secondary path through the capacitances or the interwinding or substrate parasitics. Current in the direct path flows along the length of the inductor, whereas it flows laterally in the secondary path. Our aim is to enhance the current flow through the direct path and to minimize the current flow through the secondary path, in order to obtain more inductive behavior. Let us assume that the inductors are used in differential circuits and that the terminal peak AC swing is V_P (i.e., differential $2V_P$). In a single-ended inductor, this voltage gradually drops across the inductor turns because of impedance, so we should simply assume that the turns get nodal voltages such as V_P, $0.8V_P$, $0.6V_P$, $0.4V_P$, and $0.2V_P$ (we assume equal drop for a five-turn inductor). Ultimately, the inductor terminals go to AC ground, no matter whether they are connected to circuit supply or circuit ground. In this case, the voltage difference across adjacent spatial turns is $0.2V_P$, which is responsible for current flow through the lateral interwinding capacitor. The inductive path is a single turn length, and the capacitive path is determined by the lateral separation of two turns (spacing). However, in the case of a differential inductor, to traverse from one turn to its adjacent turn, the inductive current flows through all the turns (a much longer path compared with the single-ended inductor structure and, hence, a larger impedance). The voltage difference is $2V_P$, for the outermost spatial turns, which leads to much higher current through the capacitive path. Hence, the effect of capacitance is more dominant in the case of a differential inductor, and their self resonating frequencies are lower. Thus, the differential inductors are used in the 3–5 GHz range. In a differential inductor, the center tap must be designed to carry twice the current limit for each of the single-ended segments of the inductor. The central tap of a differential inductor can be used to provide bias to the circuit under consideration. Differential inductors provide symmetric loading to the circuits at each port. In the case of single ended inductors, there are two terminals: (1) AC terminal and (2) underpass. In circuit implementation, the AC terminal is driven, as they usually provide a higher Q factor. Usually, in the construction of inductors, several metal layers are strapped to obtain a lower series resistance for Q factor enhancement.

1.8.6.3.5 Q Factor versus Frequency. The frequency dependence of the inductor Q factor is an important aspect for RF circuit design, and at very low frequencies, inductor loss is usually a fixed quantity, determined by its resistivity. As frequency increases, the inductive impedance ($2\pi fL$) increases and hence the Q factor leads to $Q \propto f$. As frequency increases even more, skin effect comes into the picture, and the current tends to be more concentrated toward the outer periphery of the inductor turns. The skin effect impacts the inductor metallization from all directions, and as the thickness of the AC current flow is given by $\delta = 1/\sqrt{\pi f \mu \sigma}$, the resistance increases with a square root dependency ($r \propto \sqrt{f}$). As the inductive impedance grows linearly, the Q factor grows in a square root dependence $Q \propto \sqrt{f}$. At even higher frequencies, various capacitive coupling to lossy silicon substrate occurs, and Q falls as an inverse square law dependence $Q \propto f^{-2}$. These dependencies can also be observed by plotting the Q factor versus frequency in a log scale plot. In between the transition region, Q peaks at a certain frequency, and this peak Q frequency is a very important parameter for circuit designers, as this is the optimum performance point of an inductor given a certain inductance, and area considerations. Usually it is desired that peak Q is maintained over a broad range of frequencies. The circuit performances are usually dependent on the inductors as a factor of Q^2 (or Q, depending on the circuit); hence, a humble 5% improvement in Q may improve circuit performance by 10% (1 dB).

Use of an inductor does not have to be restricted to the circuits and systems based on amplifiers only (hence, implying wireless type systems). Inductors are used in many digital circuits and systems, where the bandwidth enhancement is obtained using inductors. However, it must be observed that compared with the wireless systems, digital circuits and systems are wideband in nature, need to include all harmonics of the clock, and require a relatively lower Q. Otherwise, they would selectively amplify a specific frequency content. Such is the case for inductively loaded inverters, multiplexers, and selector circuits operating at the 20–40 Gbps range [19, 20]. Fundamentally, inductors help realize higher impedance at a specific frequency of interest, and hence, they reduce the power consumption of these circuits. Any use of inductor in high speed digital system should be strictly observed for area considerations. However, in practice, the resonating impedance cannot be increased arbitrarily. The inductor has its own parasitic components itself, and usually the impedance realized is somewhere in the 200–400-Ω range.

1.8.6.3.6 Mathematical Analysis of Inductors. The search for a solution to the electromagnetic fields resulting from an inductor has been an interesting area of research for a long time. The analysis becomes complicated because of the wide variation in inductor geometries. Current generation inductors use turn spacing to be much smaller than the turn thickness and width, and the solution can be obtained in a two-dimensional (2D) current distribution, and a mesh can be obtained using Kirchoff's laws to solve them. To compute the various components of an inductor in a lumped element representation, the current and charge distributions need to be computed. These distributions vary across the cross section of the inductor turns, and they must be obtained by "mesh"ing the inductor using a minimum grid. Using numerical computation techniques (contributions from individual mesh points with appropriate weight factor), the charge and current distributions can be shown to have peaks at the edges of turn cross sections. Once these steady-state current and charge distributions are obtained, they can be used in conjunction with some standard inductor solver configuration such as Greenhouse, and so on in order to obtain a full solution of the inductor. Although the computation is interesting in nature, because of the coverage and focus of the book, we encourage readers to refer to [27–30].

Inductor Q, however, needs to be observed at the desired frequency. Inductors do not include any voltage variable component in any of their subcomponents, and hence, they are extremely linear in nature. However, because of lithographic limitations, the metal resistances may vary significantly, which leads to changes in the Q factor. It should be noted at this stage that the inductance does not vary w.r.t. process corners, as it is related to flux linkage, which, in turn is dependent on the number of turns and outer diameter.

Inductive impedance is usually given by 6 Ω per nH per GHz. This number can be used in calculating the impedance at commonly used wireless center frequencies such as 0.9 G, 2.4 G, and 5.2 G.

At lower GHz ranges, on-chip spiral inductors provide much lower Q because of low-frequency operations (and often a bondwire inductance is preferred). To realize the same impedance, a lower frequency inductor also needs to have a higher inductance for the same Q factor. This high inductance leads to significant area consumption in the case of on-chip inductors. Such a large inductor would also exhibit more parasitic capacitance values, which leads to tuning range limitations in the case of tank circuits in amplifiers and VCOs. At the same time, they are susceptible to creating electromagnetic cross-talk.

1.8.6.3.7 Active Inductors. An alternative arrangement can be obtained using analog circuit techniques using a "gyrator-C" approach, as shown in Figure 1.24. A capacitance connected at the interfacing node of two antiparallel transconductance stages provides an inductive impedance at the other end [shown in Figure 1.24(a)]. A similar arrangement is possible in a feedback topology, where g_m of transistor M_P is multiplied by the gain of the amplifier. The inductance can be properly controlled, and it must be observed that the inductance in all of these configurations is a ratio of capacitance to transconductance; hence, the lower the transconductance (implying lower power), the higher is the inductance. Active inductors are attractive at lower RF frequencies, as a passive counterpart would not only consume more area, but it would also exhibit a poorer Q factor because of the lower frequency of operation.

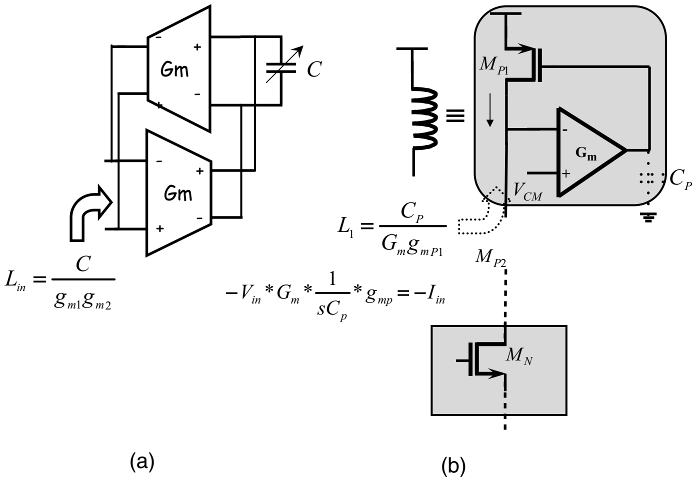

Gm
Gm
C
$L_{in} = \frac{C}{g_{m1}g_{m2}}$
(a)
≡
M_{P1}
G_m
V_{CM}
C_P
$L_1 = \frac{C_P}{G_m g_{mP1}}$
$-V_{in} * G_m * \frac{1}{sC_p} * g_{mp} = -I_{in}$
M_{P2}
M_N
(b)

Figure 1.24. (a) Gyrator-C configuration and (b) active inductor implementation.

Fundamental advantages of active inductors over the passive inductors are as follows:

1. Lower area consumption
2. No electromagnetic cross-talk

However, active inductors provide the following performance degradations as well:

1. Additional power consumption
2. Linearity degradations caused by more active components
3. Noise degradations
4. May require higher supply voltages for operation, depending on the configuration

However, as the inductance is a ratio of two dissimilar quantities, it would vary over process corners, unlike the passive inductors. Some calibration circuits would need to be used to guarantee the process-invariant behavior of the inductance.

1.8.6.4 Transformers. With the illustrations on inductors, we now illustrate the usage of transformers. A transformer is commonly used in an RF circuit to obtain either current or voltage gains. Transformers are DC isolated, and several configurations are possible as shown in Figure 1.25. The secondary terminals can use a different DC bias, and they are suitable in interfacing two fully differential circuits operating at different DC common mode levels. The DC isolation also makes them attractive for use in a feedback network feedback network without the need of DC blocking capacitor [32]. Voltage and current gains of a transformer are related by the

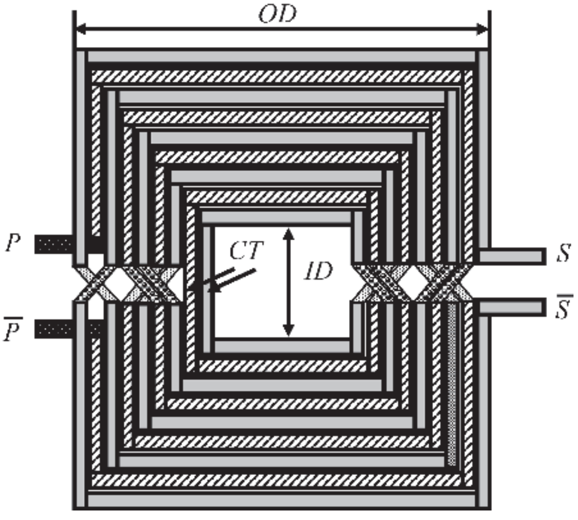


Figure 1.25. Physical construction of a transformer.

turns ratio (N) of transformers, whereas the impedance is related by N^2. Thus, it can act as an impedance buffer to reduce the effect of capacitive parasitics. The primary and secondary terminals of transformers provide inductive impedance, which can be resonated with various capacitances, such as device output capacitance, and so on. Another use of transformer is in the single-ended to differential transformation (balun) with current or voltage gains, and this operation can be performed with minimum imbalance between the two differential terminals. This implementation is important in the front end, as the LNA takes an input single-ended signal, which should be transformed to differential as early as possible with lowest amount of imbalance. Circuit elements based on active components tend to provide more imbalance than their passive counterparts.

The key element to successful implementation of transformer is the coupling coefficient of flux linkage from primary to secondary, denoted by $K(<1)$. Also, they are area inefficient. An efficient way of implementing transformers is to use an autotransformer configuration. A differentiator in this configuration is the fact that the terminals are not isolated in terms of DC voltages. However, this can be easily obtained from a simple inductor structure, which takes advantage of mutual coupling between the terminals.

At further higher frequencies (well into the mmW regime), inductors assume a distributed component model, rather than the standard lumped element model that we have been illustrating until now. In that situation, inductors are mostly microstrip line components. If we were to unwind the spiral inductor, it would certainly be a long line. Spiral winding helps in improving mutual coupling and increases inductance in a given area under considerations. At the same time, the inductance per unit length is lower because of interwinding capacitances. However, at frequencies comparable with the wavelengths of the operating frequencies, these capacitances would be prohibitive, and the physical dimension of the inductor would be comparable with the wavelength, so we can obtain larger inductance per unit length without the area penalty.

Another interesting component at mmW frequency range is called a “stub.” A stub is essentially a transmission line used to realize “short” or “open” impedances in a frequency dependent manner. They also provide characteristic impedances, which is repeatable with frequency, implying that a stub providing an “open” circuit impedance would provide the same impedance (theoretically) at frequencies f, $2f$, $3f$, and so on, while providing a “short” circuit impedance at $1.5f$, $2.5f$, and so on.

1.8.7 Evaluation Testbenches

Prior to designing circuits, it is important to obtain some useful information about the technology components. These testbenches provide useful information about the technology to be used and its performance metrics. The following testbenches are absolutely essential for circuit design:

1. Transistor f_T, f_{MAX}, ON resistance, OFF capacitance, minimum noise factor
2. Parasitic capacitance of passive components (resistance, capacitance, inductance)

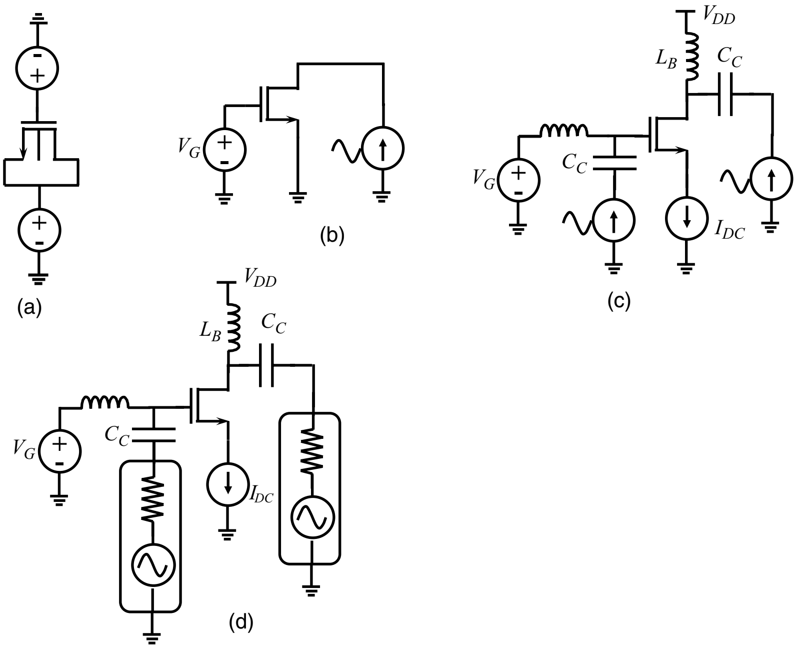


Figure 1.26. Various testbenches to characterize transistors.

3. Impedance testbenches
4. Nonlinearity testbenches

Figure 1.26 illustrates the various testbenches used to characterize the transistor. In fact various types of transistors in the technology include high speed transistor, high breakdown voltage transistor, lightly doped transistor, thick gate oxide transistor, unadjusted threshold voltage (native) transistor, and so on. f_T, f_{MAX} values should be obtained for the actual dimension of the transistor used in the circuit. The ON resistance of the MOS transistor and the OFF capacitance of the MOS transistors are important circuit considerations. ON resistance is pertinent in the cases of (1) passive mixer and (2) MOS switches in a binary capacitor array. In the case of a passive mixer circuit, ON resistance indicates the signal loss and the achievable Q factor of the mixer connected with a coupling capacitor. In the case of a switched capacitor binary array, it implies the Q factor of switchable components. ON resistance depends on the gate to source voltage, threshold voltage, and the transistor geometry. As the voltage at source terminal changes, the ON resistance would also vary accordingly, and signal-dependent variation of the ON resistance would provide important design insight. The OFF capacitance of a MOS transistor depends on the transistor geometry and also varies with voltage values. The OFF capacitance is a measure of the junction capacitances (drain to bulk or source to bulk). It is illustrated in the testbench (a)–(b). One can set the input voltage to different values to obtain the on resistance and off capacitance.

Testbench (c) can be used to set the gate voltage and bias currents separately, and the transistor's current gain, output impedance (g_m, g_{ds}), and so on, can be obtained. The biasing inductors and capacitors are of large values in order to provide high impedance to the DC biasing source and low impedance to the AC, respectively. The magnitude as well as the Q factor of these impedances is important for circuit design. It is usually obtained by separating real and imaginary parts of the input impedance obtained by simulating it with a 1-A current source. Testbench (d) can be used to estimate noise factor, and so on using 50 Ω ports at the input and output. This may provide insight to various types of microwave parameters for designing amplifiers and so on. Current sources are used to measure impedance, as current sources provide high output impedance, and minimum loading to the circuit, whereas voltage sources load the circuit by providing zero impedance. The minimum noise factor indicates the minimum achievable noise figure of a MOS transistor (with appropriate geometry and current consumption). An amplifier's noise figure is given by a parabolic curve w. r.t. impedance states. The minimum noise factor provides the lowest possible noise figure for specific transistor geometry at a given bias current. In reality, the minimum noise factor is never achieved because of the consideration of simultaneous noise and input return loss for LNAs.

Figure 1.27 illustrates component-level testbenches, which includes resistance, capacitance, inductance, and so on. The parasitic element of a resistance determines bandwidth shrinkage, whereas the parasitic capacitance of a capacitor (known as the "bottom plate capacitor") indicates a shrinkage in tuning range of integrated narrowband circuits (VCOs, filters). Assuming parasitic capacitance to be evenly distributed across the device terminals, a two-terminal device can be shorted across the functional terminals (the two terminals of a resistor, inductor, or capacitor). Parasitic

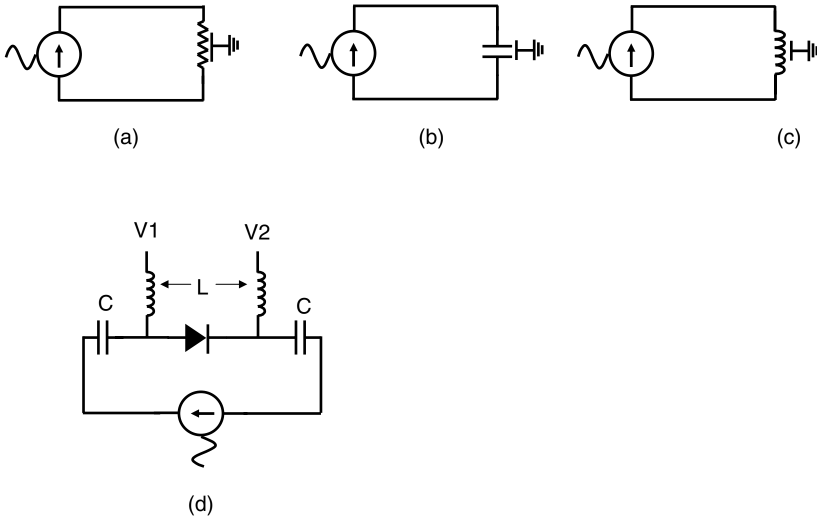


Figure 1.27. Various testbenches to characterize technology components as well as parasitic elements of (a) resistor, (b) capacitor, (c) inductor, and (d) diode.

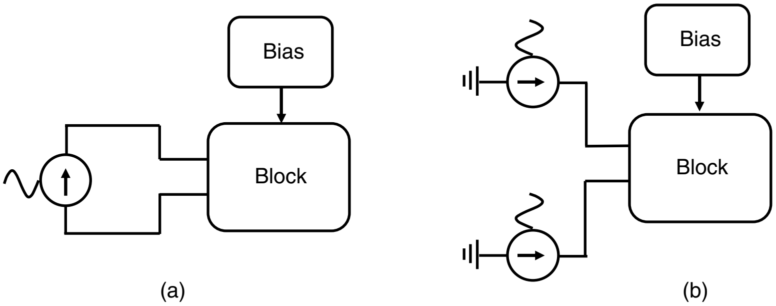


Figure 1.28. Various testbenches to characterize impedance of building blocks: (a) differential mode and (b) common mode.

capacitances indicate loading at RF, and they determines the maximum usable frequency for a component (self-resonating frequency for an inductor). Like any component, along with the parasitic capacitance value, the *Q* factor of the capacitance is important for circuit design considerations. In the case of varactors, capacitance versus voltage characteristics (known as *C–V* curves) are important to obtain information on the maximum and minimum capacitance values, the *Q* factor, as well as on potential AM-to-PM conversion. Figure 1.27(a)–(c) illustrates the resistance, capacitance, inductance, and so on, whereas (d) illustrates the input impedance determination of a diode with bias voltage applied as (V_1–V_2). To obtain the impedance, this voltage can be swept from negative to positive, and the voltage variation of the impedance, as well as its *Q* factor, can be obtained. In the figure, both the inductor and the capacitor are of large values to provide DC bias voltage and AC short without loading each other. This is a fully differential characterization, which is more accurate in terms of the voltage variation of parasitic components at each port. Similar approach can be taken for the other voltage variable components (e.g., varactors).

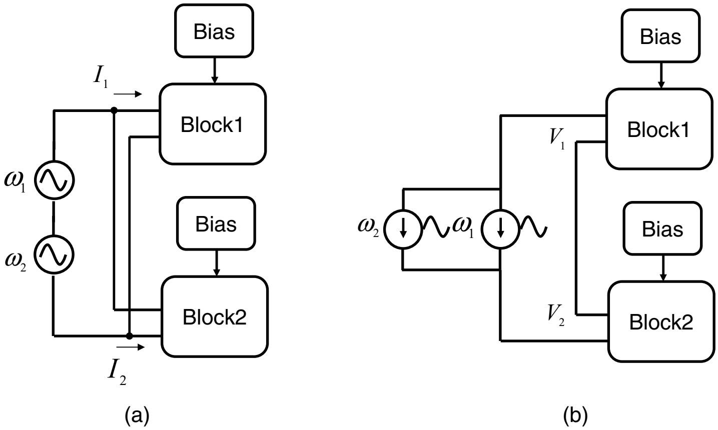


Figure 1.29. Testbenches to characterize nonlinearity: (a) with a two-tone voltage input and (b) with a two-tone current input.

In integrated circuit implementations, signal paths, as well as signal generation paths are usually differential in nature to reduce the common mode noise perturbation. Hence, the common mode and differential mode input impedances, as well as their Q factors, would need to be obtained for a circuit design. For an active circuit, proper bias condition must be provided to characterize the input and output impedances. In an integrated mixed signal environment, our aim is to amplify differential signals and reject common mode signals as much as possible. In this regard, the circuit interfaces are designed in a way such that the common mode path has lowest possible Q. If this is not the case, any finite common mode current resulting from mismatched differential circuit would lead to common mode instability. Figure 1.28 shows two testbenches to obtain differential and common mode impedances.

In a series/parallel combination of two or more components, it is imperative to know which component provides linearity limitation. Examples of these would be a parallel combination of three functional blocks in a highly linear system and so on. Such information can be obtained by providing two tones in the voltage domain (a series combination of two voltage sources) across the parallel network, and observing intermodulation terms in the individual branch currents. Two estimate linearity, two mechanisms can be used: (1) provide a two-tone voltage input (series combination) to a parallel combination of two blocks, and (2) provide a two-tone current input (parallel combination) to a series combination of two blocks. Usually, for intermodulation tests, the circuits are characterized at their linear operation range. Nonlinearities are embedded as part of signal-dependent behavior of the impedance (input/output). With two-tone voltage input as stimuli, the discrete Fourier transform (DFT) of input currents are performed to understand the block that is most limiting in terms of linearity. While performing a linearity simulation, the simulator options must be specified for the desired accuracy. It is usually obtained by providing a specific timestep of analysis in a time domain simulation. Figure 1.29 shows testbench to evaluate linearity performance using two tone voltage and current stimuli.

1.9 KEY CIRCUIT TOPOLOGIES

In this section, we will illustrate the fundamental circuit basis functions. These cores can be implemented using any type of transistors and can be used anywhere in a complicated system.

1.9.1 Differential Circuits

Differential circuits consume twice the area compared with single-ended counterparts, but they are used extensively in integrated circuits because of several key advantages such as follows:

1. Rejection of common mode signals (substrate noise, even order terms).
2. No additional hardware to create 180° phase-shifted version (inversion).

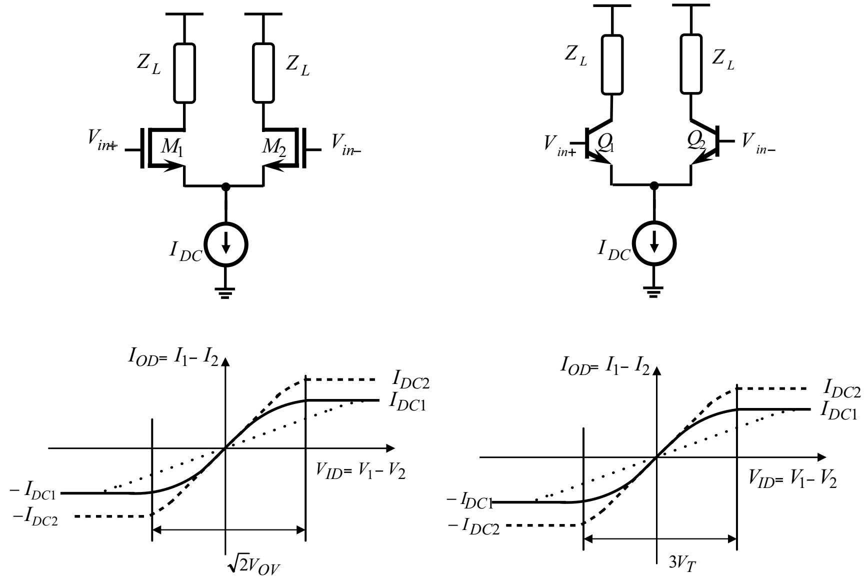


Figure 1.30. Basic differential pair topology.

3. Insensitivity to package parasitics connected at common mode terminals.
4. Increasing speed of digital circuits, and possibility of using lower supply voltage in I/O rails.

First in this category is the differential input transconductor stage, which may use bipolar or MOS devices at its input, as shown in Figure 1.30. We assume that the tail current is kept constant, and there is no voltage headroom problem across the transistors (they operate in active and saturation regions in the cases of bipolar and MOS, respectively, $V_{CE} > V_{CE,SAT}$, and $V_{DS} > V_{DS,SAT}$). Such configurations are also referred to as "source coupled" stages. In terms of large signal characteristics, the current/voltage relationships are given as follows:

$$\frac{I_{c1}}{I_{c2}} = e^{V_{id}/V_T}$$

$$I_{TAIL} = \frac{1}{\alpha_F}(I_{c1} + I_{c2})$$

$$V_{od} = V_{c1} - V_{c2} = \alpha_F I_{TAIL} R_L \tanh\left(\frac{-V_{id}}{2V_T}\right)$$

Thus, the input–output DC transfer characteristics in the case of a bipolar diff-pair is dependent on the thermal voltage $V_T = \frac{kT}{q}$. As can be observed from the plot shown in Figure 1.30, this stage is "hard-switched" when the differential input voltage exceeds $3V_T = 78$ mV. The slope around the zero point in the transfer curve plane determines the small signal amplification. The amount of deviations of this slope from a constant value (a perfectly linear curve would provide a constant derivative) determines small signal linearity (observed by providing two tones at the input). Large signal linearity is usually determined by signal clipping (essentially any arm of the differential pair running out of current). The derivative of the *I–V* characteristics determines the transconductance of the differential pair stage.

Hard-switched differential pairs behave as limiters in the case of driving mixers and so on, in order to reduce process and temperature variation of the input signal waveform. However, they are usually nonlinear in nature, and they use a small linear range. These problems can be alleviated by providing resistive feedback to linearize the input stage without sacrificing some headroom. A resistive degeneration linearizes the input stage.

A similar configuration can be performed using any type of transistors (MOS/MES FET, etc.). For the sake of simplicity, we will assume square law *I–V* characteristics to obtain the trends and insights in circuit design. In deep submicron technologies, the *I–V* curves are much different from a square law representation. From the individual gate-source voltages, we obtain

$$V_{id} = \frac{\sqrt{I_{d1}} - \sqrt{I_{d2}}}{\sqrt{\frac{K_N}{2}\left(\frac{W}{L}\right)}}$$

$$I_{TAIL} = I_{d1} + I_{d2}$$

$$I_{d1,2} = \frac{I_{TAIL}}{2} \pm \frac{K_N}{4}\left(\frac{W}{L}\right) V_{id} \sqrt{\frac{4I_{TAIL}}{K_N(W/L)} - V_{id}{}^2}$$

Both transistors operate in the saturation region when $|V_{id}| \leq \sqrt{\frac{2I_{TAIL}}{K_N(W/L)}} \Rightarrow |V_{id}| \leq \sqrt{2}V_{OD}$. Various curve families are shown in Figure 1.30. Once again, the linear range can be extended by using resistors at the sources of the transistors. When resistors are prohibitive in terms of headroom and noise, inductors can be used. As the differential input voltage range for linear operation is higher, MOS differential stages usually offer higher linearity compared with their bipolar counterparts with no degeneration added. With degeneration, these considerations would differ. In terms of a "hard-switched" differential pair (driving mixer switches, etc.), it can be observed that the bipolar transistors would require only a 78-mV differential signal swing in order to switch, whereas MOS can easily require 100–200 mV. Lowering of this voltage leads to larger transistor dimension and to increased loading on the signal generation network. Thus, bipolars prove to be attractive in these cases.

Apart from headroom, one needs to be careful about the pole created by the *RC* equivalent circuit designed by the degeneration resistor and the source capacitor. In high-frequency circuits, inductive degeneration is used, and this may resonate with the source capacitance as well. Careful attention should be provided in order to avoid such scenarios.

Tail impedance plays a critical role in differential circuits, as the common mode rejection performance is heavily dependent on its value. The extent of rejection experienced by unwanted signals is determined by the common-mode rejection ratio

(CMRR), and it is given by CMRR $= 1 + 2g_m R_T$. In reality, this is obtained by a transistor (part of a current mirror) or a small resistor. In a fully differential amplifier configuration, the output capacitance is not important, as it experiences an AC ground due to input signal symmetry. However, in reality, imbalance exists because of asymmetry in the device layout, input duty cycle error, and it is desired that the output capacitance be as small as possible.

Voltage offset for differential amplifiers is always referred to the input terminals. This is similar to small signal noise, and the overall mismatch is computed in terms of components in the amplifier stages, and then it is divided by the amplifier gain. In computation of offset voltages, we assume the components to be slightly mismatched and apply a differential voltage at the input to ensure that the output voltage is zero. Under these conditions, a MOS differential stage can be solved to obtain

$$V_{OS} = \Delta V_t - \frac{V_{OD}}{2}\left(\frac{\Delta R_L}{R_L} + \frac{\Delta W/L}{W/L}\right)$$

1.9.2 Translinear Circuits

The second widely used variety is translinear circuits, which are illustrated in Figure 1.31. A simple illustration of this principle can be provided by considering bipolar transistors, in which an equal number of clockwise and anti-clockwise transistors is connected from a reference point to the ground.

As the base-emitter voltages of bipolar transistors are given as a logarithmic function of the collector currents, summation of the clockwise and the anti-clockwise path V_{be}s can be equaled to provide

$$\prod_{CCW} I_i I_j = \prod_{CW} I_i I_j$$

In actual implementations, these currents would be a sum of DC biasing current and an AC current, providing versatile functions such as multiplication, division, and so on. Translinear circuits can also be realized using MOS transistors, and in modern technologies, the subthreshold operation of MOS transistors becomes analogous to bipolar transistors because of their exponential characteristics.

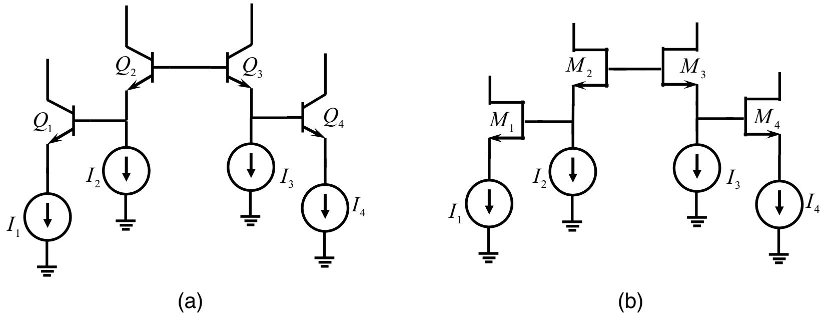


Figure 1.31. Translinear circuits using (a) bipolar and (b) MOS.

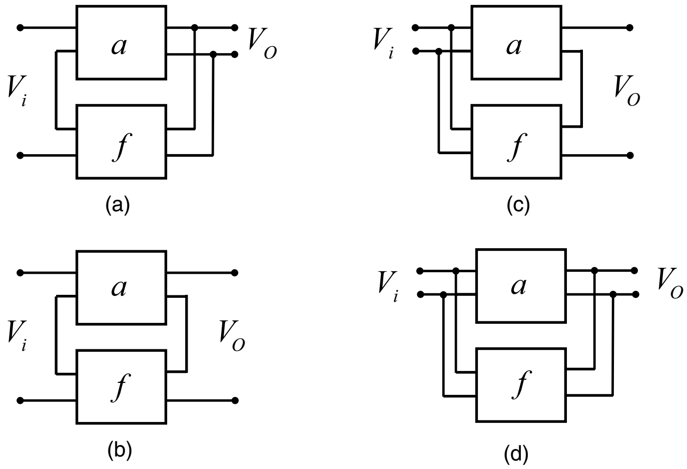


Figure 1.32. Feedback configurations: (a) series-shunt, (b) series-series, (c) Shunt-series, and (d) Shunt–shunt.

1.9.3 Feedback Circuits

Feedback plays an important role in integrated circuits and systems. Negative feedback is commonly used in linearizing amplifiers, reducing input impedance value, with the compromise of noise addition in the circuit. Positive feedback is essential in determining oscillation start-up conditions. Depending on the placement of the feedback component, their nomenclature is followed as follows: (1) Series–shunt, (2) shunt–shunt, (3) shunt–series, and (4) series–series. These topologies use voltage or current as sampling and feedback variables as appropriate. These configurations are illustrated in Figure 1.32. Negative feedback is used extensively in circuit linearization, however, they always contribute noise to the main circuit. Positive feedback has been used extensively in oscillator circuits.

1.9.3.1 Feedback in OP-AMPs. We will pay close attention to feedback loop of an OP-Amp, which is a shunt–shunt topology. In this case, the output current is sampled through the feedback resistor R_F and injected to the input terminal. The feedback resistor reduces the gain of the OP-Amp, and an open loop gain of 60 dB reflects to a close-loop gain of R_F/R_s, where R_S is the source resistance. It can be observed that although the OP-Amp consumes DC current, the gain is set by the resistor ratios! The DC current in the OP-Amp branches ensures that it can provide the large signal current swings. The input impedance is given by $Z_{in} = \frac{R_F}{1-A_v}$, where A_v denotes the open-loop voltage gain of the OP-Amp. If we were to assume the open-loop response of the OP-Amp to have a single pole response, then $A_v = \frac{A_{v0}}{1+s/\omega_P}$, which leads to inductive behavior in the input impedance of the OP-Amp.

1.9.3.2 Virtual Ground. The presence of low input impedance is referred to as “virtual ground.” A low impedance implies a “short” between the two terminals in their electrical signal equivalence but not physically. The OP-Amp is capable of sinking any amount of current as required. In reality, the current does not flow to ground, but it flows through the feedback impedance path. Thus, we have the name "virtual ground."

1.9.3.3 Miller's Theorem. Analysis of electrical circuits, where input and output terminals are coupled through an impedance element, can be performed using Miller's theorem. It is a generalized analysis, and it can be performed with any impedance connected in the feedback path. Using Miller's theorem, we establish one-to-one equivalence between the two circuits, as illustrated in Figure 1.33. Using Miller's theorem, we aim to decouple the input output–connection by using mathematical representation. This approach helps in the analysis of the circuit's input and output nodes. We first solve the two systems w.r.t. nodal equations, and we use voltage equivalence to obtain

$$Z_{O,M} = \frac{Z_f - Z_O/|A_v|}{1 + 1/|A_v|} \approx \frac{Z_f}{1 + 1/|A_v|}$$

$$Z_{O,M} = \frac{1 + Z_O/Z_f}{1 + |A_v| + Z_O/Z_f(1 - |A_v|)} \approx \frac{Z_f}{1 + |A_v|}$$

The details of this derivation are provided in Appendix A(7).

In the above discussion, we have assumed that the feedback impedance is larger compared with the series output impedance, which is a reasonable assumption. Conceptually, Norton's equivalent circuit can also be obtained, and the amplifier

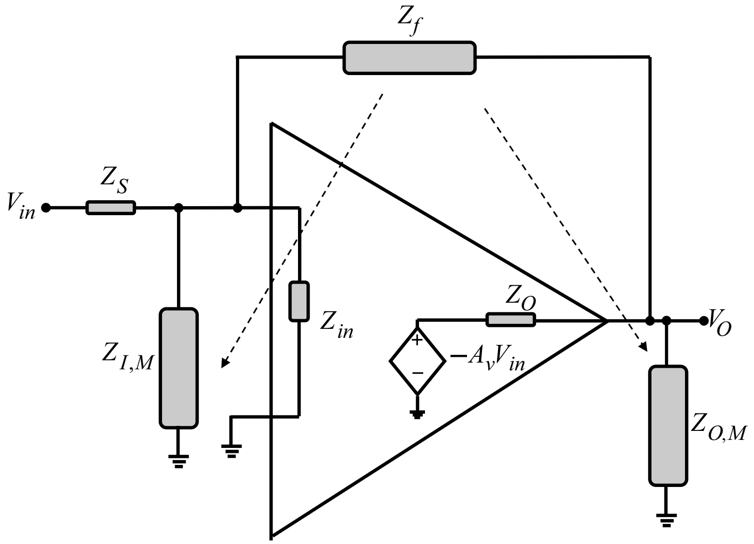


Figure 1.33. Illustration of virtual ground and Miller's impedance.

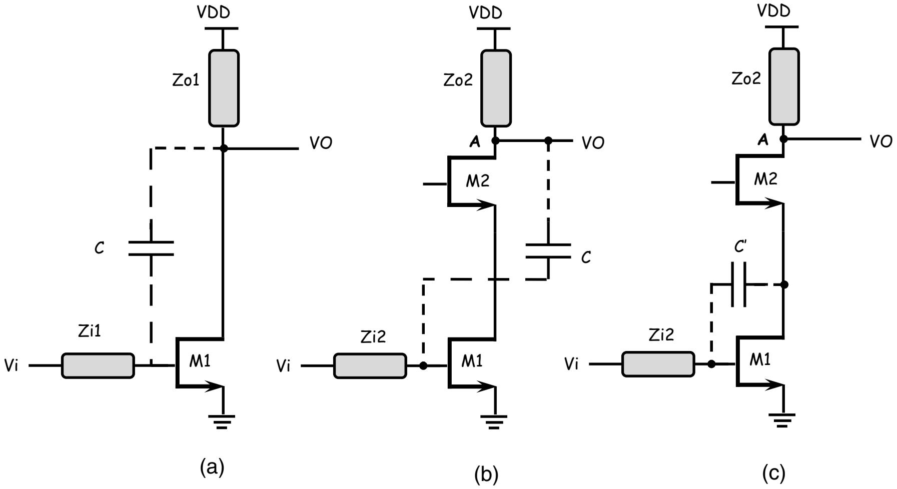


Figure 1.34. Miller effect consideration leading to cascode topology.

can be represented by the current gain. Miller's theorem provides important insights into several feedback circuits. The previous section on OP-Amp's virtual ground formulation was also an illustration of Miller's theorem. It can be observed that, if the feedback impedance is capacitive, it would lead to an equivalent input capacitance equal to the original capacitor multiplied by the open-loop gain of the amplifier. At RF frequencies, this leads to significant bandwidth reduction. Even if a capacitor is not connected deliberately, the parasitic capacitance of the device would provide the same effect. This is shown in Figure 1.34.

1.9.4 Cascode Circuits

To alleviate this problem, a cascode topology is often adopted. The purpose of a cascode transistor is to isolate the input and output networks, as shown in Figure 1.34. Assuming a voltage gain of 24 dB (linear factor of 16), without using cascode, the effective capacitance seen at the input is 16C. Assuming the same geometry of the main and the cascode devices, the voltage gain resulting from the main device is close to unity (ratio of transconductances). According to Miller's theorem, the input referred capacitance is 2C, which reduces the loading by 8 times. For this reason, cascoding is almost always used in RF circuits. Several other advantages exist apart from (1) bandwidth enhancement and (2) reverse isolation. As the input and output terminals are decoupled, cascoding helps in separate optimization of input and output matching, and it achieves high output impedance to provide high voltage gain. However, compared with a non-cascoded variant, a cascode device consumes more headroom, which leads to reduced signal swing. In the non-cascoded variant, the Miller capacitor reduces the output impedance, but the output can swing higher at the expense of increase current consumption.

1.9.5 Common Source, Common Gate, and Common Drain Stages

We will now cover a few basic circuit topologies, which can be easily analyzed in the analog domain and have been used as part of circuits in any building blocks. These topologies are known as (1) common source, (2) common gate, and (3) common drain (or source follower). They are named according to the terminal that is common in the input and the output networks. For the case of the common source amplifier, the source terminal is "common" between input and output and so on. Important performance parameters associated with this stage are as follows:

1. Gain, linearity, and noise
2. Input common mode range
3. Output common mode range
4. Current consumption and bandwidth
5. Signal handling capability

Input common mode implies the amount of signal swing at the input. It also implies the possibility of direct coupling among various blocks. A low common mode ($<V_{DD}/2$) is suitable for a PMOS input stage, whereas a high common mode is suitable for the NMOS-type input stage. This consideration is important at low frequency, as the stages cannot be capacitively coupled because of the area constraints. To interface two stages, we either require a level shifter or complementary stages (NMOS-based stage interfacing PMOS and so on). Gain bandwidth product is an important performance criteria for such amplifiers and is dependent on the bias current and load impedance. The signal handling capability is denoted by $V_{DSAT}$and by the bias current of the amplifier. Higher V_{DSAT} implies higher linearity. Linearity can also be improved by using degeneration resistors (or inductors at high frequencies), which provide negative feedback around the transconductor stage, reducing the overall gain. However, any capacitive component in the degeneration path leads to formation of negative impedance at the input, which leads to stability issues. The multiplication factor is $\frac{g_m}{j\omega C_{gs}}$, and an inductor leads to real input impedance, a resistor leads to capacitive input impedance, and a capacitor leads to negative impedance (causing stability concerns!). Hence, any parasitic capacitance at the source terminals needs careful attention in the design phase.

An advantage of using common source type topology is the amplification in both voltage and current domain. Other topologies, such as common gate, provide voltage amplification at the output (dependent on the load) but not current amplification. The input impedance of the common gate stage illustrated in Figure 1.35 is given by $Z_{in} = \frac{r_{dsL} + r_{ds1}}{1 + g_{m1} r_{ds1}}$. Thus, $Z_{in} \sim \frac{1}{g_{m1}}$, in the case of low load impedance, which is usually the case for RF circuits.

At lower frequencies, active loads are commonly used, and the input impedance can be $Z_{in} \sim \frac{2}{g_{m1}}$, assuming equal output impedance from the core transistor and the load. As the input impedance is lower, the voltage swing at the input is also reduced (leading to lower voltage variation of input capacitance). Current transfer is linear, which causes high linearity at this stage. The input impedance is wideband in nature, and this

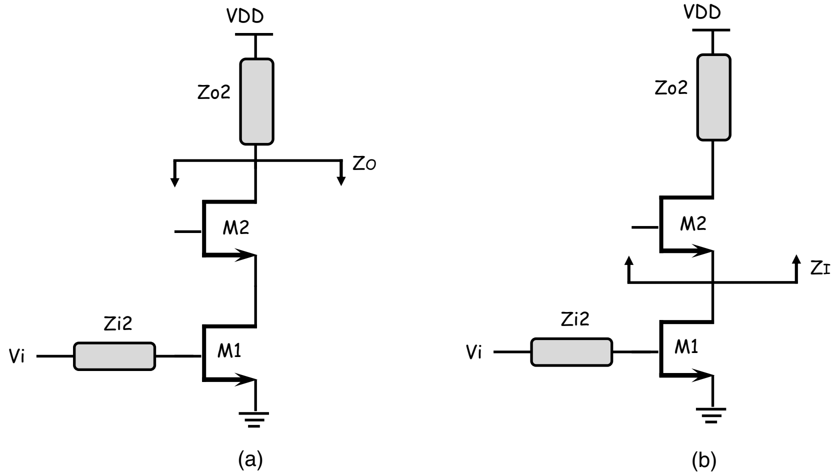


Figure 1.35. Impedance calculations.

can be optimized to 50 Ω in order to provide a power match. As the driving impedance is also 50 Ω, the minimum achievable noise figure is ~3 dB. Common source stages are usually preferred for LNAs, with simultaneous noise and power matching design considerations, as will be illustrated later.

Common drain or source/emitter followers (shown in Figure 1.36) are primiarily used as an impedance buffering stage, and they are capable of providing a voltage shift. In these cases, the source/emitter terminal voltage follows the gate/base, and no phase change occurs in terms of voltage transfer. However, the maximum voltage amplification out of this stage is usually lower than 1, and it is usually dependent on the g_m and the source/emitter impedances Z_s. It is desired that $g_m Z_s >> 1$. To meet this

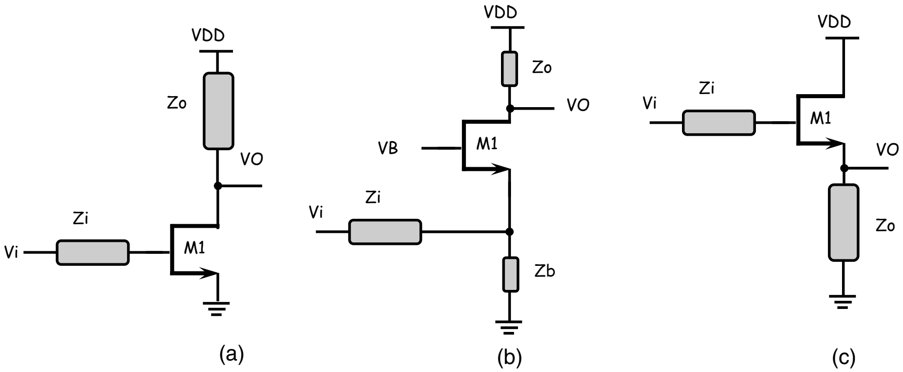


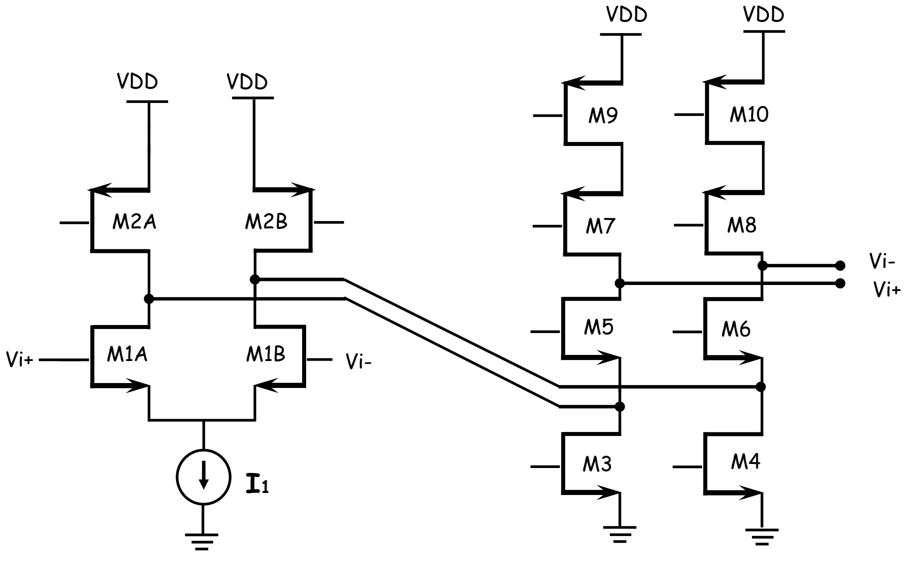
Figure 1.36. Common single transistor amplifier topologies: (a) CS, (b) CG, and (c) CD.

requirement, these stages can consume significant current. Also, because of body bias and other nonlinearity factors in V_t, the voltage gain is usually slightly lower than unity (usually 0.7–0.8) in UDSM CMOS nodes. These stages provide no voltage amplification, but they add noise, which leads to dynamic range degradations. However, they are effective in impedance buffering. As the voltage gain $A_v \approx 1$, the output capacitance is reflected back to the input by an amount $C_L(1-A_v)$, which causes a small impedance referred to at the input. The input impedance of these amplifiers is high, so they are perfectly suitable for driving large loads (off-chip) at low frequencies for measurement purposes. In this situation, enough gain is placed before them, so the resulting noise would not have much impact on system performance, and at the same time, the output capacitance of 5 pF would appear as 1 pF (with a voltage gain of 0.8) at about 1 Mhz, providing 160 kΩ, which would imply much reduced levels of loading. Similarly, their low output impedance suggests almost perfect voltage transfer to the subsequent driven networks.

1.9.6 Folded Cascode Topology

In low-voltage circuits, a configuration often known as "folded-cascode" is used to meet headroom requirements, as shown in Figure 1.37. In this configuration, current is "folded" through a high impedance to the delivering impedance. DC currents in these branches should be able to withstand the AC current swing.

Any cascode stage boosts the output impedance by the factor $g_m r_{ds}$, which leads to an output impedance of $g_m r_{ds} R$. In a stacked transistor configuration, the output



(a)

Figure 1.37. NMOS input folded cascode circuit configuration.

impedance of a transistor can be boosted as well, which results in the to $g_m r_{ds} r_{ds2}$. In the case of a bipolar transistor stack, the output impedance is βr_O. However, these impedances are only valid in the low-frequency regime of operation. At RF, these are shunted through output capacitances.

1.10 GAIN/LINEARITY/NOISE

Any building block in a mixed signal communication system can be attributed in terms of its gain, linearity, and noise contribution. Any circuit design for a certain functionality evolves around a systematic design compromise among these variables. Noise contribution can be considered in terms of a linear and nonlinear operation, as appropriate. These parameters are always considered in a cascaded system. The first block in the chain dominates in terms of noise contribution, whereas the subsequent stages determine the linearity. Hence, the linearity is usually associated with the last stage in a cascaded chain. Two types of linearity can be attributed: (1) small signal and (2) large signal. Small-signal, linearity is determined by the slope of the DC transfer characteristics near the origin whereas large-signal behavior is usually associated with the clipping behavior of signals, in an amplifier stage. Small-signal linearity terms are usually attributed in terms of two-tone inputs with closely spaced frequencies, and they are attributed as IIP3, IIP2, and so on. IIP3 relates to nonlinear terms in the output because of cubic nonlinearity, whereas IIP2 results from second-order nonlinearity effects, and is a function of component mismatches, duty cycle errors, and so on. In practical systems, output power levels of third-order intermodulation products are obtained and plotted as a function of input power level. In a double logarithmic scale plot, IM3 has a slope of 3, whereas the fundamental power has a slope of 1. IIP3 is given as the value of input power where these two cross each other. Physically, nonlinearity implies the redistribution of total available power from fundamental component to spectral harmonics. IM2 has a slope of 2, and the corresponding intersection with fundamental power is attributed as IIP2. Linearity is also related to the out-of-band blockers, and they need to be filtered using resonating tanks or passive filtering techniques as the signal propagates through the receiver chain. It must also be noted that while referring to IIP2 or IIP3 or any IIP products in (dBm), the referring impedance should be $50\,\Omega$ (or $75\,\Omega$ in the case of video standards). Otherwise they should be represented in terms of voltage or current (dBV or dBI) as appropriate.

1.10.1 Noise and Intermodulation Tradeoff

It can be observed that the linearity of cascaded blocks depends heavily on the phase of the signals traversing through them. Because of the phase shifts in the cascaded building blocks, it is possible to obtain cancellation of intermodulation terms. They also experience phase rotations because of various phase-shifting combinations such as *RC*, *CR*, and so on. Any filtering of out-of-band blockers relaxes the out-of-band linearity requirements.

Although phase is important for intermodulation, in terms of addition/subtraction, noise always adds in an uncorrelated fashion. The noise of cascaded systems is determined by the individual noise factors (linear scale) and the gain of the preceeding stages. As can be observed, providing more gain in the first few stages reduces the noise figure, but the increased level of signal causes linearity degradation. Thus, for a given power consumption, noise and linearity always pose tradeoffs in system design.

1.10.2 Narrowband and Wideband Systems

In narrowband communication systems, noise contributes more in order to degrade SNR, compared with nonlinearity. However, this scenario changes in the case of wideband communication systems, especially using multicarrier modulation techniques. Instead of evaluating nonlinearity terms based on two input tones, now we can use three or more, which leads to formation of triple order beats [terms located at $(f_1 + f_2 - f_3)$, $(f_1 - f_2 + f_3)$, etc., in addition to formation of $(2f_1 - f_2)$ and $(2f_2 - f_1)$]. It can be easily observed [shown in Appendix A(8)] that the magnitudes of triple beats is 6 dB higher than the original IM3 terms. Hence, in wideband systems, the intermodulation floor rises faster compared with narrowband systems, which causes higher power consumption in the building blocks.

CONCLUSION

In this chapter, we have provided the readers with the fundamental concepts to understand the basic principles of communication systems and circuit design. Key analysis methods have been illustrated along with circuit simulators and system design parameters. We have tried to focus on the basis functions that occur in all communication circuits and systems. Although the technology platforms keep changing, and various communication technology standards keep evolving, fundamentals are applicable everywhere, and they can provide a basic tool for understanding the principles of design. In the subsequent chapters, we will apply these concepts to solve complicated systems.

REFERENCES

Communication Systems

[1] S. Haykins, *Communication Systems*, John Wiley and Sons, 1995.

[2] J.G. Proakis, *Digital Communications*, McGraw-Hill, 2nd edition, 1989.

[3] G.L. Stuber, *Principles of Mobile Communication*, 2nd edition, Kluwer Academic Publisher, 1996.

[4] B. Sklar, “A structured overview of digital communications-A tutorial review-part I,” *IEEE Communications Magazine*, Vol. 21, No. 5, Aug 1983, pp. 4–17.

[5] B. Sklar, “A structured overview of digital communications-A tutorial review-part II,” *IEEE Communications Magazine*, Vol. 21, No. 7, Oct 1983, pp. 6–21.

[6] William C.Y. Lee, *Mobile Cellular Telecommunications*, 2nd edition, McGraw-Hill, 1995.

Electromagnetics

[7] R.E. Collin, *Foundations for Microwave Engineering*, 2nd edition, John Wiley and Sons, 1992.

[8] D.M. Pozar, *Microwave Engineering*, 2nd edition, John Wiley and Sons, 1998.

[9] D.J. Griffiths, *Introduction to Electrodynamics*, 2nd edition, Prentice Hall, 1989.

[10] J.R. Reitz, F.J. Milford, and R.W. Christy, *Foundations of Electromagnetic Theory*, 3rd edition, Addison Wesley, 1980.

[11] R.F. Harrington, *Field Computation by Moment Methods*, Oxford University Press, 1993.

[12] S. Ramo, J.R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, 3rd edition, John Wiley and Sons, 1994.

Basic Mathematical Analysis

[13] E. Kreyszig, *Advanced Engineering Mathematics*, 8th edition, John Wiley and Sons, 2001.

Digital/Mixed-Signal Circuit Design

[14] H. Taub and D. Schilling, *Digital Integrated Electronics*, McGraw-Hill, 1977.

[15] J.P. Uyemura, *Digital MOS Integrated Circuits*, Kluwer Academic Publisher, 1999.

[16] R.J. Baker, H.W. Li, and D.E. Boyce, *CMOS Circuit Design, Layout, and Simulation* volumes 1 and 2, Prentice Hall, 2002.

[17] P.E. Allen and D.R. Holdberg, *CMOS Analog Circuit Design*, 2nd edition, Oxford University Press, 2004.

[18] D.A. Johns and K. Martin, *Analog Integrated Circuit Design*, John Wiley and Sons, 1997.

[19] P.R. Gray, P.J. Hurst, S.H. Lewis, and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 4th edition, John Wiley and Sons, 2001.

[20] S. Gondi, J. Lee, D. Takeuchi, and B. Razavi, “A 10Gb/s CMOS adaptive equalizer for backplane applications,” *IEEE International Solid State Circuits Conference*, Vol. 1, Feb 2005, pp. 328–601.

[21] T. Dickson, E. Laskin, I. Khalid, R. Beerkens, J. Xie, B. Karjica, and S. Voinigescu, “A 72Gb/s 2^{31}-1 PRBS generator in SiGe BiCMOS technology,” *IEEE International Solid State Circuits Conference*, Vol. 1, Feb 2005, pp. 342–602.

Electronic Devices

[22] B.G. Streetman and S. Banerjee, *Solid State Electronic Devices*, 5th edition, Prentice Hall, 2000.

[23] E.O. Johnson, "Physical limitations on frequency and power parameters of transistors," *IRE International Convention Record*, Vol. 13, Part 5, Mar 1965, 27–34.

[24] K.-H. To, Y.-B. Park, T. Rainer, W. Brown, and M.W. Huang, "High frequency noise characteristics of RF MOSFETs in subthreshold region," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, June 2003, pp. 163–166.

[25] J.D. Cressler (editor), *Silicon Heterostructure Handbook – Materials, Fabrication, Devices, Circuits, and Applications of SiGe and Si Strained-Layer Epitaxy*, CRC Press, 2005.

[26] G. Baldwin, et al., "90 nm CMOS RF technology with 9.0V I/O capability for single chip radio," *IEEE VLSI Symposium*, 2003.

[27] J.Y. Yang, et al., "0.1 um RFCMOS on high resistivity substrates for system on chip applications," *IEEE IEDM*, 2002.

Inductors

[28] H.M. Greenhouse, "Design of planar rectangular microelectronic inductors," *IEEE Transactions on Parts, Hybrids, and Packaging*, Vol. PHP-10, No. 2, Jun 1974, pp. 101–109.

[29] E. Pettenpaul, H. Kapusta, A. Weisgerber, H. Mampe, J. Luginsland, and I. Wolff, "CAD models of lumped elements on GaAs up to 18 Ghz," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 36, No. 2, Feb 1988.

[30] A.M. Niknejad and R.G. Meyer, "Analysis, design, and optimization of spiral inductors and transformers for Si RF ICs," *IEEE Journal of Solid State Circuits*, Vol. 33, No. 10, Oct 1998.

[31] A.M. Niknejad and R.G. Meyer, "Analysis of Eddy-current losses over conductive substrates with applications to monolithic inductors and transformers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 49, No. 3, Jan 2001.

[32] J.R. Long and M.A. Copeland, "The modeling, characterization and design of monolithic inductors for Silicon RF ICs," *IEEE Journal of Solid State Circuits*, Vol. 32, Mar 1997, pp. 357–369.

[33] D.J. Cassan and J.R. Long, "A 1-V transformer-feedback low noise amplifier for 5-Ghz wireless LAN in 0.18-um CMOS," *IEEE Journal of Solid State Circuits*, Vol. 38, No. 3, Mar 2003, pp. 427–435.